Statistisk modellering af ekstremværdier

Statistical Modeling of Extreme Values

Copenhagen Business School 2016

Cand.merc.(mat.)

Kandidatafhandling

Pernille Louise Hansen

Helle Johansen

Vejleder: Dorte Kronborg

Afleveret den 29. april 2016

148.811 anslag/103 sider

Abstract

The focus in this thesis is to study Extreme Value Theory (EVT) and statistical methods, to describe extreme observations, with regard to financial risk management. These different statistical methods is used to model independent and identically distributed (i.i.d.) data and clusters of volatility. For the selection of the extreme values, the Method of Block Maxima and Peaks over Threshold (POT) are introduced, which are two different approaches used to select extreme values. The main focus of the first part is the assumption of i.i.d. data, where the focus is on the POT method and the generalized pareto distribution (GPD). To model the GPD and estimate the parameters and risk measures, several different techniques are used. The estimation methods Method-of-Moments (MOM), Elemental-Percentile-Method (EPM), Probability-Weighted-Method (PWM), Maximum Likelihood Estimation (MLE) and L-Moments (LMOM) method are implemented in daily prices for the Vestas stock, wherein the risk measures Value-at-Risk (VaR) and Expected Shortfall (ES) are used. The achieved results of five different estimation methods are: the MOM and EPM method is not a valid choise for modeling the extreme observations. PWM, MLE and LMOM are preferred. These methods provide approximately the same values for VaR and ES. The focus in the second part of the thesis is point processes and clusters of volatility, where the Poisson process and Hawkes Self-Exciting POT are introduced. The Poisson point process gives the same results as the i.i.d. POT model, and the Hawkes POT models, which take clustering into account, result in almost identically estimates, however smaller VaR and ES measures.

Indhold

1 Indledning **4**

1.1 Motivation 4

1.2 Problemformulering 6

1.3 Metode 6

1.4 Afgrænsning 10

I Ekstremværdi teori **11**

2 Maksima **12**

2.1 Konvergens 12

2.1.1 Konvergens i fordelinger - Svag konvergens 12

2.1.2 Konvergens i sandsynligheder 13

2.1.3 Næsten sikker konvergens 14

2.1.4 Konvergens til typer af fordelinger 15

2.2 Egenskaber ved maksima 15

2.2.1 Maks-stabile fordelinger 20

2.2.2 Maximum Domain of Attraction 21

3 Ekstremværdifordelinger **21**

3.1 Blok Maksima 21

3.2 Den generaliserede ekstremværdifordeling 23

3.2.1 Udledning af fordelinger 25

3.3 Peaks-over-Threshold metoden 28

3.4 Den generaliserede paretofordeling 29

3.5 Bestemmelse af threshold 33

3.5.1 Mean Excess Plot . . . 34

4 Estimation **37**

4.1 Method-of-Moments . . . 38

4.2 Elemental-Percentile-Method . . . 40

4.3 Probability-Weighted-Moments . . . 44

4.4 L-Moments-Method . . . 46

4.5 Maksimum-Likelihood-Estimation . . . 49

4.6 Estimation af risikomål . . . 52

4.6.1 Kritik af VaR . . . 54

4.7 Test af estimater . . . 55

4.7.1 Standard Error og konfidensinterval . . . 56

4.7.2 Bias . . . 57

4.7.3 Mean Sqared Error . . . 57

5 Punktprocesser **59**

5.1 Generelt om punktprocesser . . . 60

5.2 Poisson punktprocessen . . . 62

6 Self-Exciting punktprocesser **66**

6.1 Hawkes . . . 67

6.2 Hawkes POT . . . 68

6.3 Hawkes POT med uforudsigelige mærker . . . 68

6.4 Hawkes POT med forudsigelige mærker . . . 70

6.5 Risikomål . . . 71

6.6 Test af punktproces modeller . . . 72

II Analyse **74**

7 Analyse af ekstremværdier **74**

7.1 Data 74

7.2 Fastsættelse af threshold 77

7.3 Estimation af parametre 79

7.4 Estimation af risikomål 86

7.5 Punktprocesser 89

8 Diskussion og perspektivering **97**

9 Konklusion **101**

Litteratur **103**

1 Indledning

1.1 Motivation

I den finansielle sektor er risikostyring et stort og vigtigt fokusområde, og dermed et interessant emne at studere. Hver gang finansielle institutioner, som for eksempel pensionskasser og banker, låner penge ud eller investerer i nye aktiver, påtager de sig en risiko. Det kan have store konsekvenser, hvis denne risiko ikke modelleres korrekt. Er de finansielle institutioner ikke villige til at påtage sig en form for risiko, er sandsynligheden for et fordelagtigt afkast meget lille, hvorfor risiko er en naturlig del af de finansielle institutioners hverdag. Statistisk modellering af risiko er dermed en vigtig forudsætning for, hvordan finansielle institutioner klarer sig på markedet.

Inden for risikostyring modelleres der ofte på tab, som kan inddeles i forventede tab og uforventede tab. De forventede tab kan anses som værende en omkostning, og kan direkte medtages i beregningerne, mens de uforventede tab er uforudsigelige, og dermed udgør en risiko, som statistisk kan modelleres.

Det er relevant for de finansielle institutioner at kunne styre og måle risiko, for dermed at sikre deres portefølje mod store tab. Det handler for institutioner om at opretholde reglerne opstillet af Bank for International Settlements (BIS) vedrørende risici og solvens. BIS har til formål at sikre nok kapital, hvilket gøres ved at udstede lovmæssige kapitalkrav til institutionerne[1]. Det er selvfølgelig også i institutionernes egen interesse at kunne holde styr på de finansielle risici, idet et enkelt ekstremt tab kan have store konsekvenