CHALMERS

GÖTEBORGS UNIVERSITET

Modellutvärdering av Thomas- och Matérnklusterprocesserna för nervmönster i neuropati

Model evaluation of the Thomas and Matérn cluster processes for nerve patterns in neuropathy

Examensarbete för kandidatexamen i matematik vid Göteborgs universitet
Kandidatarbete inom civilingenjörsutbildningen vid Chalmers

Michal Palak
Ruben Seyer
Jesper Söderberg
Alexander Thorén

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2021

Modellutvärdering av Thomas- och Matérnklusterprocesserna för nervmönster i neuropati

Examensarbete för kandidatexamen i matematisk statistik vid Göteborgs universitet

Alexander Thorén

Kandidatarbete i matematik inom civilingenjörsprogrammet Kemiteknik med fysik vid Chalmers

Jesper Söderberg

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk matematik vid Chalmers

Michal Palak Ruben Seyer

Handledare: Aila Särkkä
Konstantinos Konstantinou

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2021

Förord

Först och främst tackar vi naturligtvis våra handledare Aila Särkkä och Konstantinos Konstantinou, utan vilka vi aldrig ens hade fått ner ett enda ord på pappret. Våra opponenter Alexander Lehnberg, Malte Rundqvist, Emanuel Abada, Lukas Johansson och Sulaiman Carneil förtjänar också tack för deras insats i att göra detta arbete så bra som möjligt. Vi tackar också examinatorerna Ulla Dinger och Maria Roginskaya. Sist men inte minst tackar vi alla andra som bidragit med synpunkter under handledningstillfällen eller i läsutbyten: Andreas Eriksson, Mårten Åsberg, Mattias Mattsson, Simon Sundqvist, Carl Wiede, Emma Pettersson, Jennifer Krogh, Magnus Gustafsson, Ivan Flensburg, Linda Hoang, Holger Johansson, Alex Matsson, Mats Richardson och Alexander Samuelsson.

En loggbok har förts över alla gruppmedlemmars prestationer. Huvudförfattare till respektive avsnitt förtecknas i tabellen nedan, där författare för avsnitt på toppnivå avser den text innan första delavsnittet.

§	Avsnitt	Huvudförfattare
	Populärvetenskaplig presentation	Alexander
	Sammanfattning och abstract	Michal
1	Inledning	Ruben
1.1	Syfte	Ruben
1.2	Problem	Jesper
1.3	Avgränsningar	Jesper
2	Spatiala punktprocesser	
2.1	Modeller	Jesper
2.2	Ripleys K-funktion	Michal
2.3	Närmsta granne-funktionen G	Alexander
2.4	Parameterskattningar och Monte Carlo-anpassningstest	Ruben
3	Analys och resultat	
3.1	CSR-test på punktmönstren	Jesper
3.2	Individuella parameterskattningar	Jesper
3.3	Individuella anpassningstest	Michal
3.4	Gruppvisa parameterskattningar	Ruben
3.5	Gruppvisa anpassningstest	Alexander
3.6.1	Utforskande analys av förgreningspunkterna	Ruben
3.6.2	Experimentell hierarkisk modell	Alexander
4	Diskussion och slutsats	
4.1	Thomas- eller Matérnklusterprocess	Michal
4.2	Skillnader mellan grupperna	Alexander
4.3	Förkastning av de gruppgemensamma modellerna för sjuka	Jesper
4.4	Experimentell hierarkisk modell	Alexander
4.5	Generaliserbarhet och framtida arbete	Ruben
4.6	Slutsats	Ruben
A	Appendix	
A.3–A.5	Bilagor om statistikor	Alexander
A.6	Rayleighfördelning	Alexander
A.7	Exempelräkning ERL	Ruben

Populärvetenskaplig presentation

År 2019 led 463 miljoner vuxna människor med diabetes i världen; år 2045 förutsägs detta antal vara 700 miljoner [1]. Personer som lider av diabetes kan drabbas av diabetesneuropati, en sjukdom som skadar nerverna. Bland annat medför denna sjukdom att de kvarvarande nerverna blir mer grupperade än hos friska. I detta projekt undersöker vi hur man med hjälp av olika statistiska modeller kan kvantifiera och jämföra mönster av nervändpunkter hos både friska individer och individer som lider av diabetesneuropati.

Diabetesneuropati kan ligga latent i många år innan diagnos. Det är av vikt att diagnostisera sjukdomen i tidigt stadium för att kunna ge behandling som fördröjer utvecklingen av symptom. Genom att anpassa och utvärdera dessa modeller hoppas vi kunna få nya insikter om hur diabetesneuropati påverkar kroppens nervsystem.

Nervsystemet är ett komplext nätverk som löper genom hela kroppen och koordinerar kroppens olika funktioner. Nerverna leder elektrokemiska signaler mellan kroppens olika delar och låter oss bland annat se, höra, lukta samt känna. En av skillnaderna mellan nervceller och andra celler i kroppen är de trådliknande utskott som leder signaler till och från nervcellen. Varje nervcell har ett axon som leder ut signaler och flera dendriter som leder signaler till nervcellen [2]. Axoners diameter varierar mellan 0,1 och 20 mikrometer [3]. Grövre axoner är omgivna av ett elektriskt isolerande lager av myelin, vilket är en fettrik substans [4]. Diabetesneuropati innebär att axonerna försämras samt att myelinlagret skadas [5].

De mönster som bildas av dessa nervers ändpunkter kan beskrivas som en spatial punktprocess. Det finns många olika typer av spatiala punktprocesser; den mest grundläggande är Poissonprocessen. Denna process placerar slumpmässigt ut ett antal punkter på en yta där det genomsnittliga antal punkter bestäms av en parameter. Den typ av processer som undersöks här heter klusterprocesser.

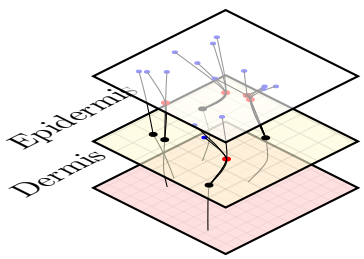
Klusterprocesser byggs upp genom att först skapa en punktprocess vars punkter ses som föräldrar. För varje förälder genererar vi sedan ytterligare ett punktmönster centrerat på föräldern. I vårt fall kommer vi framförallt att undersöka och jämföra Thomas- samt Matérnprocesserna.

Thomas- och Matérnprocessernas föräldrar är båda Poissonprocesser. Skillnaden mellan de två är att Thomasprocessens barn är fördelade enligt normalfördelningen runt en föräldrapunkt medan Matérnprocessens barn är likformigt fördelade i en cirkel runt en föräldrapunkt.

Den data vi kommer att använda är strukturerad som i figur 1. Framförallt konstruerar vi modellerna med hjälp av ändpunkterna men vi undersöker även till viss grad hur förgreningspunkterna påverkar modellerna. I figuren ser vi att de flesta nervträd endast har en förgreningspunkt men i verkligheten kan de ha många fler. Vi har endast använt data på de första förgreningspunkterna.

Det finns ett antal funktioner för att analysera och kvantifiera spatiala punktmönster. De som kommer användas i detta projekt är framförallt *Ripleys* K-funktion samt *närmsta granne-funktionen*. K-funktionen beskriver antalet punkter inom en radie r. Närmsta granne-funktionen beskriver sannolikheten att en punkts närmsta granne ligger inom ett avstånd r. Med hjälp av dessa funktioner kan man kvantifiera och jämföra graden av klustring inom vår data.

Med hjälp av *minimum contrast-metoden* anpassar vi våra modeller. *Minimum contrast-metoden* försöker hitta parametervärden som minimerar skillnaden mellan det teoretiska och uppskattade värdet på en utvald statistika. En statistika är ett värde som kvantifierar och karaktäriserar ett urval. I vårt fall använder vi oss av K-funktionen. Efter att ha anpassat modellerna genomförs ett *Monte Carlo-test* för att utvärdera dem.



Figur 1: Schematisk illustration av nervstrukturerna som studeras, med baspunkter i svart, förgreningspunkter i rött, och ändpunkter i blått.

Sammandrag

Den globala epidemin av diabetes och prediabetes har lett till en epidemi av komplikationer relaterade till dessa sjukdomar. En av de mest förekommande komplikationerna är diabetesneuropati som utmärks av axonal degeneration av nervtrådar och kan orsaka känselbortfall och smärtor. Med hjälp av konfokalmikroskopi har man kunnat observera att nervändarna i ytterhuden hos sjuka patienter tenderar att vara mer klustrade än hos friska. Därför är det viktigt att få bättre förståelse för degenerationsprocessen och den spatiala strukturen i nervmönster för att kunna upptäcka diabetesneuropati i tidigt stadium.

I detta arbete undersöks hur två matematiska modeller, Thomas- och Matérnklusterprocesser representerar nervdata i form av punktmönster, från åtta friska och sju sjuka patienter genom att anpassa båda modeller till individuell data och gruppvist. Modellerna anpassas till data från nervändarna med hjälp av minimum contrast-metoden. Vi använder K-funktionen som sammanfattningsfunktion vid anpassning av modeller, vilket beskriver det förväntade antalet grannpunkter inom radie r från godtycklig punkt i punktprocessen. Därefter testas lämplighet av de anpassade modellerna med ett sammansatt ”global envelope” test och analys av modeller med hjälp av all data.

Resultat från individuella anpassningstestet indikerar att Thomas och Matérnprocesser kan vara lämpliga för modellering av nervänderna och att det inte finns en väsentlig skillnad i prestanda mellan de två processerna. Båda processer klarade den gruppgemensamma anpassningstestet för den friska gruppen men förkastas för den sjuka gruppen vilket kan bero på större variationer hos den sjuka gruppen. Från analys av fullständig data framgick det att båda modeller tenderar att överskatta antalet kluster.

Abstract

The global epidemic of diabetes and prediabetes has led to the epidemic of complications related to these diseases. One of the most common complications is diabetic neuropathy which is characterized by axonal degeneration of nerve fibers and can cause pain and loss of sensation. With the help of confocal microscopy, it has been possible to observe that the nerve endings in the outer skin of sick patients tend to be more clustered than of healthy subjects. Therefore, it is important to gain a better understanding of the degeneration process and the spatial structure of nerve patterns in order to be able to detect diabetic neuropathy at an early stage.

This work examines how two mathematical models, Thomas and Matérn cluster process, represent nerve data in the form of point patterns, from eight healthy and seven sick patients by fitting both models to both individual data groupwise. The models are fitted to the data from nerve endings using Minimum contrast method. We use the K function as a summary function when fitting the model, which describes the expected value of the number of neighboring points within radius r from a arbitrary point in the point process. The suitability of the fitted models is then tested with a composite “global envelope” test and analysis of models using the complete data.

Results from the individual goodness of fit test indicate that Thomas and Matérn processes may be suitable for modeling nerve endings and that there is no significant difference in performance between the two processes. Both processes passed the groupwise goodness of fit test for the healthy group but are rejected for the sick group, which may be due to greater variation in the sick group. Analysis of complete data showed that both models tend to overestimate the number of clusters.

Notation

$\mathbb{E}[X]$	väntevärdet av slumpvariabel X
$\mathbb{P}\{A\}$	sannolikhet för händelse A
$\mathcal{N}(\mu, \sigma^2)$	normalfördelningen med väntevärde μ och varians σ^2
$\mathcal{N}(\boldsymbol{\mu}, \Sigma)$	multivariata normalfördelningen med väntevärden $\boldsymbol{\mu}$ och kovariansmatris Σ
$\mathbf{1}\{P\}$	indikatorfunktionen, 1 om påstående P sant och 0 annars
$\wedge$	logisk OCH-operator, sann om båda operander är sanna
∂S	randen av en mängd S i ett metriskt rum
$\lvert S \rvert$	längd/area av en 1-/2-dimensionell mängd S
$\#S$	antalet element i mängden S (till skillnad från längd-/areamått)
$\mathbf{X}$	punktprocess, en stokastisk mängd av punkter
W	observationsfönster för ett punktmönster från en punktprocess
λ	intensitet, förväntat antal observationer per enhetsyta
κ	intensitet för föräldraprocessen i klusterprocess
μ	förväntat antal dotterpunkter per kluster i klusterprocess
σ	standardavvikelse i Thomasprocessen för dotterpunkternas placering
R	radie i Matérnklusterprocessen för dotterpunktsområde
τ	allmän skalparameter för klusterprocess dvs. σ eller R
$G(r)$	närmsta granne-funktion för en stationär punktprocess $\mathbf{X}$
$K(r)$	Ripleys K-funktion, sammanfattningsfunktion för en stationär punktprocess $\mathbf{X}$
$L(r)$	transformation $\sqrt{K(r)/\pi}$ av Ripleys K-funktion
$\hat{}$	empirisk skattning av variabel eller funktion
$\bar{}$	gruppgemensam skattning av variabel eller funktion
$\lVert u - v \rVert$	Euklidiska normen, avståndet mellan punkterna u, v i planet
$\mathrm{d}(u, S)$	$\min\{\lVert u - x_i \rVert : x_i \in S\}$, minsta avståndet från punkten u till punkt i mängden S
$S_{\ominus r}$	mängden S krympt r, definierad $\{x \in S : \mathrm{d}(x, \partial S) \geq r\}$
$S_{\oplus r}$	mängden $S \subseteq U$ utökad r, definierad $\{x \in U : \mathrm{d}(x, S) \leq r\}$

Förkortningar

CSR Complete spatial randomness, då punkterna i mönstret är oberoende slumpmässigt fördelat i observationsfönstret, dvs. enligt en Poissonprocess (se avsnitt 2.1.1).

ERL Extreme rank length, mått för hur extrem en vektor är relativt en uppsättning vektorer (se avsnitt 2.4.3).

Innehåll

1 Inledning

Neuropati är en nervsjukdom som innebär att nervtrådarna successivt skadas utifrån och in, i regel med början längst ut i benen. Vanliga symtom är känselbortfall, domningar, torr hud och försämrad balans. En vanlig orsak är diabetes, men själva mekanismen bakom förloppet är ännu inte helt känd [6]. När diagnos ställs har ofta sjukdomen pågått under lång tid och de skador som hunnit uppstå upptäcks då sent. Vissa typer av diabetesneuropati medför irreversibla skador på nervsystemet [7].

Tidigare forskning har visat att förlust av nervändarna inte sker helt slumpmässigt; i stället verkar nervtätheten minska i överhuden jämfört med friska personer [8]. Aspekter som har studerats är till exempel förekomsten av nervfiber per area och nervlängd per area [9]. Nervändarna är förstås naturligt klustrade på grund av nervsystemets trädstruktur, men det förklarar inte skillnaderna mellan grupperna, och sjukdomstillståndets klustringseffekt förekommer eventuellt på en annan skala. Bland annat tycks de återstående nervändarna bilda kluster i större utsträckning än hos friska personer [8]. För att bättre förstå dessa skillnader är det av intresse att försöka kvantifiera dem och modellera dem med lämpliga statistiska verktyg.

På senare tid har spatiala punktprocesser, där varje punkt utgörs av en nervände i överhuden, använts för detta ändamål, bland annat eftersom deras konstruktion tar hänsyn till det faktum att nervändarna förgrenats från en cellkropp. Parameterskattningar för denna typ av modeller är speciellt problematiskt; den vanliga metoden med maximum likelihood-skattare för parametrarna går inte att använda, eftersom de relevanta likelihood-funktionerna inte kan uttryckas på sluten form. Därför behövs en alternativ lösning: minimum contrast-metoden, där parametrarna anpassas så att teoretiska karakteristiker hos processerna motsvarar de empiriska observationerna så väl som möjligt, som i sin tur för med sig nya val och problem som påverkar resultatet (se avsnitt 2.4). Föregående kandidatarbeten inom detta område [10, 11] har fokuserat på just dessa parameterskattningar och valet av bl.a. integrationsgräns och statistika.

Vårt fokus ligger i stället på själva modellerna. I detta arbete undersöker vi och utvärderar Thomas- och Matérnklusterprocessen som modell för nervmönster. Båda dessa modellerar kluster men skiljer sig åt i hur nervändarna placeras relativt så kallade föräldrapunkter: Thomasprocessen använder en isotropisk (dvs. rotationssymmetrisk) normalfördelning, medan Matérnklusterprocessen använder en likformig fördelning inom en viss radie.

De anpassade modellerna kan bidra till en bättre förståelse av de friska mönstren, mönstren som uppkommer av sjukdomen samt metoder för diagnostik och behandling. Tidig klassifikation av mönstren är synnerligen intressant eftersom sjukdomen kan ligga vilande innan symtom bryter ut och diagnos kan ställas, samtidigt som tidig behandling kan undvika att neuropatin förvärras [7]. Sådana metoder som kan användas kliniskt är naturligtvis det långsiktiga målet inom problemområdet. Vår utvärdering av modellerna och modellparametrarna från våra anpassningar kan ligga till grund för framtida arbete inom klassifikation av nervmönster. Det är också möjligt att ytterligare beskrivning av och skillnader mellan friska och sjuka mönster av nervändpunkter kan hjälpa till att klargöra varför nervsystemet genomgår de förändringar som det gör till följd av skadorna. Slutligen kan även tillämpningen av modellerna på problemet leda till nya statistiska insikter om dem.

1.1 Syfte

Syftet med detta arbete är att jämföra två olika modeller för nervändarna baserade på klusterprocesser, Thomas- och Matérnklusterprocessen, utifrån en given uppsättning data från biopsier (vävnadsprov) på två grupper, friska och sjuka. Genom att förstå modellerna, anpassa dem till data, och analysera parameterskattningarna både individuellt och för de två grupperna, kan lämpligheten hos modellerna och skillnaderna i nervändarna mellan grupperna kvantifieras, så att mönstren bättre förstås.

1.2 Problem

Den huvudsakliga uppgiften med arbetet är att anpassa spatiala modeller till nervändpunkter, för att sedan analysera skillnader i modellerna för friska och sjuka nervmönster och därigenom utvärdera modellerna. Eftersom nervändarna tenderar att vara klustrade undersöks två klusterprocesser, Thomas- och Matérnklusterprocessen. Båda modellerna anpassas till given data från åtta personer

som är friska samt sju personer som lider av måttlig diabetesneuropati. Modellanpassningen görs både individuellt och gruppvis. För att se vilken modell som passar observationerna bäst i var och ett av fallen undersöks anpassningsgraden för dem och jämförs mellan de två processerna.

Först verifieras att nervpunkterna inte är slumpvis utplacerade, det vill säga förkasta att punkterna karaktäriseras av CSR, ”complete spatial randomness”. Vidare vill vi skatta parametrarna för Thomas- samt Matérnklusterprocessen för att bygga upp modellerna, dels för varje individuellt punktmönster men också för hela grupperna friska samt måttligt sjuka. En analys av modellerna mot data med Monte Carlo-anpassningstest görs för att utvärdera om modellerna kan beskriva observationerna, och om så är fallet, vilken av processerna som ger bäst anpassningsgrad.

1.3 Avgränsningar

I arbetet diskuteras inte insamling av data och inte heller bearbetning av data, då undersökningen sker i ett redan givet dataset och fokuset i arbetet ligger på att utvärdera och jämföra modellerna. Vidare analyseras i modellerna endast ändpunkter på nerverna och inte hela nervfibrerna eller andra strukturer, då de endast anpassas på information om ändpunkterna. Arbetets fokus ligger inte på att klassificera sjukdomsbilden utan på att analysera och utvärdera Thomas- och Matérn-klusterprocesserna för datan. I framtida forskning skulle resultatet kunna användas som grund i syfte att förbättra detektion samt klassificering av graden av neuropati.

2 Spatiala punktprocesser

I detta avsnitt redogör vi för den teori som krävs för att underbygga den senare analysen. Vi sammanfattar kort relevant teori inom spatiala punktprocesser i allmänhet och teorin bakom modellerna, samt beskriver de statistikor som används i skattning av parametrarna och i undersökning av anpassningsgraden av modellerna. I appendix A.3 och A.4 beskriver vi två andra statistikor som ofta används, men som inte används i vårt arbete. Vi beskriver också hur parametrarna skattas, både för individuella nervmönster och för hela grupper sedda som replikat, och hur anpassningstest kan göras på modellerna vi tar fram för att fastställa huruvida de har förmågan att beskriva observationerna.

2.1 Modeller

En spatial punktprocess är en samling av slumpmässiga punkter och ett utfall av en sådan process kallas ett punktmönster. Det finns många olika punktprocesser och beroende på vilken som används skapas det punktmönster med olika struktur. En punktprocess kan skapa punktmönster vilket uttrycks i $n = 1, 2, \ldots$ dimensioner. I detta arbete behandlas endast punktmönster som ligger i $\mathbb{R}^2$, det vill säga att punktprocesserna också befinner sig i $\mathbb{R}^2$.

Spatiala punktprocesser kan användas som statistiska modeller för punktmönster. Exempel på när punktmönster uppkommer är då positioner markeras med punkter där en viss art av träd växer, var guld hittats i berggrunden eller positioner av hus där anmälan om magbesvär skett. Punkterna kommer då bilda ett punktmönster. Med hjälp av en punktprocess kan en statistisk undersökning utföras där parametrar i processen skattas från de insamlade datapunkterna för att skapa en modell för datan. En sådan modell skulle kunna ligga som grund till att få veta var någonstans guldfyndigheter kan förväntas hittas eller för att snabbare hitta om hälsofarliga nivåer av ämnen finns i exempelvis bostäders dricksvatten.

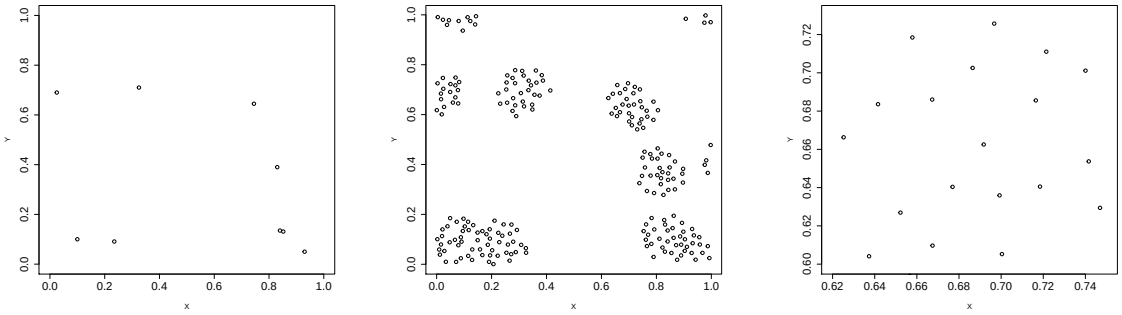
Spatiala punktprocesser brukar delas in i tre huvudgrupper beroende av vilken typ av mönster realisationen av processen får: CSR, reguljära och klustrade. CSR står för ”complete spatial randomness” vilket innebär att punkterna är helt spatialt slumpmässigt utlagda utan någon struktur samt att punkterna är helt oberoende av varandra.

Reguljära punktmönster är mönster där punkterna ligger ordnade. Ett reguljärt punktmönster fås till exempel vid undersökning av cellers position och där en punkt sätts där cellens kärna är lokaliserad. Då cellen har en volym finns en viss radie där andra cellkärnor inte kan befinna sig. Då celler ligger nära varandra ges ett mönster där punkter ligger ordnat med liknande avstånd från varandra, det vill säga ett reguljärt mönster.

Ett klustrat punktmönster är ett mönster som karaktäriseras av att punkter ligger klumpade i ett eller flera kluster. Klustrade punktmönster uppkommer ofta vid undersökningar av växtriket. Ett av

de mest klassiska exemplen när klustermönster uppkommer är vid granskning av trädpopulationer. De flesta frön från ett så kallat föräldraträd hamnar i närområdet vilket leder till att dotterplantorna med större sannolikhet börjar växa nära föräldraträdet. Dotterplantorna kommer då att bilda kluster runt föräldraplantorna.

För att få relevanta resultat är det viktigt att välja en process som ger rätt typ av punktmönster. Det är också viktigt att förstå att skalan och vilket objekt som studeras kan ändra vilken typ av punktmönster som uppkommer. Ett exempel på när skala och objekt spelar stor roll på punktmönstrets utseende är vid beskrivning av cirkulära bakterier som faller slumpmässigt ner i en petriskål och sedan börjar celldela. Efter en tid lokaliseras bakterierna och det sätts ut en punkt i mitten av varje cell. Vid analys av en liten del av petriskålen kommer ett reguljärt mönster uppvisas på grund av cellens volym enligt ovan, se figur 2c. Undersöks däremot ett större område kommer ett klustrat punktmönster att bildas med kluster där ursprungsbakterien landade, se figur 2b. Om positionerna för de ursprungliga bakterierna (föräldrapunkter) markeras kommer ett CSR punktmönster att uppstå, se figur 2a.



(a) Positionen av föräldrapunkter har en slumpmässig bild vilken påvisar ett mönster som är CSR.

(b) Bakterieodlingen uppvisar en tydlig klustring av dotterpunkter.

(c) Närbild av en del av ett kluster visar en ordnad struktur.

Figur 2: Genererad bild av en hypotetisk bakterieodling.

Inom statistisk analys av punktmönster finns det ett flertal olika punktprocesser att använda. I denna studie kommer två olika klusterprocesser att undersökas, Thomasprocessen och Matérnklusterprocessen [12, s. 463, 13, s. 376]. Processerna som undersöks under arbetet kommer att vara *stationära* samt *isotropa*. En isotropisk process är rotationsinvariant, vilket betyder att fördelningen inte ändras vid rotation [12, s. 146–147]. En punktprocess är stationär om den är invariant vid förflyttning [12, s. 147]. Om punktprocessen är stationär kommer genomsnittliga antalet punkter per enhetsyta, intensiteten λ, vara konstant.

2.1.1 Poissonprocessen

Inom spatiala punktprocesser är Poissonprocessen en viktig process som ligger till grund för mer komplicerade spatiala punktprocesser. Ett informellt sätt att beskriva en Poissonprocess är en process vars punktmönster realiseras som ett mönster av punkter vilka är oberoende och helt slumpmässigt utsatta, det vill säga CSR. För att skapa föräldrapunkter för både Thomas- och Matérnklusterprocesser används Poissonprocessen. Visualisering av en Poissonprocess ses i figur 3a och 4a. Formellt definieras en homogen Poissonprocess av [14, s. 61]

1. för något $\lambda > 0$ och en godtycklig yta A i $\mathbb{R}^2$, antalet punkter n är Poissonfördelade med väntevärdet $\lambda |A|$ där $|A|$ är arean av A.

2. givet antalet punkter n, är alla punkterna oberoende och likformigt fördelade på A.

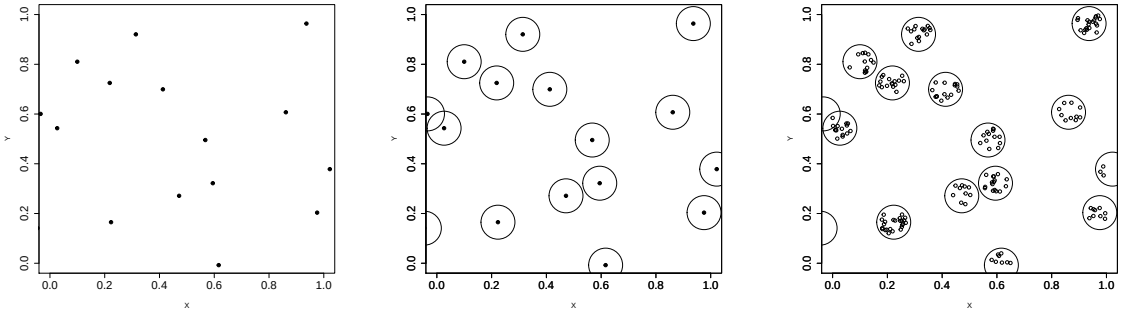
2.1.2 Matérnklusterprocessen

Matérnprocessen är en klusterprocess, där föräldrapunkter genereras först genom en Poissonprocess med en intensitet κ. Vid undersökning i $\mathbb{R}^2$ genereras dotterpunkter likformigt inom en disk med

konstant radie R runt föräldrapunkterna . Dotterpunkterna fördelas endast inom disken vilken har arean $R^2\pi$, fördelningen sker med en Poissonprocess med intensiteten $\mu/R^2\pi$. I disken runt en föräldrapunkt blir μ det förväntade antalet dotterpunkter. Utanför disken är intensiteten noll och det förväntade antalet dotterpunkter är $\mu\kappa\,|A|$ för hela området A. Efter att dotterpunkterna är genererade så tas föräldrapunkterna bort.

I Matérnklusterprocessen tas hänsyn till dotterpunkter som kan uppstå inne i område A medan föräldrapunkten är lokaliserad utanför området som studeras. För att ta hänsyn till detta genereras föräldrapunkterna på ett större område $A_{\oplus R}$ utökat med avståndet R, vilka genererar dotterpunkter som ovan beskrivet. Endast dotterpunkter som hamnar inom område A studeras.

Processen är stationär samt isotropisk om föräldrapunkter och dotterpunkter skapas som beskrivet ovan. Det finns inte någon positionsberoende variation i fördelningen av punkterna, eftersom föräldrapunkterna är CSR över hela området samt att dotterpunkterna är fördelade med en Poissonprocess över disken. En visualisering av en Matérnklusterprocess ses i figur 3.

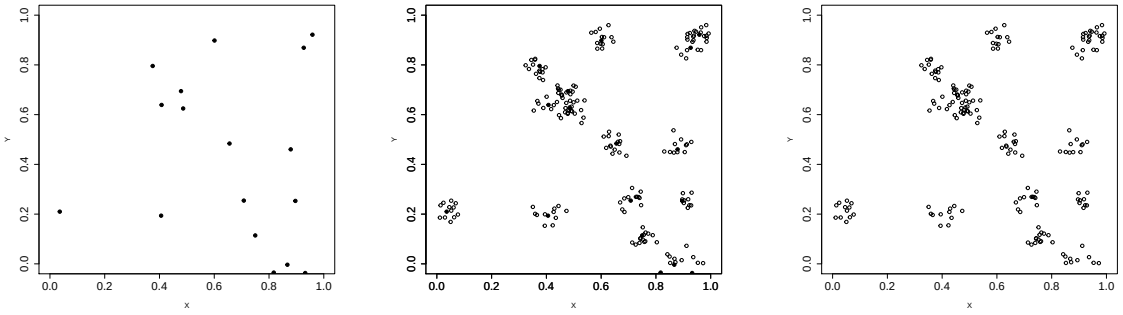


(a) Position för föräldrapunkterna (b) Diskar med områden där intensiteten inte är noll (c) Dotterpunkter utsatta samt föräldrapunkterna borttagna

Figur 3: Visualisering av en Matérnprocess, med intensiteten av föräldrapunkterna $\kappa = 10$, radien för disken $R = 0.05$ samt genomsnittliga antalet dotterpunkter $\mu = 15$.

2.1.3 Thomasprocessen

Thomasprocessen är en klusterprocess vilken är uppbyggd så att föräldrapunkter genereras enligt en Poissonprocess, CSR, med en intensitet κ. Från föräldrapunkterna genereras Poi(μ) dotterpunkter per kluster vilka är oberoende och normalfördelade $\mathcal{N}(0, \sigma^2 I)$, där 0 är föräldrapunkten. Produkten $\sigma^2 I$ bestämmer spridningsytan av klustret där σ^2 är variansen och I är en identitetsmatris, 2×2 för punktmönster i två dimensioner. Efter att dotterpunkterna är utsatta tas föräldrapunkterna bort. Med detta fås då ett genererat klustrat punktmönster enligt Thomasprocessen.



(a) Position för föräldrapunkterna (b) Dotterpunkter utsatta runt föräldrapunkterna (c) Föräldrapunkterna borttagna

Figur 4: Visualisering av en Thomasprocess, med intensiteten av föräldrapunkterna $\kappa = 10$, variansen $\sigma^2 = 0.025$ samt genomsnittliga antalet dotterpunkter $\mu = 12$.

Likt Matérnprocessen tas dotterpunkter med där föräldrapunkterna ligger utanför observationsfönstret A. Då Thomasprocessens kluster är normalfördelade runt föräldrapunkten med variansen σ^2 är det teoretiskt möjligt att en föräldrapunkt har en dotterpunkt långt bort, relativt σ. Föräldrapunkter genereras på ett område där en längd d anges för att vidga varje sida av A. Vanligt värde på d är 4σ, med det är sannolikheten $p \approx 1/31000$ att en dotterpunkt landar i A då föräldrapunkten ligger på avståndet 4σ från A [12, s. 468]. Uppbyggnaden av ett punktmönster med Thomasprocessen visualiseras i figur 4.

Då dotterpunkterna i en Thomasprocess har en normalfördelning runt en föräldrapunkt $\mathcal{N}(0, \sigma^2 I)$ där föräldrapunkten ligger i 0 enligt ovan, blir processen stationär och isotropisk. Föräldrapunkterna är fördelade enligt CSR vilket gör att det inte är någon fördelningsvariation i någon riktning. Dotterpunkterna har inte heller någon fördelningsvariation då placeringen endast beror på avståndet till föräldrapunkten och inte riktning.

2.2 Ripleys K-funktion

Ripleys K-funktion är en sammanfattningsfunktion som beskriver beteendet hos spatiala punktprocesser. Den beskriver det förväntade antalet punkter inom en radie r från en godtycklig punkt i processen. För en stationär punktprocess $\mathbf{X}$ med intensitet λ är funktionen definierad enligt [12, s. 205]

$$K(r) = \lambda^{-1}\mathbb{E}[\text{antalet grannar till } u \text{ inom avstånd } r \mid u \in \mathbf{X}]. \tag{1}$$

För en helt slumpmässig process (CSR) blir det förväntade antalet grannar inom avstånd r lika med produkten mellan arean av cirkeln med radie r och intensiteten, nämligen $\lambda\pi r^2$. Alltså blir värdet av funktionen $K(r) = \lambda^{-1}(\lambda\pi r^2) = \pi r^2$ för en helt slumpmässig process. För en klustrad punktprocess blir $K(r) > \pi r^2$ och för en reguljär punktprocess blir $K(r) < \pi r^2$.

En vanligt förekommande transformation av K-funktionen är L-funktionen definierad enligt

$$L(r) = \sqrt{\frac{K(r)}{\pi}}. \tag{2}$$

L-funktionen används för att stabilisera variansen och den stabiliserade $L(r) - r$ funktionen används för att underlätta visuell bedömning av grafen då $L(r) - r = 0$ för CSR [12, s. 207].

2.2.1 Skattning

För att skatta K-funktionen skall den empiriska $\hat{K}$-funktionen användas. Den är definierad enligt

$$\hat{K}(r) = \frac{|W|}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbf{1}\{d_{ij} \leq r\} e_{ij}(r), \tag{3}$$

där $|W|$ är arean av observationsfönstret, n är antalet punkter, $d_{ij} = \|x_i - x_j\|$ är avståndet mellan punkter x_i och x_j, och e_{ij} är vikten för korrektion av kanteffekter som beskrivs i avsnitt 2.2.2.

2.2.2 Korrektion av kanteffekter

För punkter som ligger inom ett avstånd r från observationsfönstrets kanter riskerar vi att underskatta det riktiga antalet grannpunkter eftersom grannpunkter kan ligga utanför observationsfönstret. Det problemet förekommer ofta vid analys av spatiala punktprocesser och kallas för kanteffektproblemet. Ett exempel på kanteffektproblemet visas i figur 5.

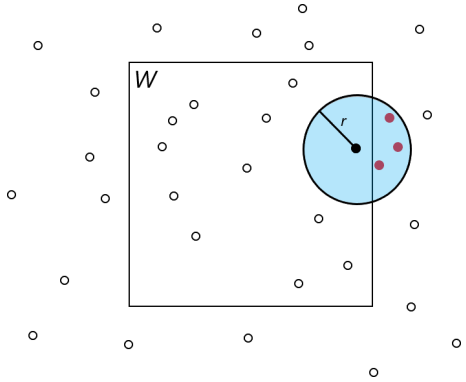
För att ta hänsyn till kanteffekter använder vi oss av vikten e_{ij} för den isotropa korrektionen av kanteffekter [12, s. 216] där

$$e_{ij}(r) = \frac{1}{p(x_i, r)}, \tag{4}$$

där

$$p(x_i, r) = \frac{|W \cap \partial C(x_i, r)|}{2\pi r} \tag{5}$$

är andelen av omkretsen av cirkeln $C(x_i, r)$ som ligger i observationsfönstret W.



Figur 5: Illustration av kanteffekter. Inom avstånd r från den svarta punkten ligger tre grannpunkter. Alla grannpunkterna ligger dock utanför observationsfönstret W. Det observerade antalet blir därför noll, vilket inte stämmer med det faktiska antalet.

2.2.3 Replikat

För att bestämma gemensam skattning $\bar{K}$ av $\hat{K}_1, \hat{K}_2, ..., \hat{K}_m$ för replikerad data med m punktmönster, beräknar vi ett viktat medelvärde. För $\hat{K}$ med den isotropa korrektionen för kanteffekter blir skattningen av $\bar{K}$ [12, s. 681]

$$\bar{K}(r) = \frac{\sum_{i=1}^{m} n_i(n_i - 1)\hat{K}_i(r)}{\sum_{i=1}^{m} n_i(n_i - 1)}, \tag{6}$$

där n_i är antalet punkter i punktmönstret i. Det finns flera sätt att välja vikter, här använder vi vikten $n_i(n_i - 1)$ för att inte anta samma intensitet för de olika punktmönstren.

2.2.4 Teoretiska K-funktionen för Thomas och Matérnprocessen

Den teoretiska K-funktionen används vid skattning av parametrar för Thomas och Matérnprocessen som beskrivs i 2.4.2, där vi anpassar modellens teoretiska K-funktion till den empiriska $\hat{K}$-funktionen (3) från data.

För Thomasprocessen med parametrar $\boldsymbol{\theta} = (\kappa, \sigma, \mu)$ ges den teoretiska K-funktionen av [12]

$$K(r) = \pi r^2 + \frac{1}{\kappa}\left(1 - \exp\left(-\frac{r^2}{4\sigma^2}\right)\right). \tag{7}$$

För Matérnprocessen med parametrar $\boldsymbol{\theta} = (\kappa, R, \mu)$ ges den teoretiska K-funktionen av [12]

$$K(r) = \pi r^2 + \frac{1}{\kappa} h\left(\frac{r}{2R}\right), \tag{8}$$

där

$$h(z) = \begin{cases} 2 + \frac{1}{\pi}\left[(8z^2 - 4)\arccos(z) - 2\arcsin(z) + 4z\sqrt{(1-z^2)^3} - 6z\sqrt{1-z^2}\right] & z \leq 1 \\ 1 & z > 1 \end{cases}. \tag{9}$$

2.3 Närmsta granne-funktionen G

Närmsta granne-funktionen $G(r)$ för en stationär punktprocess $\mathbf{X}$ beskriver sannolikheten att avståndet från en godtycklig punkt $x \in \mathbf{X}$ till sin närmsta granne är mindre än eller lika med r. För en stationär punktprocess $\mathbf{X}$ definieras funktionen som [12, s. 262]

$$G(r) = \mathbb{P}\{\mathrm{d}(u, \mathbf{X} \setminus u) \leq r \,|\, u \in \mathbf{X}\}, \tag{10}$$

där

$$\mathrm{d}(u, \mathbf{X}) = \min\{\|u - x_i\| : x_i \in \mathbf{X}\}, \tag{11}$$

alltså det minsta avståndet från en punkt u till en annan punkt som tillhör $\mathbf{X}$.

Om $\mathbf{X}$ är en Poissonprocess med intensitet λ är

$$G_{\text{pois}}(r) = 1 - \exp\left(-\lambda\pi r^2\right). \tag{12}$$

Uppskattade $\hat{G}(r)$ plottas ofta mot denna funktion som referens.

2.3.1 Skattning

Närmsta granne-funktionen kan uppskattas från data som en empirisk fördelningsfunktion. För att uppskatta funktionen räknas alla punkter vars närmsta granne ligger inom ett avstånd r [12, s. 285],

$$\hat{G}(r) = \frac{1}{\#\{\mathbf{X} \cap W\}} \sum_{i=1}^{n} \mathbf{1}\{\mathrm{d}(x_i, \mathbf{X} \cap W \setminus x_i) \leq r\}. \tag{13}$$

Notera att denna uppskattning inte tar hänsyn till kanteffekter vilket resulterar i att den får negativa systematiska fel. Till exempel är det möjligt att i ekvation (13)

$$\mathrm{d}(x_i, \mathbf{X} \setminus x_i) \leq \mathrm{d}(x_i, \mathbf{X} \cap W \setminus x_i), \tag{14}$$

alltså att det sanna minsta avståndet till en punkt är mindre än det observerade minsta avståndet.

2.3.2 Kaplan-Meier-kantkorrektion för G(r)

Det sanna avståndet till en punkts närmsta granne ges av

$$T_i = \mathrm{d}(x_i, \mathbf{X} \setminus x_i)$$

men när vi endast kan observera de punkter i vår tillgängliga data, alltså de punkter som faller i observationsfönstret W, kan vi inte alltid observera detta avstånd. Det observerade avståndet till den närmsta grannen blir

$$t_i = \mathrm{d}(x_i, \mathbf{X} \cap W \setminus x_i).$$

I vissa fall kommer detta avstånd att vara större än det sanna avståndet, alltså $t_i \geq T_i$, som vi kan se i figur 5. Låt

$$c_i = \mathrm{d}(x_i, \partial W)$$

vara avståndet till fönstrets kant och definiera det *observerade brytningsavståndet*

$$\widetilde{t_i} = \min(t_i, c_i).$$

Detta kan liknas med en ballong som blåses upp runt punkten x_i och sprängs när den antingen rör en annan punkt eller fönstrets kant [12, s. 289]. Om $\widetilde{t_i} = c_i$ är avståndet till fönstrets kant mindre än avståndet till punktens närmsta granne. Det observerade avståndet blir då större än eller lika med det sanna avståndet, $c_i \leq T_i \leq t_i$. Om istället $t_i \leq c_i$ vet vi att $t_i = T_i$, alltså att det sanna närmsta avståndet har observerats. Låt nu också

$$D_i = \mathbf{1}\{t_i \leq c_i\}$$

vara en indikator som visar om det sanna avståndet har observerats. Kaplan-Meier uppskattningen av $G(r)$ ges då av [12, s. 289]

$$\hat{G}_{km}(r) = 1 - \prod_{s \leq r} \left(1 - \frac{\#\{i : \widetilde{t_i} = s, D_i = 1\}}{\#\{i : \widetilde{t_i} \geq s\}}\right). \tag{15}$$

Täljaren i bråket ovan är antalet punkter vars sanna avstånd till sin närmsta granne är exakt s och nämnaren är antalet punkter vars observerade brytningsavstånd är minst s. Enligt ballongliknelsen är detta bråk sannolikheten att ballongen sprängs när ballongen har blåsts upp till radie s givet att den överlever minst så länge [12, s. 290]. Parentesen uttrycker då sannolikheten att överleva längre än s. Produkten av alla avstånd s blir då sannolikheten att överleva vid tidpunkt större än r, alltså $1 - G(r)$.

2.4 Parameterskattningar och Monte Carlo-anpassningstest

Ett av huvudmålen i arbetet är att utifrån data skatta parametrar $\boldsymbol{\theta} = (\kappa, \tau, \mu)$ för de två spatiala punktprocesserna, där τ är skalparametern dvs. σ för Thomasprocessen och R för Matérnklusterprocessen. Tyvärr går inte standardverktyget *maximum likelihood-metoden* att tillämpa, eftersom likelihood-funktionen i allmänhet inte kan uttryckas på sluten form [15, s. 176]. Till skillnad från likelihood-funktionen så finns det för våra processer däremot en enkel form för K-funktionen (och parkorrelationsfunktionen, se appendix A.3). Vi använder därför en annan metod, där vi försöker anpassa $\hat{\boldsymbol{\theta}}$ så att första och andra ordningens karakteristiker (dvs. intensiteten respektive K-funktionen eller parkorrelationsfunktionen) teoretiskt motsvarar de skattade dito i modellen [12, s. 470]. Till sist utforskar vi metoder för att utvärdera anpassningen av modellerna.

2.4.1 Skattning av intensiteten

För punktmönster från stationära processer (likt våra) skattas enkelt intensiteten genom den empiriska punkttätheten

$$\hat{\lambda} = \frac{n}{|W|}, \tag{16}$$

där n är antalet datapunkter i punktmönstret och W är observationsfönstret [13, s. 189].

För en serie replikat för vilka vi har skattningar $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$ och observationsfönster $W_1, \ldots, W_m$ kan man sedan bestämma en gemensam skattning [13, s. 261]

$$\bar{\lambda} = \frac{\sum_{i=1}^{m} \hat{\lambda}_i \, |W_i|}{\sum_{i=1}^{m} |W_i|}. \tag{17}$$

2.4.2 Minimum contrast-metoden

Minimum contrast-metoden beskrivs enklast som en slags olinjär kurvanpassning [12, s. 470]. Vi skattar parametrarna κ, τ i $\boldsymbol{\theta}$ så att de teoretiska värdena $S_{\boldsymbol{\theta}}(r)$ för vår utvalda andra ordningens statistika S (t.ex. K-funktionen) motsvarar de empiriska dito $\hat{S}(r)$ så väl som möjligt, genom att vi försöker minimera *kontrasten*

$$D(\boldsymbol{\theta}) = \int_a^b \left| \hat{S}(r)^q - S_{\boldsymbol{\theta}}(r)^q \right|^p dr \tag{18}$$

som funktion av parametrarna $\boldsymbol{\theta}$, för några metodparametrar $0 \leq a < b$ och $p, q > 0$ [12, s. 483–484]. Notera att S i regel inte beror på μ, som därför inte kan skattas direkt genom denna metod; med hjälp av skattningarna $\hat{\lambda}, \hat{\kappa}$ skattas sedan μ enligt $\hat{\mu} = \hat{\lambda}/\hat{\kappa}$ [13, s. 375].

2.4.3 Monte Carlo-test av anpassningen

Efter att vi skattat modellparametrarna behöver vi utvärdera själva anpassningen. Vi har inget klassiskt kvantitativt mått på anpassningens kvalitet för klustermodeller [12, s. 486], men vi kan åtminstone göra hypotestester om modellernas förmåga att förklara observationerna. Ett *Monte Carlo-test* i det här sammanhanget går till som följer: [12, s. 384–386]

1. Vi simulerar N stycken punktmönster från den modell som utgör nollhypotesen (dvs. nollhypotesen är att observationerna beskrivs av modellen).

2. Dessa och det observerade punktmönstret (med index 0) reduceras till $N + 1$ stycken värden $T_0, T_1, \ldots, T_N$ med hjälp av någon teststatistika för vilken mindre värden anses mer extrema och osannolika under nollhypotesen.

3. Testets p-värde bestäms till $p = \frac{1}{N+1} \sum_{i=0}^{N} \mathbf{1}\{T_i \leq T_0\}$, och nollhypotesen förkastas om denna underskrider någon förutbestämd signifikansnivå α.

Det hela bygger på en enkel princip av symmetri. Under nollhypotesen torde de $N+1$ olika mönstren vara helt statistiskt utbytbara, och vi kan således inte avgöra vilket av dem som är den faktiska

observationen. Sannolikheten att det observerade punktmönstret ger åtminstone den k:te minsta teststatistikan blir därför precis $k/(N+1) = p$ [12, s. 385–387].

Valet av teststatistika är förstås avgörande för testets karaktär. En enkel strategi är att betrakta värdet av någon statistika $\hat{S}(r)$ för punktmönstret, lämpligen en annan än den använd i skattningen, för ett fixt r, och låta teststatistikan vara värdets minsta avstånd från ändarna i en sorterad lista av $\hat{S}(r)$ för alla mönstren [12, s. 391]. Detta kräver dock ett rimligt antagande om vilket avstånd r som är av intresse. I stället vill vi ta hänsyn till att statistikorna för punktmönstret faktiskt är funktioner, och därmed utveckla denna strategi till att omfatta många värden. Däremot är det, till skillnad från enstaka värden, inte uppenbart hur funktioner (i praktiken diskretiserade vektorer av funktionsdata) kan rangordnas, eftersom trenden kan variera mellan funktionskurvorna. Flera olika metoder har utvecklats som är olika lämpliga för olika dataset (se [16]). Vi beskriver här metoden *extreme rank length (ERL)*, som är enkel men lämplig för ett större antal simulationer [16, 17].

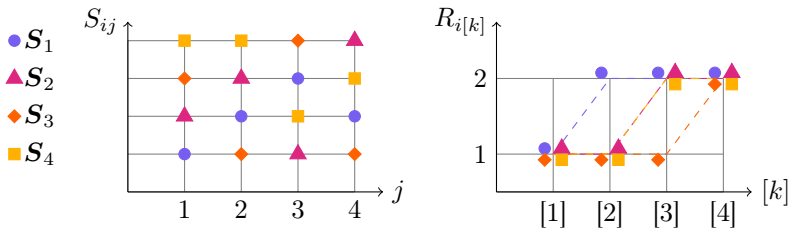
För varje punktmönster $i = 0, \ldots, N$ i ett test beräknar vi från statistikan en vektor av värden $\boldsymbol{S}_i := (S_{i1}, \ldots, S_{im}) := (\hat{S}_i(r_1), \ldots, \hat{S}_i(r_m))$ utifrån någon bestämd diskretisering $r_1, \ldots, r_m$ som har gjorts. Vi definierar först *rangen* av ett värde t_i i en lista $t_1, \ldots, t_N$ som dess position i listan sorterad i stigande ordning. Således har det minsta värdet rang 1 och det största rang N. Vid lika värden definieras rangen som genomsnittet av positionerna [16]. Vi definierar sedan den *punktvisa extrema rangen*

$$R_{ij} = \min\{r_{ij},\ N + 1 - r_{ij}\} \tag{19}$$

där r_{ij} är rangen av S_{ij} i listan $S_{0j}, \ldots, S_{Nj}$. Observera att både ovanligt stora och ovanligt små värden av S_{ij} betraktas som ”extrema”. Eftersom det är möjligt att R_{ij} inte är lägst (och mest extrem) för ett givet i för alla j behöver vi ett sätt att skilja lika punktvisa extrema ranger åt. Därför definierar vi den *globala extrema rangen*

$$E_i = \frac{1}{N} \sum_{i'=0}^{N} \mathbf{1}\{\exists d \leq m : (\forall k < d : R_{i'[k]} = R_{i[k]}) \wedge (R_{i'[d]} < R_{i[d]})\} \tag{20}$$

där $R_{i[k]}$ är det värde R_{ij} som har position k i listan $R_{i1}, \ldots, R_{im}$ sorterad i stigande ordning (utan hänsyn till lika värden, som kan ha godtycklig inbördes ordning). Om flera punktmönster i ger upphov till samma lägsta punktvisa extrema rang för ett avstånd j betraktar vi alltså för hur många j de har det lägsta värdet, därefter hur många j de har det näst lägsta värdet osv. (se figur 6 och exempel i appendix A.7). Måttet E_i är normaliserat mellan 0 och 1, ger mycket sällan lika värden på större datamängder [16, 17], och är det vi använder som teststatistika T_i.



Figur 6: Illustration av globala extrema rangen. Från råvärden S_{ij} t.v. bestäms de punktvisa rangerna R_{ij} (t.ex. $R_{11} = 1,\ R_{21} = 2,\ R_{31} = 2,\ R_{41} = 1$), som sedan sorteras till $R_{i[k]}$ t.h. för varje serie. Ordningen mellan E_i kan då utläsas grafiskt från kurvorna; i detta fall $E_1 = 3/4,\ E_2 = 1/4,\ E_3 = 0,\ E_4 = 1/4$. (I ett litet exempel som detta med få värden är lika E_i ej oväntat.)

Ytterligare en viktig detalj påverkar styrkan av testet avsevärt: medan vår nollhypotes bygger på parametrar som skattas från observationer, kommer den faktiska observationen från motsvarande modell (om den nu är giltig) med okända ”sanna” parametrar. Punktmönstren är därför egentligen inte utbytbara, och de skattade parametrarna ger med större sannolikhet teststatistikor fördelade runt värdet för observationen [12, s. 388]. För att kompensera för detta fenomen gör vi flera sådana tester för att uppskatta rätt p-värde: [16]

1. Vi gör ett huvudtest med N simuleringar för de skattade modellparametrarna, vilket ger ett p-värde p_0 enligt ovan.

2. Från modellen simuleras $M-1$ mönster som vi skattar parametrarna för.

3. Dessa $M-1$ uppsättningar parametrar utgör nu nollhypoteser i $M-1$ stycken test med N simuleringar, med p-värden $p_1, \ldots, p_{M-1}$ enligt ovan.

4. Det korrigerade testets p-värde blir då $p' = \frac{1}{M}\sum_{i=0}^{M-1} \mathbf{1}\{p_i \leq p_0\}$, och nollhypotesen (att modellen är lämplig) förkastas om denna underskrider någon förutbestämd signifikansnivå α.

Testen kan visualiseras grafiskt som s.k. *envelope*, där kurvan för observationens teststatistika lämnar detta envelope om och endast om p-värdet understiger en förutbestämd signifikansnivå α (vilket då skulle innebära att nollhypotesen förkastas).

2.4.4 Multipeltestning

Monte Carlo-testet beskrivet ovan testar en nollhypotes mot en enstaka observation, dvs. ett enstaka punktmönster. I datan har vi dock inom varje grupp nervmönster från flera patienter. Ska vi utföra tester på alla observationerna måste vi också justera signifikansnivån därefter, eftersom vi annars riskerar att göra det sammanslagna testet för känsligt. Om vi inom en grupp har N punktmönster, och därför utför N tester enligt ovan, korrigerar vi signifikansnivån α enligt Šidák [18, 19] till $\bar{\alpha} = 1 - (1-\alpha)^{1/N}$. Om nollhypotesen förkastas i ett av testen vid nivå $\bar{\alpha}$ kan vi då förkasta den för gruppen som helhet vid nivå α.

3 Analys och resultat

Vi genomför vår analys på ett dataset av nervmönster (se [20]) från dels åtta stycken friska personer, betecknade NORMAL, dels sju stycken personer med måttliga symptom av diabetesneuropati, betecknade MODERATE. Punktmönstren från deras nervändar behandlas både individuellt och gruppvis enligt avsnitt 1.2. Utöver själva ändpunkterna har vi även data om de första förgreningspunkterna och data om klustertillhörighet, som behandlas i de sista delavsnitten. Alla punktmönstren visas i figur 14 och information om antalet punkter och observationsfönster listas i tabell 1, båda i appendix A.2. För själva genomförandet använder vi oss av `R` [21] med paketen `spatstat` [12] (version 2.1-0) och `GET` [16, 17]. Instruktioner för att se koden finns i appendix A.1.

3.1 CSR-test på punktmönstren

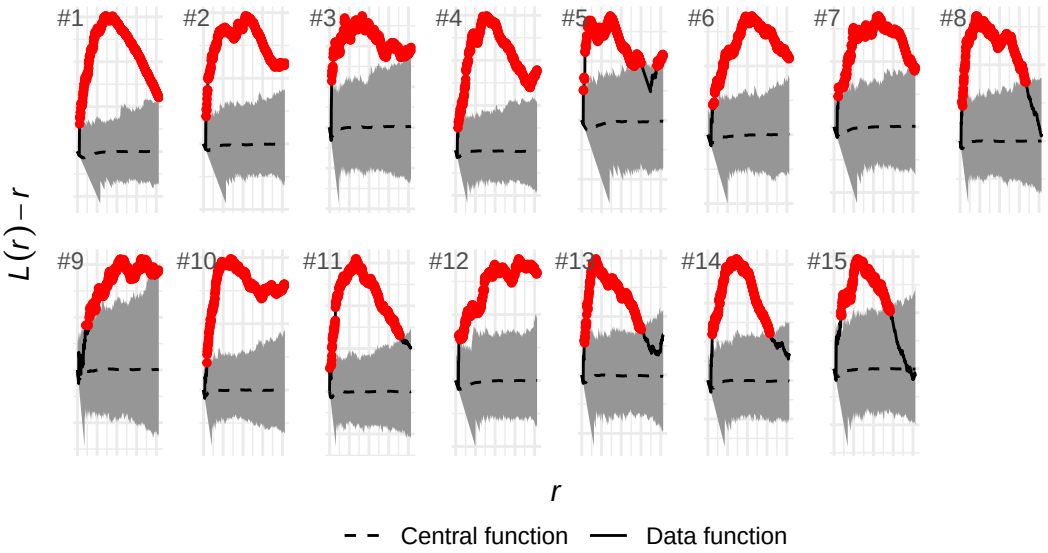
Som grund för arbetet vill vi kvantitativt påvisa att ändpunktmönstren är klustrade, vilket vi gör med ett *Monte Carlo-test*. I testet jämför vi ett punktmönster med simuleringar på punktmönster vilka är CSR. Det är endast placeringen av punkterna i mönstren vi vill jämföra. Vi får fullständigt utbytbara mönster om vi betingar antalet simulerade punkter på antalet observerade punkter [14, s. 15]. Detta medför att testet inte är sammansatt enligt avsnitt 2.4.3. I testet skapas det för varje punktmönster ett envelope av 999 simuleringar med $L(r)$ som statistika.

I `spatstat` skattas $K(r)$ vilket sedan används för att skatta $L(r)$ enligt avsnitt 2.2. Envelopet används sedan i `global_envelope_test`. Resultatet kan ses i figur 7. Det ses tydligt att alla mönster ligger till stor del ovanför CSR-området skapat av envelopet, vilket ger att samtliga punktmönster är klustrade.

3.2 Individuella parameterskattningar

Vid anpassning av parametrarna κ samt μ används `kppm` i `spatstat` vilken skattar parametrar för en vald punktprocess till ett punktmönster. Parameterskattningen utförs numeriskt med funktionen `mincontrast` i `spatstat` beskriven i avsnitt 2.4.2, vilken används i många andra funktioner och utnyttjar inbyggda metoder för olinjär optimering i `R`.

Standardparametrarna i `spatstat` för `mincontrast` i ekvation (18) är $p = 2$, $q = 1/4$, och värden på a, b bestäms utifrån observationsfönstret och $\hat{\lambda}$ (se [12]). De skattade parametrarna för punktmönstren kan ses i tabell 2 i appendix A.2, eller visualiserade med ett lådagram i figur 10. I lådagrammet finns en outlier för båda modellerna markerad med en prick. Det avvikande värdet är från punktmönster 3 vilket tillhör gruppen MODERATE.



Figur 7: CSR-(anpassnings)test för punktmönstren (95%, $N = 999$ simulationer).

3.3 Individuella anpassningstest

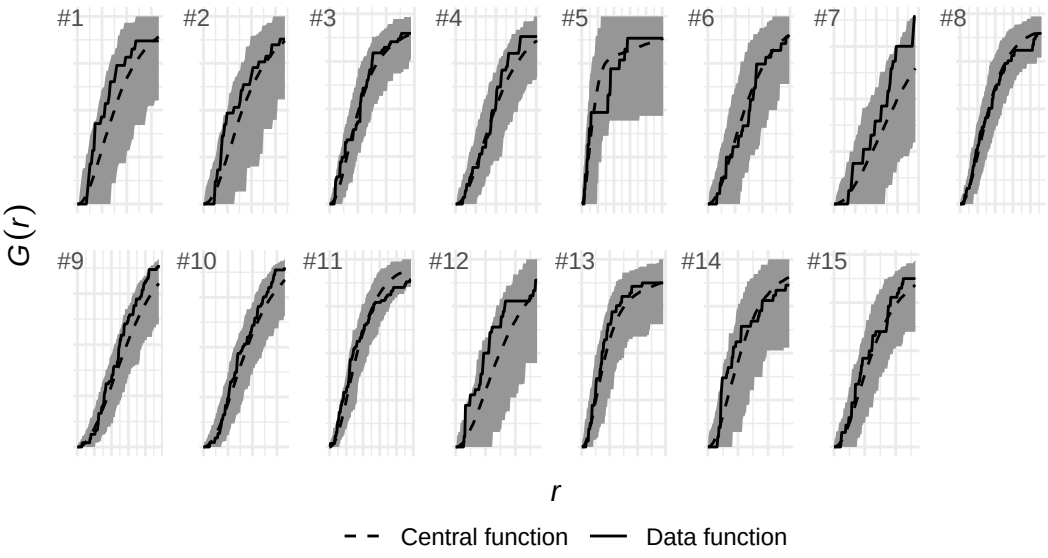
Vi genomför individuella anpassningstest för att undersöka om skattningar från föregående avsnitt är lämpliga för observerad data. Testen görs enligt avsnitt 2.4.3, med den individuella skattade modellen som nollhypotes och signifikansnivå $\alpha = 0.05$. Här används statistikan närmsta granne-funktionen eftersom vi använder K-funktionen vid anpassning av modeller med minimum contrast-metoden, och vill undvika användning av samma statistika för undersökning av modellens anpassningsgrad. Vi utför testet med hjälp av `R`-paketet `GET` [16, 17], där det korrigerade testet görs genom funktionen `multiGET.composite`. I den metoden beräknar vi vektorer $\boldsymbol{G}_i$ för statistikorna med funktionen `envelope` i `spatstat`, och använder funktionen `GET.composite` i `GET` för ”global envelope” test med sammansatt nollhypotes där vi använder `type='erl'` för att använda *extreme rank length* metoden.

Resultatet från de individuella testen visas i figur 8, där vi kan se envelopes för K-funktionen för de anpassade Thomas- och Matérnprocesserna till punktmönster 1–15. I varje test accepteras nollhypotesen om datakurvan stannar inom det gråa området för alla r. Vi kan se att alla tester visar att nollhypotesmodellen, dvs. Thomas eller Matérn, verkar vara en lämplig modell för data.

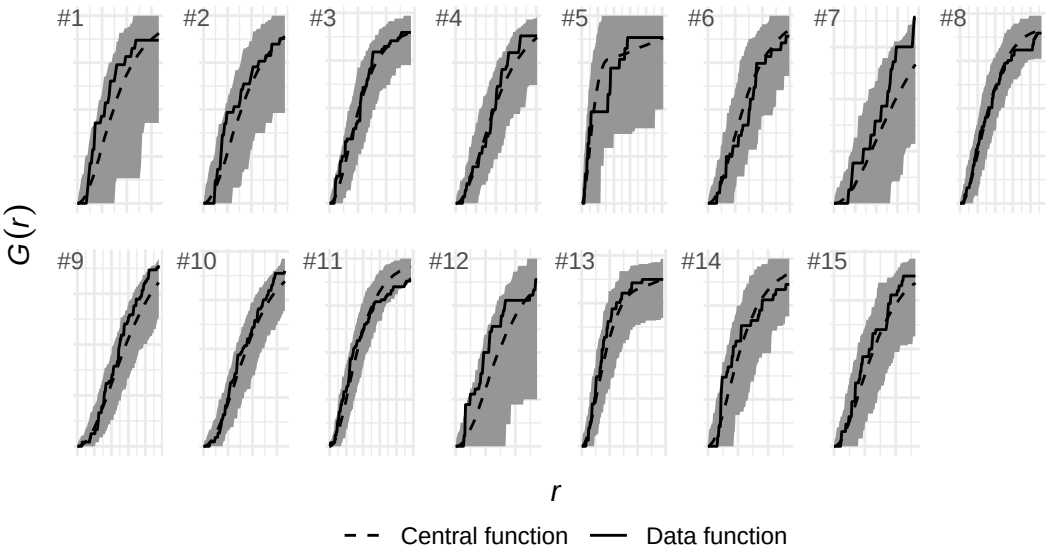
3.4 Gruppvisa parameterskattningar

Vi kan också betrakta punktmönstren i grupperna MODERATE respektive NORMAL som replikat tagna från gruppen, och på så sätt ta fram gemensamma parameterskattningar för de två processerna. Först tas $\bar{K}$-funktionen fram för vardera grupp enligt avsnitt 2.2.3 med hjälp av funktionen `pool` från `spatstat`. Därefter bestämmer vi $\bar{\lambda}$ enligt avsnitt 2.4.1 med en egen funktion `intensitybar` och skattar parametrar enligt avsnitt 2.4.2 med $\bar{K}$ som statistika. Det sistnämnda görs genom `thomas.estK` och `matclust.estK`, som använder den tidigare beskrivna `mincontrast` på samma sätt som för de individuella skattningarna. Skattningen görs alltså som i avsnitt 3.2; det är endast en annan uppsättning underliggande data med observationerna sammanförda gruppvis.

De framtagna $\bar{K}$-funktionerna visas (transformerade) i figur 9. Resultatet av skattningarna redovisas i tabell 3 i appendix A.2 och figur 10, med de teoretiska $K(r)$ från de skattade parametrarna i figur 9. I den sistnämnda figuren anmärker vi att den empiriska gruppskattningen för MODERATE går under referenslinjen för de första få mikrometrarna, och att dess maximum är högre och på större skala (större r) jämfört med NORMAL. Modellernas teoretiska funktioner är lika, men något förskjutna jämfört med den empiriska.

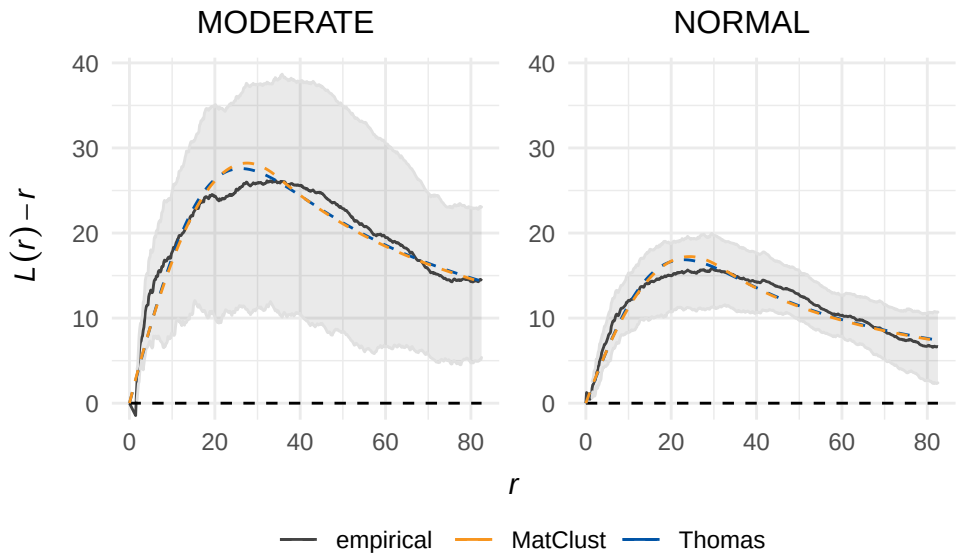


(a) Thomasprocessen.



(b) Matérnklusterprocessen.

Figur 8: Envelopes för individuella anpassningstest (95%, $N = M - 1 = 499$ simulationer).



Figur 9: Gruppvis jämförelse av $\bar{L}(r) - r$ skattat från nervmönstren och en teoretisk $L(r) - r$ med parametrar skattade från data, överlagda på ett fält som visar spridningen av $\hat{L}(r) - r$ (dvs. empirisk skattning) för alla mönster i gruppen.

3.5 Gruppvisa anpassningstest

Nästa steg är att testa den gruppvisa modellen mot alla tillgängliga mönster. För att göra detta använder vi oss av vår egna funktion `multiGET.composite` på samma sätt som i avsnitt 3.3. Denna funktion tar en mängd punktmönster och genomför sedan ett globalt envelope test på varje punktmönster. Vår nollhypotes i detta fall är den sammansatta modellen baserad på alla mönster och vi stöter igen på problemet som beskrevs i andra halvan av avsnitt 2.4.3.

Resultaten kan vi se i figur 11a och figur 11b där de röda markeringarna visar var kurvan lämnar envelopet. Som vi kan se förkastas båda modellerna för MODERATE. Det test som är närmast att förkastas transformeras tillbaka och ger de p-värden vi ser i figurerna. För Thomas får vi $p \approx 0.014$ och $p \approx 0.216$ för MODERATE respektive NORMAL. För Matérn är motsvarande värden $p \approx 0.014$ och $p \approx 0.337$.

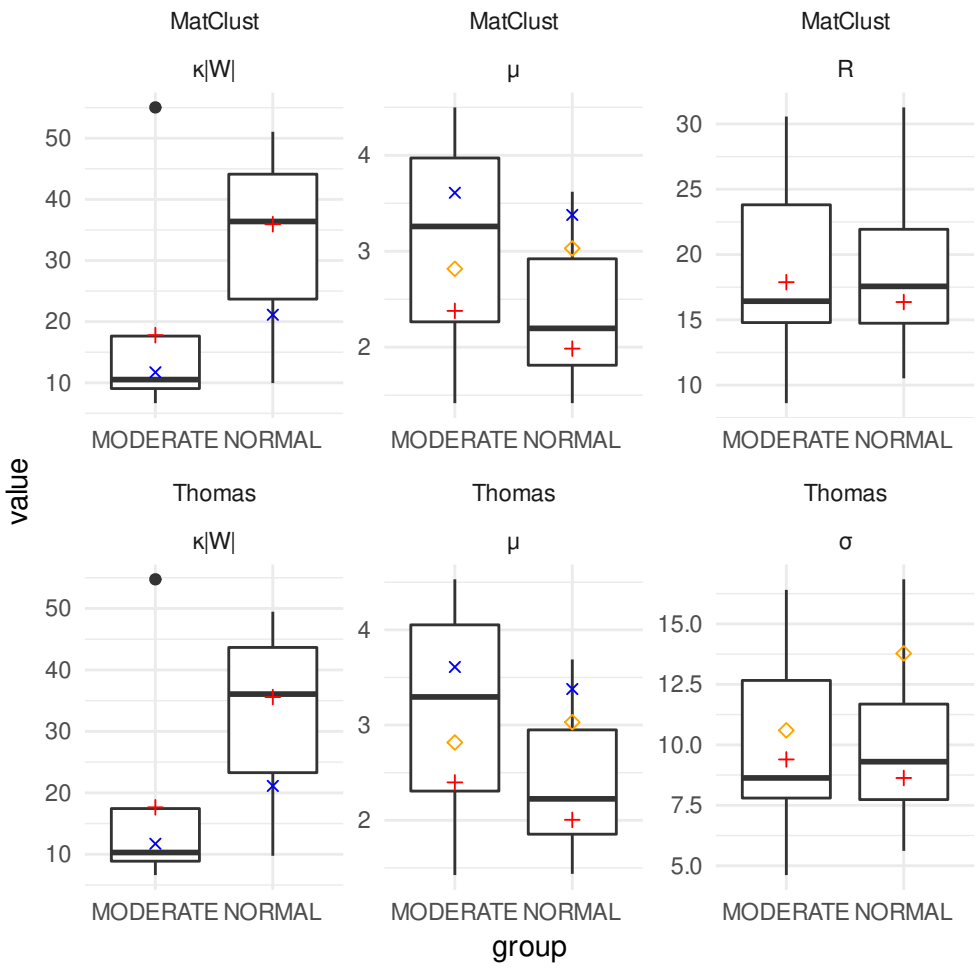
3.6 Utökad analys med förgreningspunkter och klusterdata

I den vanliga formuleringen av klusterprocesser observeras endast dotterpunkter, men utöver de mönster av ändpunkter vi studerar har vi också data för de första förgreningspunkterna och vilka ändpunkter och förgreningspunkter som tillhör vilket kluster. Vi analyserar den extra datan för att hjälpa oss tolka och utvärdera modellernas lämplighet.

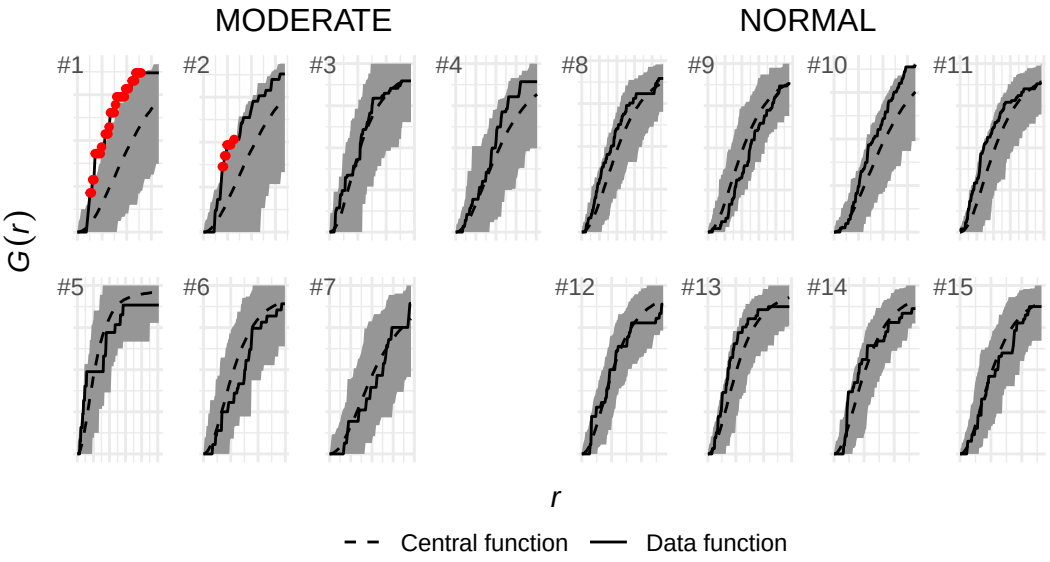
3.6.1 Utforskande analys av förgreningspunkterna

En rimlig tolkning av förgreningspunkterna är som föräldrapunkter till ändpunkterna i de två punktprocesserna. I så fall bör de enligt avsnitt 2.1 motsvara en Poissonprocess. Vi kan därför göra ett CSR-test på precis samma sätt som i avsnitt 3.1 fast för föräldrapunktsmönstren för att kontrollera detta. Resultaten visas i figur 12, där alla mönster utom #1 och #6 klarar detta test.

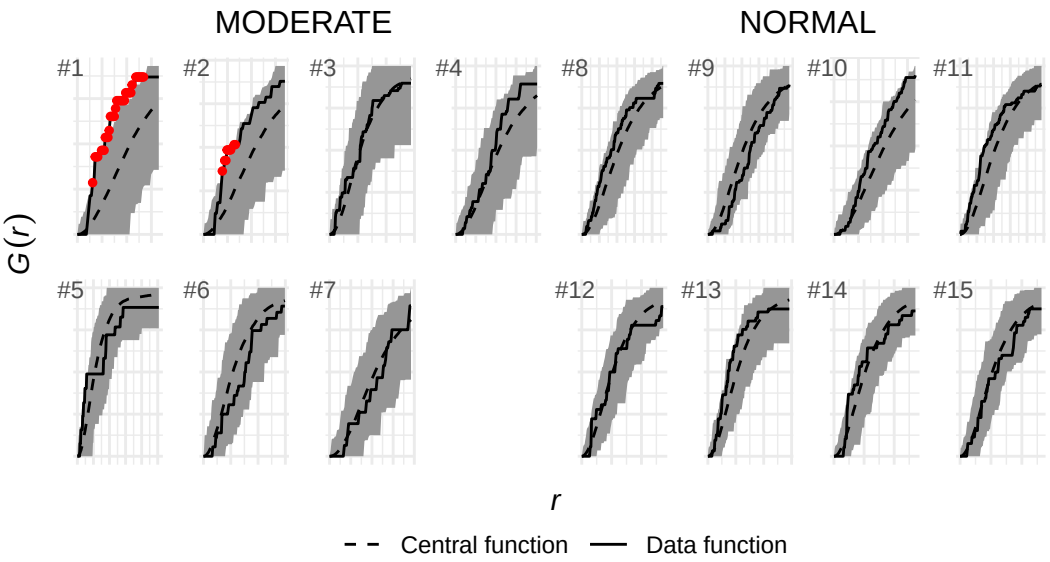
En annan intressant jämförelse är den naturliga skattningen för κ från observationerna av föräldrapunkterna, eftersom det motsvarar intensiteten för de mönstren. Skattningarna tas fram från observationerna och slås samman gruppvis enligt avsnitt 2.4.1. Av detta följer dessutom en alternativ skattning $\bar{\mu} = \bar{\lambda}/\bar{\kappa}$ enligt tidigare samband. Värden för denna listas i tabell 4 i appendix A.2. I figur 10 och 13 jämförs resultaten av de olika skattningarna gruppvis med väntevärden från de skattade modellerna. Notera att enligt modellparametrarna förväntas i de flesta fall fler kluster än vad förgreningspunkterna tyder på.



Figur 10: Lådagram över de individuella parameterskattningarna för alla nervmönster, uppdelade gruppvis, där κ skalats med observationsfönstrets area (så att värdet motsvarar antal föräldrapunkter/kluster). De överlagda korsen ($+$) är motsvarande gruppvis parameterskattning från avsnitt 3.4, kryssen ($\times$) är motsvarande skattning baserad på förgreningspunkterna från avsnitt 3.6.1, och romberna ($\diamond$) är motsvarande skattning baserat på markerade kluster enligt pseudomodellerna som introduceras i avsnitt 3.6.2. Skattningarna för R enligt pseudomodellerna har plockats bort eftersom de är väldigt stora. Värdena återfinns i tabell 2 till 5 i appendix A.2.

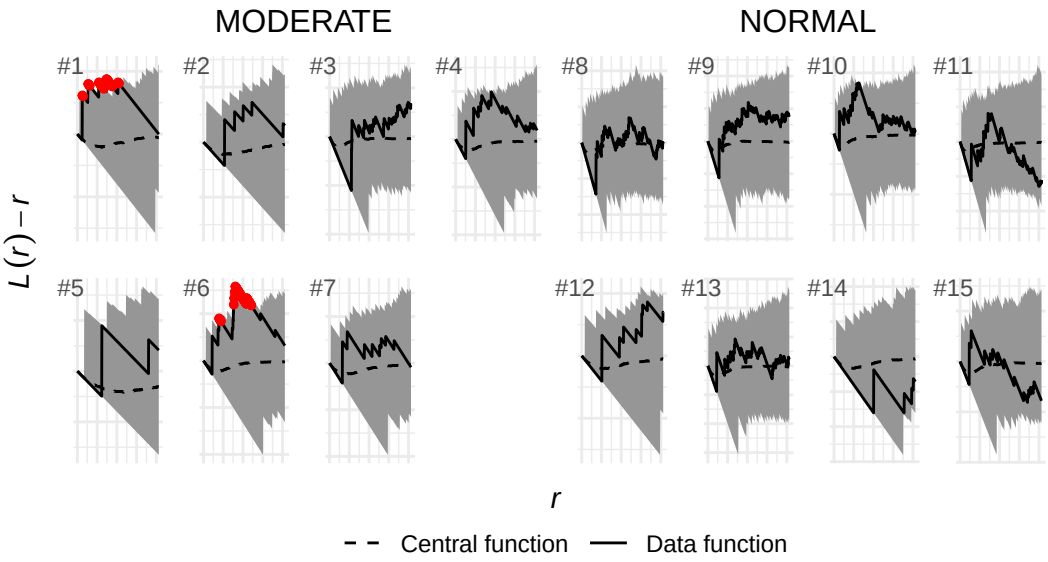


(a) Thomasprocessen. MODERATE $p \approx 0.014$, NORMAL $p \approx 0.216$.

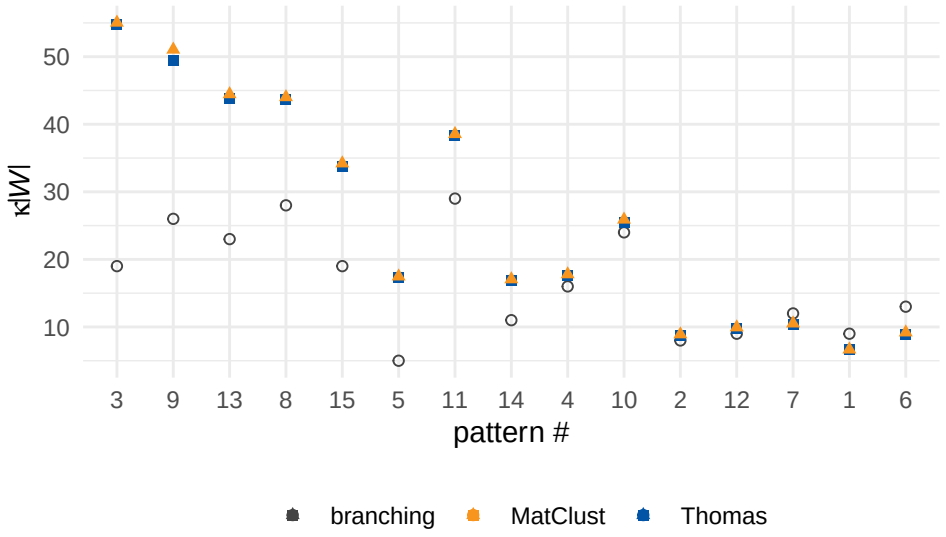


(b) Matérnklusterprocessen. MODERATE $p \approx 0.014$, NORMAL $p \approx 0.337$.

Figur 11: Envelopes för gruppvisa anpassningstest (95% gemensamt, $N = M - 1 = 499$ simulationer).



Figur 12: CSR-(anpassnings)test för förgreningspunkterna (95%, $N = 999$ simulationer).



Figur 13: Avvikelser mellan anpassad $\bar{\kappa}\,|W|$ från modellerna för ändpunktsmönstren och antalet observerade förgreningspunkter, ordnade efter (tecknad) differens.

3.6.2 Experimentell hierarkisk modell

En alternativ modell kan konstrueras med hjälp av det kompletta datasetet. Som vi ser i figur 12 klarar inte alla mönster CSR-testet. Med denna data kan vi betinga på förgreningspunkterna för att eliminera föräldrapunkternas fördelningspåverkan på resultatet samt jämföra de nya skattningarna.

Vi undersöker även med denna modell om en Matérn- eller Thomas liknande process är bäst. Det är viktigt att redan nu påpeka att vår modell inte är en Matérn- eller en Thomasprocess då den underliggande föräldraprocessen inte är en Poissonprocess. När vi här säger pseudo-Matérn menar vi hur dotterpunkterna simuleras, alltså likformigt fördelade i en cirkel med radie r centrerad på en föräldrapunkt. Med pseudo-Thomas menar vi att dotterpunkterna är normalfördelade med väntevärde 0 och standardavvikelse σ runt en föräldrapunkt.

För att uppskatta μ går vi igenom alla förgreningspunkter och räknar hur många ändpunkter som hör till dem. Om $y_1, \ldots, y_N$ är förgreningspunkterna och X_i är mängden ändpunkter för y_i får vi då

$$\hat{\mu} = \frac{\sum_{i=0}^{N} \#\{X_i\}}{N}. \tag{21}$$

För att uppskatta radien i Matérnprocessen räknar vi alla avstånd från ändpunkterna till sin respektive förgreningspunkt. Då avstånden är likformigt fördelade används det största värdet som uppskattning av övre gränsen [22, s. 286],

$$\hat{R} = \max_i \{ \max_{x_j \in X_i} \{\|x_j - y_i\|\}\}. \tag{22}$$

För att uppskatta skalan för Thomasprocessen räknar vi återigen alla avstånd. Från avsnitt 2.1.3 vet vi att det sanna väntevärdet för avstånden är 0. Med hjälp av den kunskapen räknar vi sedan ut standardavvikelsen som sedan används som uppskattning av skalan

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=0}^{N} \sum_{x_j \in X_i} \|x_j - y_i\|^2}{2 \sum_{i=0}^{N} \#\{X_i\}}} \tag{23}$$

enligt ekvation (51) i appendix A.6. För att uppskatta κ räknar vi antalet förgreningspunkter och delar det med observationsfönstrets yta vilket är samma skattning som använts tidigare,

$$\hat{\kappa} = \frac{N}{|W_i|}. \tag{24}$$

Med dessa uppskattade parametrar kan vi nu simulera modellen betingat på förgreningspunkterna. Först måste vi simulera de föräldrapunkter som kan ligga utanför observationsfönstret. Enligt avsnitt 2.1.2 får vi fönstret $W_{\oplus \hat{R}} \setminus W$ för pseudo-Matérn och enligt avsnitt 2.1.3 får vi $W_{\oplus 4\hat{\sigma}} \setminus W$ för pseudo-Thomas. Vi simulerar nu en Poissonprocess med intensitet $\hat{\kappa}$ i dessa fönster och låter dessa punkter vara ytterligare föräldrar i nästa steg. De uppskattade parametrarna ser vi som de gula romberna i figur 10 och listas i tabell 5 i appendix A.2.

För både Matérn- och Thomasprocessen är antalet dotterpunkter för en föräldrapunkt Poissonfördelat. För varje punkt i ett av förgreningspunktsmönstren tar vi ett slumptal från $\text{Poi}(\hat{\mu})$ och låter det vara antalet dotterpunkter. För pseudo-Matérnprocessen placerar vi dessa punkter i en cirkel med radie $\hat{R}$ centrerad på förgreningspunkten. För pseudo-Thomasprocessen placeras de $N(0, \hat{\sigma})$-fördelat runt förgreningspunkten.

4 Diskussion och slutsats

4.1 Thomas- eller Matérnklusterprocess

Alla punktmönster från både NORMAL och MODERATE gruppen visar tecken på klustring som vi kan se i figur 7, vilket visar att användning av klusterprocesser är rimligt. Resultat från de individuella testen indikerar att nervändpunkterna i ytterhuden kan modelleras med Thomas- och Matérnklusterprocesser. Alla anpassade modeller från både Thomas och Matérnprocesser ser ut som relevanta modeller för den individuella datan. Vi noterar dock att några datakurvor ligger

på kanten av det gråa området för vissa värden av r (se figur 8), vilket indikerar att det kan finnas problematiska mönster som inte kan modelleras så bra med dessa processer. Intensiteten av föräldrapunkter för vissa punktmönster skattas mycket högre än intensitet av förgreningspunkter som vi ser i figur 13, vilket leder till att antalet kluster kan vara överskattade. Detta kan bero på att modellerna ser vissa kluster i datan som en gruppering av flera mindre kluster.

Vi antar stationäritet eftersom observerad data är en mindre del av ett stort område, men vi ser tydlig predominans i ena delen av observationsfönstret hos vissa mönster se figur 14 i appendix. Detta kan indikera att en djupare undersökning av homogenitet hos punktmönster för nervdata kan krävas.

Resultaten från de individuella och gruppvisa anpassningarna skiljer sig knappast mellan Thomas och Matérnprocessen, även de skattade intensiteterna är väldigt lika för båda processer. Hur dotterpunkter genereras verkar alltså ha en minimal betydelse för resultaten och det är svårt att upptäcka skillnader mellan de simulerade punktmönstren med hjälp av de sammanfattningsfunktionerna som användes i detta arbete.

4.2 Skillnader mellan grupperna

Enligt vår hypotes ska de mönstren som kommer från MODERATE vara mer klustrade än de som kommer från NORMAL. Som vi ser till vänster i figur 10 verkar MODERATE ha färre nervträd. Vidare ser vi dock att MODERATE har fler nervtrådar per träd. Det genomsnittliga antalet nervändpunkter hos ett MODERATE mönster är 42,29. För ett NORMAL mönster är det motsvarande antalet 71,39. Vi ser alltså att det finns en stor skillnad i antal nerver mellan individer som lider av diabetesneuropati och friska individer.

Det är intressant att se att μ är större för MODERATE än friska. En möjlig orsak till detta är att de kvarstående nervträden blir tvungna att kompensera genom att öka antalet ändpunkter. Till höger i figur 10 ser vi även att skalan för båda modellerna är väldigt lika.

Vi kan se i figur 9 att den empiriska skattningen av $\hat{L}(r) - r$ för MODERATE gruppen har maximum vid större r än vad NORMAL gruppen har vilket stämmer överens med vad vi ser i figur 10. För MODERATE antar grafen maximum vid $r \approx 34{,}32\,\mu\text{m}$ och för NORMAL vid $r \approx 29{,}49\,\mu\text{m}$. För MODERATE är det maximala värdet ca. 26,15 medan det är ca. 15,84 för NORMAL. Vi ser att MODERATE gruppen antar högre värden för nästan alla r. Detta indikerar att den sjuka gruppen är mer klustrad än den friska och stämmer överens med vår hypotes.

Intressant nog ser vi att grafen för MODERATE gruppen runt $r \approx 0$ går under 0 vilket skulle kunna indikera regularitet eller inhibition av någon sort. Det är möjligt att nervändpunkterna försöker täcka så mycket hudarea som möjligt vilket resulterar i att de inte ligger för nära varandra. Dock ser vi inte samma beteende för NORMAL gruppen. Hade det varit så att nervändpunkterna inte ligger för nära varandra på grund av att de försöker täcka så mycket area som möjligt borde vi observera samma beteende för NORMAL gruppen. Det skulle kunna vara av intresse att undersöka detta mer noggrant i framtiden.

4.3 Förkastning av de gruppgemensamma modellerna för sjuka

De gruppgemensamma modellerna skapade med Thomas- och Matérnklusterprocessen för datan i den friska gruppen verkar passa bra för alla punktmönster, men modellerna förkastas för punktmönstren från de sjuka då de inte passar för mönstren #1 och #2. Att de gruppgemensamma modellerna inte är tillräckligt bra kan bero på flera bidragande orsaker, vilka kommer diskuteras nedan.

Klassifikationen ”måttliga symptom” är en kvalitativ bedömning i vilken en varierad grad av neuropati innefattas. Detta gör att MODERATE i grunden har större varians än NORMAL vilken är mer homogen och inte har samma problem. Skillnaden i varianserna ses tydligt i figur 9, där spridningen av $\hat{L}(r) - r$ är mycket större för gruppen MODERATE än för NORMAL.

Då vissa punktmönster har färre punkter leder det till att en lägre vikt ges för deras egenskaper i den gemensamma modellen. I ett vidare arbete skulle det kunna undersökas om man kan förbättra modellen genom att använda en annan viktning för att beakta egenskaper från mönstren med få punkter till en annan grad än innan.

Vi har också fått data för förgreningspunkter vilket är punkter på första stället där nerven delar sig. Från detta stället har inte ändpunkterna så stor utbredning och förgreningspunkterna

ses som föräldrapunkter, se figur 14 i appendix A.2. I ett steg i analysen utfördes ett CSR-test på förgreningspunkterna. Anledningen till att det är intressant att studera föräldrapunkternas fördelning är för att både Thomas- och Matérnklusterprocessen bygger på att föräldrapunkterna är CSR. I figur 12 ser vi att mönster #1 inte klarar CSR-testet, vilket är ett av de i MODERATE som också misslyckas för gruppmodellen. Dock är förgreningspunkterna för mönster #6 inte heller CSR, men passar för gruppmodellen. Mönster #2 som inte passar för gruppmodellen har förgreningspunkter som är CSR.

Vid fallet att föräldrapunkter är klustrade skulle vara ett argument mot användandet av dessa punktprocesser då de har föräldrapunkter som är CSR. Alternativt att förgreningspunkterna inte kan ses som föräldrapunkter. Dock ser vi att punktmönster #6 beskrivs av gruppgemensamma modellen. Så icke CSR förgreningspunkter behöver inte påverka modellens förmåga att beskriva ändpunkterna. Däremot påverkas den biologiska förklaringsförmågan av den kompletta trädstrukturen. Viktigt att betona är att datan är för begränsad för att sådana slutsatser skulle kunna göras.

4.4 Experimentell hierarkisk modell

Från figur 12 vet vi att inte alla mönster av förgreningspunkter går att beskriva som en Poissonprocess. För att undvika att behöva göra antaganden om den underliggande processen kan vi istället bygga en hierarkisk modell. Med fullständig data får vi även alternativa skattningar på parametrarna.

Det första problemet med denna modell är att uppskattningen av μ inte tar hänsyn till kanteffekter. Det är möjligt att en förgreningspunkt ligger tillräckligt nära kanten av fönstret att en eller fler av dotterpunkterna hamnar utanför det. Detta leder till att vår uppskattning av μ kan vara konservativ. För att åtgärda detta behöver vi ignorera de förgreningspunkter som ligger för nära kanten i vår uppskattning, alltså vårt observationsfönster måste minskas till $W_{\ominus \hat{R}}$ för pseudo-Matérn och till $W_{\ominus 4\hat{\sigma}}$ för pseudo-Thomas. Ett av problemen för pseudo-Thomasprocessen är att vi aldrig kan garantera att en eller fler dotterpunkter fortfarande hamnar utanför observationsfönstret. Man kan krympa observationsfönstret ännu mer för att minska denna möjlighet men 4σ är standardvärdet för Thomasprocessen i `spatstat`. Även uppskattningarna av både R och σ är konservativa. Det är möjligt att det finns dotterpunkter utanför observationsfönstret vilket betyder att både $\hat{R} \leq R$ och $\hat{\sigma} \leq \sigma$. En uppskattning av dessa två parametrar som inte har systematiskt fel beror på de sanna värdena. Det är självklart möjligt att först uppskatta R och σ för att sedan använda dessa värden för att få en bättre uppskattning, men det systematiska felet kvarstår. Ytterligare en möjlighet hade varit att iterativt fortsätta uppskatta R och σ med hjälp av den föregående iterationen tills ett stationärt värde hittades.

Det är också möjligt att det finns nervtrådar som går direkt från baspunkten upp till överhuden. Det finns då inget sätt att skilja dessa punkter från punkter som kommer från en förgreningspunkt utanför observationsfönstret. Då dessa punkter inte kan skiljas från de andra kan de inte simuleras. Det sista problemet är att för de mönstren där CSR förkastas är det inte korrekt att simulera en Poissonprocess för föräldrar utanför observationsfönstret.

4.5 Generaliserbarhet och framtida arbete

Individerna i studien är inte ett stort slumpmässigt stickprov från en viss population, och därför kan resultaten inte direkt generaliseras till grupper eller sjukdomen i sin helhet. Ska våra resultat generaliseras vidare och användas för att kunna klassificera nervmönster till friska och sjuka, vilket kanske kan anses som slutmålet med forskningen om detta, så krävs det mycket mer data. Med små datamängder som vårat dataset är det omöjligt att validera prediktioner och undersöka metodens tillförlitlighet. Vi understryker dock att vårt syfte inte är att konstruera en klassificerare, utan att jämföra modellerna och analysera vad skattningarna säger om de mönster vi har i datasetet som en grund till framtida arbete. Vi eftersträvar att förbättra de analyser som tidigare har gjorts, t.ex. genom att testa gruppmodeller och kompensera för effekten av den sammansatta nollhypotesen i testen.

Framtida arbete med modellerna behöver inte bara innefatta arbete mot en diagnostisk slutprodukt. Det finns också ett antal intressanta aspekter i analysen som visar på möjliga utökningar av modellerna för att bättre beskriva nervmönstren, även om det finns ett värde i sig att konstruera den enklaste möjliga lämpliga modellen. I figur 9 ser vi att (den empiriska) skattningen av $L(r) - r$

för MODERATE är negativ för några få mikrometer i början, ett fenomen som inte ryms i våra modeller i nuläget. Detta indikerar en viss regularitet för de sjuka mönstren på ett mycket kort avstånd. Är det en artefakt från själva analysen eller borde modellen ta hänsyn till detta genom någon mekanism? En annan aspekt är den betydande variationen mellan individer i grupperna, i synnerhet för de sjuka mönstren, som tidigare anfördes som möjligt skäl till varför en gruppgemensam modell för MODERATE misslyckades. Genom att explicit introducera slumpmässiga individuella avvikelser från en gruppgemensam grundmodell skulle det vara möjligt att öka variationen som ryms i modellen, och på så sätt försöka beskriva individvariationen för att eventuellt studera den vidare. Ytterligare ett exempel är diskussionen om eventuella inhomogeniteter i föregående avsnitt; för att introducera den variationen i modellen krävs det en mer generell process än vad Thomas- eller Matérnklusterprocessen kan beskriva, vilket också är en utökning som kan studeras vidare.

Ett annat spår som kan utforskas vidare är huruvida en helt annan typ av modell skulle vara mer passande för nervmönstren än Thomas- och/eller Matérnklusterprocessen. Vi har fokuserat mycket på dessa på grund av klustringen som observeras i nervmönstren, och anpassningen är tillräcklig för vissa delmängder i datasetet, men många av de små tveksamheterna som behandlats kanske skulle kunna avhjälpas genom en annan modell. Dessutom fås i praktiken mycket mer data från biopsierna än bara nervändarna. Den mest uppenbara utökningen vore att på något sätt ta hänsyn till förgreningspunkterna (som, om de tolkas som föräldrapunkter, inte anses eller kan observeras i klusterprocesserna vi använder). Mer komplicerat vore en modell som tar hänsyn till trädstrukturen i sin helhet, som t.ex. den hierarkiska modellen studerad i avsnitt 3.6.2.

4.6 Slutsats

Våra resultat följer flertalet tidigare resultat från forskning som beskrevs i inledningen. För alla ändpunktsmönster i datasetet förkastades CSR till förmån för klustring, vilket väntades. Modellparametrarna som skattas från punktmönstren tyder på en generellt högre grad av klustring hos sjuka och en annan skala av klustring. En icke-intuitiv observation är att det generellt tycks finnas fler ändpunkter i varje kluster hos sjuka än hos friska. Notera dock att resultaten inte direkt kan generaliseras; däremot kan de ge förslag på nya hypoteser att utforska.

Individuella skattningar för alla ändpunktsmönster i båda modellerna accepteras av anpassningstest, vilket är i linje med resultat i tidigare arbeten. Anpassningen är förmodligen sämre än vad som tidigare troddes eftersom vi här också tar hänsyn till den sammansatta nollhypotesen, vilken tidigare hade försummats. Resultaten är emellertid inte direkt jämförbara eftersom vår analys använder ett annat dataset, så det går inte att utesluta att skillnaden hör till variation mellan dataseten. Våra nya gruppgemensamma modeller tycks utifrån anpassningstest beskriva de friska mönstren väl, men inte de sjuka mönstren. Det faktum att en gruppgemensam modell för sjuka misslyckas skulle kunna bero på att det är en mycket större variation mellan mönstren i den gruppen, eller på problem med själva modellen. Detta omöjliggör inte framtida arbete med klassifikation, eftersom sjuka mönster skulle kunna identifieras som de som inte följer en frisk gruppmodell.

Vi hittade inte någon väsentlig skillnad i prestanda mellan Thomas- och Matérnklusterprocessen. Små skillnader i parametervärden eller i anpassningsgraden kan inte med säkerhet hänföras till någon praktisk skillnad modellerna emellan. Valet mellan de två modellerna för nervdatan blir därför snarare en subjektiv fråga baserad på vilken modell som bättre motsvarar biologisk kunskap om hur nervstrukturen fungerar. Lyckade anpassningar till ändpunktsmönstren tyder på att modellerna klarar av att beskriva den spatiala strukturen i ändpunktsmönstren, men deras biologiska förklaringsförmåga kan vara begränsad eftersom de observerade förgreningspunkterna för några mönster inte klarar ett CSR-test, vilket föräldrapunkterna i modellen följer. Datan lyfter också andra osäkerheter kring modellernas antaganden, t.ex. om det kan finnas inhomogeniteter i mönstren. Det är möjligt att en bättre modell kan konstrueras med den mer detaljerade data en biopsi ger jämfört med en traditionell observation i en klusterprocess, något som också kan vara föremål för framtida forskning.

Referenser

1. International Diabetes Federation. *IDF Diabetes Atlas* 9. utg. (International Diabetes Federation, Bryssel, 2019).
2. *Nervcell - Neuron - Människans nervsystem: Uppslagsverk* `http://nervsystemet.se/nsd/structure_226`.
3. *Nervtråd - Axon - Människans nervsystem: Uppslagsverk* `http://nervsystemet.se/nsd/structure_58`.
4. *Myelin - Myelin - Människans nervsystem: Uppslagsverk* `http://nervsystemet.se/nsd/structure_219`.
5. Adamson, U. *Diabetesneuropati* NetdoktorPro.se. 4 nov. 2013. `https://www.netdoktorpro.se/diabetes/medicinska-oversikter/diabetesneuropati/`.
6. 1177 Vårdguiden. *Polyneuropati, en sjukdom i flera nerver* 10 sept. 2019. `https://www.1177.se/sjukdomar--besvar/hjarna-och-nerver/nerver/polyneuropati-en-sjukdom-i-flera-nerver/`.
7. Sjöholm, Å. *Diabetesneuropati* Internetmedicin. 18 april 2020. `https://www.internetmedicin.se/behandlingsoversikter/diabetes/diabetesneuropati/`.
8. Waller, L. A. *m. fl.* Second-order spatial analysis of epidermal nerve fibers. *Statistics in Medicine* **30,** 2827–2841 (15 okt. 2011).
9. Kennedy, W. R., Wendelschafer-Crabb, G. & Johnson, T. Quantitation of epidermal nerves in diabetic neuropathy. *Neurology* **47,** 1042–1048 (1 okt. 1996).
10. Eriksson, S., Källgren, C. & Lane, H. *Modellering av nervmönster med spatiala punktprocesser* Kandidatarbete (Chalmers tekniska högskola/Göteborgs universitet, 24 juni 2019).
11. Ekelund, H. *m. fl. Parameterskattning av spatiala klusterprocesser med en inblick i nervdata* Kandidatarbete (Chalmers tekniska högskola/Göteborgs universitet, 1 juli 2020).
12. Baddeley, A., Rubak, E. & Turner, R. *Spatial point patterns: methodology and applications with R* (Chapman och Hall/CRC Press, London, 2016).
13. Illian, J., Penttinen, A., Stoyan, H. & Stoyan, D. *Statistical Analysis and Modelling of Spatial Point Patterns* (John Wiley & Sons, Ltd, Chichester, 15 jan. 2007).
14. Diggle, P. J. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns* 3. utg. (Chapman och Hall/CRC, New York, 2013).
15. Møller, J. & Waagepetersen, R. P. *Statistical inference and simulation for spatial point processes* (Chapman & Hall/CRC, Boca Raton, 2004).
16. Myllymäki, M. & Mrkvička, T. GET: Global envelopes in R. arXiv: `1911.06583 [stat.ME]` (23 juli 2020).
17. Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H. & Hahn, U. Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79,** 381–404 (mars 2017).
18. Baddeley, A. *Simulation-based hypothesis testing on spatial point pattern hyperframes using "envelope" function in spatstat* Stack Overflow. 4 juli 2016. `https://stackoverflow.com/a/38176292`.
19. Šidák, Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association* **62,** 626–633 (juni 1967).
20. Andersson, C., Guttorp, P. & Särkkä, A. Discovering early diabetic neuropathy from epidermal nerve fiber patterns. *Statistics in Medicine* **35,** 4427–4442 (30 okt. 2016).
21. R Core Team. *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria, 2020).
22. Johnson, N. L., Kotz, S. & Balakrishnan, N. *Continuous univariate distributions* 2nd ed (Wiley, New York, 1994).

23. Stoyan, D. & Stoyan, H. *Fractals, random shapes, and point fields: methods of geometrical statistics* (John Wiley & Sons, Ltd, Chichester, 1994).

24. *Statistical distributions: Catherine Forbes ... [et al.]* 4th ed (utg. Forbes, C. S.) (Wiley, Hoboken, N.J, 2011).

A Appendix

A.1 Kod

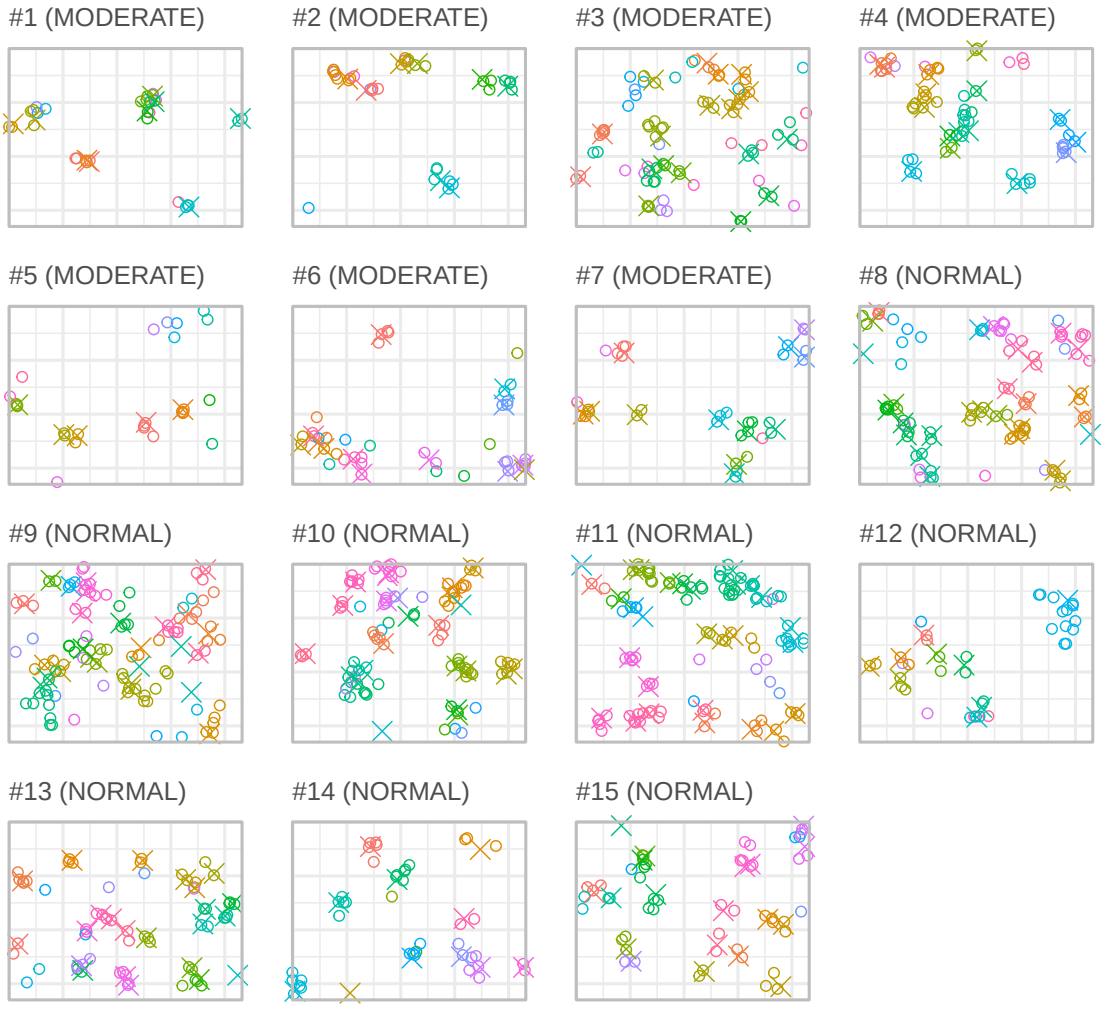
En komplett listning av all kod för analysen skulle ta mycket plats. För att göra analysen fullständigt reproducerbar är `R`-koden samt resultatfilerna bifogade som en fil i den digitala PDF-filen.

mvex01-21-01.zip

Det krävs stöd för PDF 1.3 i din läsare, vilket motsvarar en version av Acrobat från detta millennium. Däremot tillåter inte nya versioner av just Adobes läsare att öppna bifogade arkiv. En online-kopia återfinns på `https://github.com/MVEX01-21-01/kod`.

A.2 Kompletterande tabeller och figurer

Flertalet figurer och tabeller ryms ej i huvuddelen, men är fortfarande av intresse för en fullständig bild av data och resultaten och finns därför här i ett appendix.



Figur 14: Visualisering av alla punktmönster i datasetet, med cirklar för ändpunkter, kryss för förgreningspunkter och färger för kluster. Mönster #1–#7 är måttligt sjuka (MODERATE) och #8–#15 är friska (NORMAL).

Tabell 1: Egenskaper för punktmönster (av ändpunkter) i datasetet: grupp, antal punkter n, observationsfönster W i rådatans koordinatsystem, fönstrets area $|W|$.

#	Grupp	n	W [µm]	$\lvert W\rvert$ [µm^2]
1	MODERATE	30	$[-0{,}1; 432] \times [-330; 0]$	142 590
2	MODERATE	33	$[0; 432] \times [-330; 0{,}3]$	142 690
3	MODERATE	78	$[-0{,}1; 432] \times [-330; 0{,}2]$	142 680
4	MODERATE	58	$[0; 432] \times [-330; 0{,}3]$	142 690
5	MODERATE	26	$[0; 432] \times [-330; 0]$	142 560
6	MODERATE	39	$[0; 432] \times [-330; 0]$	142 560
7	MODERATE	32	$[0; 432] \times [-330; 0]$	142 560
8	NORMAL	85	$[-1; 432] \times [-330; 0]$	142 890
9	NORMAL	97	$[-0{,}35; 432] \times [-330; 0]$	142 680
10	NORMAL	89	$[-0{,}3; 432] \times [-330; 0]$	142 660
11	NORMAL	106	$[-1; 432] \times [-330; 0{,}6]$	143 150
12	NORMAL	36	$[0; 432] \times [-330; 0]$	142 560
13	NORMAL	63	$[0; 432] \times [-330; 0]$	142 560
14	NORMAL	42	$[-1; 432] \times [-330; 0{,}1]$	142 930
15	NORMAL	53	$[0; 432] \times [-330; 0]$	142 560

Tabell 2: Individuella parameterskattningar.

#	Thomas			Matérn		
	κ	μ	σ	κ	μ	R
1	$4{,}6437\times10^{-5}$	4,5306	8,6307	$4{,}6761\times10^{-5}$	4,4992	16,425
2	$6{,}1726\times10^{-5}$	3,7467	8,4947	$6{,}2464\times10^{-5}$	3,7025	16,005
3	$3{,}8369\times10^{-4}$	1,4248	7,1037	$3{,}8578\times10^{-4}$	1,4171	13,573
4	$1{,}2335\times10^{-4}$	3,2954	10,051	$1{,}2474\times10^{-4}$	3,2586	19,008
5	$1{,}2124\times10^{-4}$	1,5043	4,6154	$1{,}2262\times10^{-4}$	1,4873	8,6178
6	$6{,}2775\times10^{-5}$	4,3579	16,406	$6{,}4503\times10^{-5}$	4,2412	30,567
7	$7{,}2198\times10^{-5}$	3,1090	15,274	$7{,}3800\times10^{-5}$	3,0415	28,608
8	$3{,}0516\times10^{-4}$	1,9493	6,4289	$3{,}0801\times10^{-4}$	1,9313	12,192
9	$3{,}4674\times10^{-4}$	1,9607	16,843	$3{,}5783\times10^{-4}$	1,8999	31,267
10	$1{,}7820\times10^{-4}$	3,5010	10,046	$1{,}8152\times10^{-4}$	3,4369	18,828
11	$2{,}6774\times10^{-4}$	2,7657	8,1715	$2{,}6932\times10^{-4}$	2,7494	15,586
12	$6{,}8455\times10^{-5}$	3,6889	16,599	$6{,}9756\times10^{-5}$	3,6201	31,244
13	$3{,}0738\times10^{-4}$	1,4377	5,6157	$3{,}1180\times10^{-4}$	1,4173	10,521
14	$1{,}1805\times10^{-4}$	2,4891	9,3083	$1{,}1936\times10^{-4}$	2,4618	17,580
15	$2{,}3714\times10^{-4}$	1,5677	9,2968	$2{,}4000\times10^{-4}$	1,5490	17,545

Tabell 3: Gruppvisa parameterskattningar (+ i figur 10).

Grupp	Thomas			Matérn		
	κ	μ	σ	κ	μ	R
MODERATE	$1{,}2365\times10^{-4}$	2,3978	9,3955	$1{,}2465\times10^{-4}$	2,3787	17,881
NORMAL	$2{,}4938\times10^{-4}$	2,0050	8,6293	$2{,}5184\times10^{-4}$	1,9854	16,361

Tabell 4: Gruppvisa parameterskattningar från förgreningspunkter (× i figur 10).

Grupp	κ	μ
MODERATE	$8{,}2137\times10^{-5}$	3,6098
NORMAL	$1{,}4799\times10^{-4}$	3,3787

Tabell 5: Gruppvisa parameterskattningar från pseudomodell i avsnitt 3.6.2 (◇ i figur 10).

Grupp	μ	σ	R
MODERATE	2,8171	10,599	68,613
NORMAL	3,0296	13,780	81,109

A.3 Parkorrelationsfunktionen $g(r)$

K-funktionen beskriver den genomsnittliga mängden punkter som ligger på en skiva med radie r centrerad på en godtycklig punkt av processen, vilket står i kontrast med parkorrelationsfunktionen som beskriver den genomsnittliga mängden punkter med avstånd r från en godtycklig punkt av processen.

Parkorrelationsfunktionen definieras [13, s. 219] som

$$g(r) = \frac{K'(r)}{2\pi r} \quad r \geq 0. \tag{25}$$

Paret $K(r)$ och $g(r)$ kan liknas vid vanliga fördelnings- samt täthetsfunktioner [13, s. 219], alltså som

$$f(x) = F'(x), \quad F(x) = \int_{-\infty}^{x} f(t)\, dt. \tag{26}$$

Från (25) kan vi härleda

$$K(r) = 2\pi \int_{-\infty}^{r} xg(x)\, dx = 2\pi \int_{0}^{r} xg(x) dx \tag{27}$$

vilket vi ser liknar en fördelningsfunktion medan (25) liknar en täthetsfunktion.

I vanliga fall är korrelationsfunktioner normaliserade för att anta värden mellan -1 och 1 där 0 betyder att det inte finns någon korrelation, men detta är inte fallet för $g(r)$. Parkorrelationsfunktionen kan i stället anta värden mellan 0 och oändligheten och ett värde på 1 visar att det inte finns korrelation. Värden större än 1 visar på förekomst av klustring av punkter medan värden mindre än 1 visar på regelbundna avstånd mellan punkter [12, s. 226-227], vilket kan observeras i figur 15.

K-funktionen kan vara svårtolkad då den är kumulativ vilket gör det svårt att tolka funktionen för exakta värden. Denna skillnad kan vi se när vi jämför figur 15 med figur 16. Parkorrelationsfunktionen rekommenderas som den bästa, mest informativa statistikan vid undersökning av punktmönster [13, s. 218]. Funktionen presenterar samma information som K- och L-funktionen men är lättare att tolka. Exempelvis kan vi i figur 15 (höger) se en topp runt $r = 0.1$ vilket visar på att det finns ett stort antal punkter på detta avståndet från varandra. Om vi tittar till höger i figur 16 är denna topp inte alls lika tydlig.

I vissa fall finns det ett avstånd r_0 med

$$g(r) = 0 \quad r \leq r_0 \tag{28}$$

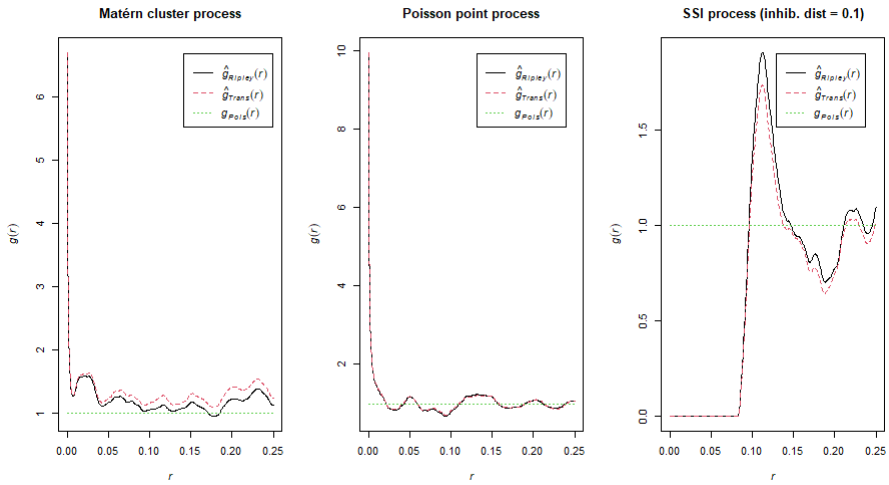
vilket kan ses i den högra plotten i figur 15. Vi kan se att $r_0 \approx 0.1$ vilket stämmer överens med inhibitionsavståndet för processen. Dessa avstånd kallas ”hard-core” avstånd och innebär helt enkelt att det inte kan finnas en granne med ett avstånd mindre än r_0.

Parkorrelationsfunktionen kan skrivas som

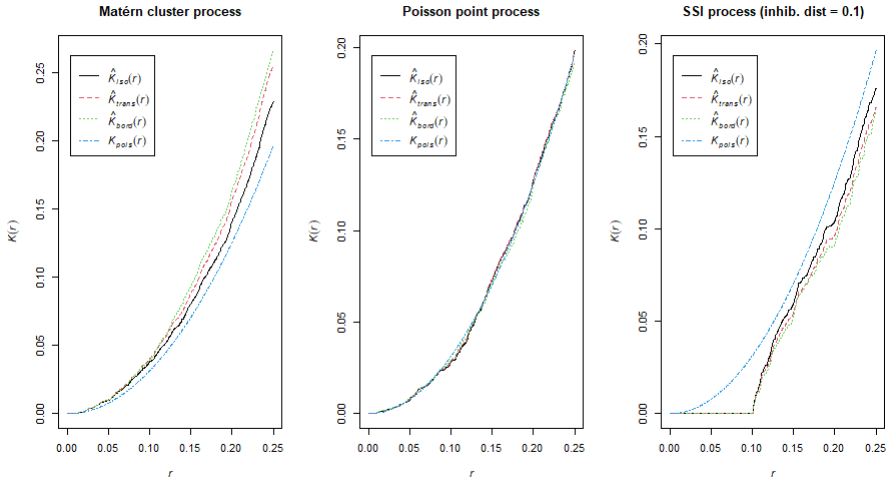
$$g(x, y) = \frac{p_2(x, y)}{p(x)p(y)} \tag{29}$$

där $p(x)$ är sannolikheten att observera en punkt av processen i en oändligt liten omgivning runt x och $p_2(x, y)$ är sannolikheten att det finns en punkt i båda de oändligt små omgivningarna runt punkterna x och y. Om den oändligt lilla omgivningen runt x har yta dx är sannolikheten att observera en punkt i denna volym

$$p(x) = \lambda \mathrm{d}x. \tag{30}$$



Figur 15: Uppskattade PKF för några punktprocesser. Klusterprocess (vänster), CSR (mitten) och reguljär process (höger).



Figur 16: Uppskattade K-funktioner för några punktprocesser. Klusterprocess (vänster), CSR (mitten) och reguljär process (höger)

Om processen är isotropisk beror sannolikheten endast på avståndet r mellan x och y vilket ger

$$g(r) = \frac{p_2(r)}{\lambda \mathrm{d}x \cdot \lambda \mathrm{d}y}. \tag{31}$$

För en Poissonprocess är händelsen att observera en punkt i båda omgivningarna runt x och y oberoende vilket betyder att

$$p_2(r) = \lambda \mathrm{d}x \cdot \lambda \mathrm{d}y \implies g(r) = \frac{\lambda \mathrm{d}x \cdot \lambda \mathrm{d}y}{\lambda \mathrm{d}x \cdot \lambda \mathrm{d}y} = 1 \tag{32}$$

vilket diskuterades ovan.

A.3.1 Skattning av parkorrelationsfunktionen

Parkorrelationsfunktionen uppskattas ofta med hjälp av kärnutjämning. Den uppskattade funktionen kan skrivas som

$$\hat{g}(r) = \frac{|W|}{2\pi r n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} k_h(r - d_{ij}) e_{ij}(r) \tag{33}$$

där $e_{ij}(r)$ är en kantkorrigering, $|W|$ är ytan av observationsfönstret W, $k_h(t)$ är utjämningskärnan med utjämningsbredd $h > 0$ och $k(x)$ som är en valfri sannolikhetsfunktion med medelvärde 0. Ofta väljer man $k_h(x)$ som Epanechnikovkärnan [12, s. 228, 23, s. 285] vilken är

$$\epsilon_h(x) = \begin{cases} \frac{3}{4h}\left(1 - \frac{x^2}{h^2}\right) & x \in [-h, h], \\ 0 & \text{annars} \end{cases} \tag{34}$$

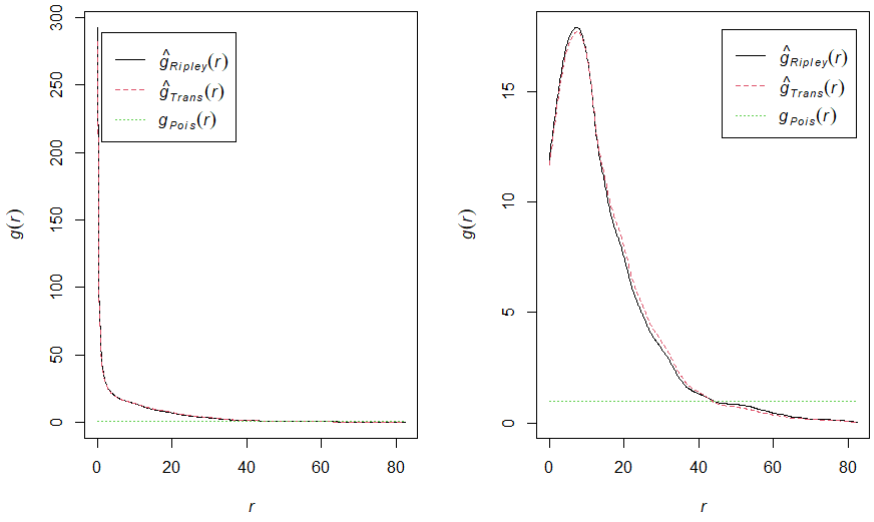
med

$$h = c\lambda^{-\frac{1}{2}} \quad \text{och} \quad c \in [0.1, 0.2]. \tag{35}$$

Som vi ser i figur (15) är uppskattningen runt $r \approx 0$ inte så exakt. Värdet på $\hat{g}(r)$ har till och med en oändlig asymptot runt $r = 0$ på grund av att man delar med r i (33). En alternativ uppskattning av $g(r)$ som inte har detta problem ges av [12, s. 229]

$$\hat{g}(r) = \frac{1}{2\pi} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{k_h(r - d_{ij})}{d_{ij}} e_{ij}(r). \tag{36}$$

En jämförelse av dessa två uppskattningar ses i figur 17.



Figur 17: Uppskattade parkorrelationsfunktioner för en individ med moderat diabetesneuropati. Uppskattning med r i nämnaren (vänster), alternativ uppskattning (höger).

A.4 Tomrumsfunktionen $F(r)$

Tomrumsfunktionen $F(r)$ beskriver sannolikheten att avståndet från en godtycklig punkt till närmsta punkt i en stationär punktprocess $\mathbf{X}$ är mindre än eller lika med ett godtyckligt avstånd r. Den är väldigt lik närmsta granne-funktionen (10) och i vissa fall är de två helt lika. Tomrumsfunktionen definieras [12, s. 262] som

$$F(r) = \mathbb{P}\{d(u, \mathbf{X}) \leq r\}. \tag{37}$$

Om $\mathbf{X}$ är en Poissonprocess är

$$F_{\text{pois}}(r) = 1 - \exp\left(-\lambda\pi r^2\right). \tag{38}$$

Uppskattade $\hat{F}(r)$ plottas ofta mot denna funktion som referens.

A.4.1 Skattning av $F(r)$

Uppskattningen av $F(r)$ [12, s. 286] liknar avsnitt 2.3.1 men istället för punkter som hör till processen räknas avstånden från godtyckliga punkter $u_1, \ldots, u_m$.

$$\hat{F}(r) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\{d(u_j, \mathbf{X} \cap W) \leq r\}. \tag{39}$$

Denna uppskattning tar heller inte hänsyn till kanteffekter.

A.5 Kantkorrigeringar för $G(r)$ och $F(r)$

Vi har tidigare diskuterat kanteffekter för K-funktionen i 2.2.2 och detta problem är även relevant vid uppskattning av $G(r)$ och $F(r)$. De vanligaste kantkorrigeringsmetoderna som används vid uppskattning av $F(r)$ och $G(r)$ är områdesminskning, Kaplan-Meier korrigering, Hanisch korrigering samt Chiu-Stoyan korrigering.

A.5.1 Områdesminskning

En metod för att få en kantkorrigerad uppskattning av $\hat{G}(r)$ är att endast ta hänsyn till de punkter vars avstånd till observationsfönstrets kant är minst r [12, s. 289]. Detta avstånd ges av

$$b_i = d(x_i, \partial W) \tag{40}$$

där ∂W är fönstrets kant. Med detta avstånd får vi uppskattningen

$$\hat{G}_{bord}(r) = \frac{\sum_{i=1}^{n} \mathbf{1}\{b_i \geq r \wedge d_i \leq r\}}{\sum_{1=1}^{n} \mathbf{1}\{b_i \geq r\}} \tag{41}$$

med

$$d_i = d(x_i, \mathbf{X} \cap W). \tag{42}$$

Detta är alltså andelen punkter som ligger inom det krympta fönstret $W_{\ominus r}$ vars avstånd till dess närmsta granne är mindre än eller lika med r. Det krympta fönstret definieras som [12, s. 215]

$$W_{\ominus r} = \{u \in W \ : \ d(u, \partial W) \geq r\}. \tag{43}$$

På liknande sätt uppskattas tomrumsfunktionen genom [12, s. 287]

$$\hat{F}_{bord}(r) = \frac{\sum_{j=1}^{m} \mathbf{1}\{d(u_j, \mathbf{X} \cap W) \leq r \wedge b_j \geq r\}}{\sum_{j=1}^{m} \mathbf{1}\{b_j \geq r\}} \tag{44}$$

med

$$b_j = d(u_j, \partial W),$$

alltså avståndet från de slumpmässigt valda punkterna till observationsfönstrets kant.

A.5.2 Kaplan-Meier-korrektion för F(r)

Kaplan-Meier uppskattningen av $F(r)$ ges av [12, s. 290]

$$\hat{F}_{km}(r) = 1 - \frac{|W \setminus \mathbf{X}|}{|W|} \exp\left(-\int_0^r \frac{\ell(\partial(\mathbf{X}_{\oplus s}) \cap W_{\ominus s}}{|W_{\ominus s} \setminus \mathbf{X}_{\oplus s}|} ds\right). \tag{45}$$

För att se likheten med (15) börjar Baddeley et al. att definiera

$$\begin{aligned} t(u) &= d(u, \mathbf{X} \cap W), \\ c(u) &= d(u, \partial W), \\ d(u) &= \mathbf{1}\{t(u) \leq c(u)\} \end{aligned}$$

och

$$\widetilde{t}(u) = \min(t(u), c(u)).$$

Vi ser att nämnaren i integralen i (45) kan skrivas som

$$|W_{\ominus s} \setminus \mathbf{X}_{\oplus s}| = \left|\{u \in W \,:\, \widetilde{t}(u) \geq s\}\right|$$

vilket är ytan som har överlevt till avstånd s. Vidare kan täljaren skrivas som

$$\ell(\partial(\mathbf{X}_{\oplus s}) \cap W_{\ominus s}) = \ell(\{u \in W \,:\, \widetilde{t}(u) = s, d(u) = 1\}),$$

vilket tolkas som längden av kurvan som dör vid avstånd s. Detta ser vi är den kontinuerliga liknelsen med täljaren i (15).

A.5.3 Hanisch-korrektion

Hanisch definierar en punktvis väntevärdesriktigt estimator $A(r) = \lambda G(r)$ som [12, s. 291]

$$A_H(r) = \sum_{i=1}^{n} \frac{\mathbf{1}\{d_i \leq b_i \wedge d_i \leq r\}}{|W_{\ominus d_i}|} \tag{46}$$

och en konsistent uppskattning av $G(r)$ som

$$G_H(r) = \frac{\sum_{i=1}^{n} \mathbf{1}\{d_i \leq b_i \wedge d_i \leq r\}}{\sum_{1=1}^{n} \mathbf{1}\{d_i \leq b_i\}} \tag{47}$$

med d_i och b_i som (42) och (40).

A.5.4 Chiu-Stoyan-korrektion

Chiu-Stoyan uppskattningen av $F(r)$ är motsvarigheten till (47) för tomrumsfunktionen och definieras som [12, s. 291]

$$F_{CS}(r) = \int_0^r \frac{\ell(W_{\ominus s} \cap \partial(\mathbf{X}_{\oplus s}))}{|W_{\ominus s}|} \tag{48}$$

där $\mathbf{X}_{\oplus r}$ definieras som [12, s. 266]

$$\mathbf{X}_{\oplus r} = \{u \in \mathbb{R}^2 \,:\, d(u, \mathbf{X}) \leq r\}. \tag{49}$$

A.6 Rayleighfördelning

*Rayleigh*fördelningen är nära släkt med normalfördelningen. Om vi har två normalfördelningar $X_1 \sim \mathcal{N}(0, \sigma^2)$ och $X_2 \sim \mathcal{N}(0, \sigma^2)$ är då

$$\text{Rayleigh}(\sigma) = \sqrt{X_1^2 + X_2^2}. \tag{50}$$

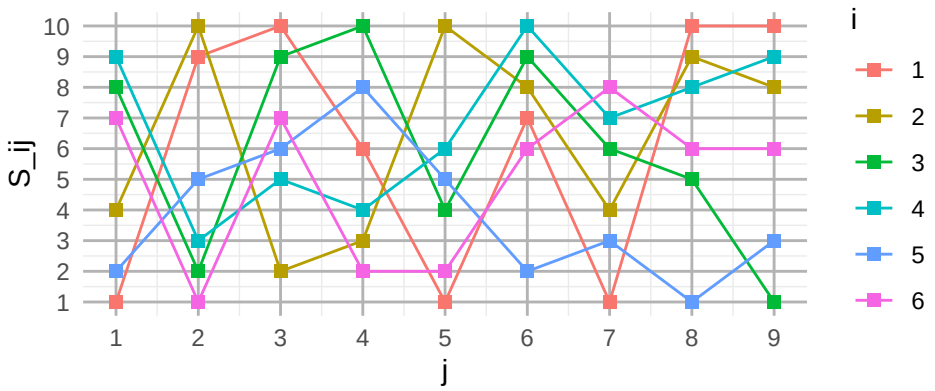
ML-skattningen för σ är [24, s. 174-175]

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{2n}} \tag{51}$$

när x_i är ett stickprov på n observationer från en Rayleigh-fördelning.

A.7 Exempelräkning ERL

Den rigorösa formuleringen av ERL-teststatistikan är inte trivial att förstå. I detta appendix finns därför ett större exempel som illustrerar principen bakom metoden för att jämföra funktionskurvor. Rådatan som vi ska beräkna teststatistikor för är $N = 6$ diskretiserade serier av $m = 9$ värden $S_{ij};\ i = 1, \ldots, N;\ j = 1, \ldots, m$, som visas i figur 18. Det är inte direkt uppenbart vid en inspektion hur extrem var och en av kurvorna är.



Figur 18: Rådata för ERL-exempel.

Först tas de punktvisa extrema rangerna R_{ij} fram enligt ekvation (19). Dessa kan enkelt läsas av på varje lodrät linje, eftersom rangen r_{ij} bara är ordningen nerifrån och upp. Som exempel betraktar vi $j = 1$. Vi har

$$r_{11} = 1,\ r_{21} = 3,\ r_{31} = 5,\ r_{41} = 6,\ r_{51} = 2 \text{ och } r_{61} = 4,$$

vilket enligt ekvation (19) ger

$$R_{11} = 1,\ R_{21} = 3,\ R_{31} = 2,\ R_{41} = 1,\ R_{51} = 2 \text{ och } R_{61} = 3.$$

Den punktvisa extrema rangen beskriver intuitivt uttryckt värdets avstånd från ändarna i ordningen. På samma sätt bestäms den för alla j, där de yttersta värdena får 1, de näst yttersta får 2 och de tredje yttersta (och därmed innersta) värdena får 3.

Nästa steg är att sortera alla de punktvisa extrema rangerna för varje vektor. Som exempel betraktar vi $i = 1$. Enligt föregående stycke beräknar vi

$$R_{11} = 1,\ R_{12} = 2,\ R_{13} = 1,\ R_{14} = 3,\ R_{15} = 1,\ R_{16} = 3,\ R_{17} = 1,\ R_{18} = 1, \text{ och } R_{19} = 1,$$

vilket sorterat med notationen i avsnitt 2.4.3 ger

$$R_{1[1]} = R_{1[2]} = R_{1[3]} = R_{1[4]} = R_{1[5]} = R_{1[6]} = 1,\ R_{1[7]} = 2, \text{ och } R_{1[8]} = R_{1[9]} = 3.$$

Det finns alltså sex index j då serie 1 är mest extrem, ett index j då den är näst mest extrem och två index j då den är tredje mest extrem. För den fortsatta beräkningen är det oväsentligt i vilka index kurvorna är olika extrema, utan det enda som spelar roll är hur många gånger.

Till sist bestäms E_i enligt ekvation (20) utifrån motsvarande sorterade $R_{i[k]}$ för alla serier. Intuitionen är att minst testvärde ska tilldelas den kurva som är mest extrem i flest index (som alltså betraktas som den mest extrema kurvan). Skulle det vara så att det finns två kurvor som är mest extrema i lika många index så avgörs det genom att jämföra antalet index de är näst mest extrema i, och så vidare. På samma sätt fortsätter detta även för de näst mest extrema kurvorna, och så vidare.

I ekvation (20) jämförs serie i mot alla serier i', och man försöker finna ett d sådant att i och i' är lika extrema i $d-1$ index, men inte i d index (då i' är mer extrem). I så fall kan man säga att i' ”vinner över” i. Värdet E_i är alltså antalet serier som ”vinner över” serie i delat med antal serier för att normalisera till $[0, 1)$, och ju mindre E_i desto mer extrem. Det bildas således en ordning på kurvorna.

Matriserna med alla R_{ij} respektive alla $R_{i[k]}$ blir (jämför den rad och kolonn vi tidigare beräknat)

$$\begin{bmatrix} 1 & 2 & 1 & 3 & 1 & 3 & 1 & 1 & 1 \\ 3 & 1 & 1 & 2 & 1 & 3 & 3 & 2 & 3 \\ 2 & 2 & 2 & 1 & 3 & 2 & 3 & 2 & 1 \\ 1 & 3 & 2 & 3 & 2 & 1 & 2 & 3 & 2 \\ 2 & 3 & 3 & 2 & 3 & 1 & 2 & 1 & 2 \\ 3 & 1 & 3 & 1 & 2 & 2 & 1 & 3 & 3 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 2 & 3 & 3 \\ 1 & 1 & 1 & 2 & 2 & 3 & 3 & 3 & 3 \\ 1 & 1 & 2 & 2 & 2 & 2 & 2 & 3 & 3 \\ 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 \\ 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 \\ 1 & 1 & 1 & 2 & 2 & 3 & 3 & 3 & 3 \end{bmatrix}.$$

Därefter försöker vi finna d för alla serier i' enligt ekvation (20), eller uppge om inget existerar. I matrisform framgår detta tydligt eftersom en rad i' ”vinner över” en annan rad i omm den kommer före i lexikografisk ordning. Som exempel beräknar vi E_3:

1. $d = 3$.
2. $d = 3$.
3. Existerar ej. En serie kan inte vara mer extrem än sig själv.
4. Existerar ej. Vi har $R_{4[7]} > R_{3[7]}$.
5. Existerar ej. Vi har $R_{5[7]} > R_{3[7]}$.
6. $d = 3$.

Eftersom vi fann d för tre stycken serier betyder det att $E_3 = 3/6 = 1/2$. På samma sätt erhålls

$$E_1 = 0,\ E_2 = 1/6,\ E_3 = 1/2,\ E_4 = 2/3,\ E_5 = 2/3 \text{ och } E_6 = 1/6.$$

En sista anmärkning är att om motsvarande beräkning görs med `GET` kommer teststatistikan skalas om för att indikera förekomsten av lika extrema värden; bara unika rader i $R_{i[k]}$-matrisen betraktas. Den inbördes ordningen mellan E_i blir dock likadan, vilket gör resultatet ekvivalent eftersom vi bara relaterar testvärdena till varandra. I praktiken, när ett mycket större antal serier med många fler värden betraktas, som i tillämpningen i rapporten, så är motsvarande lika rader i denna matris mycket ovanligt.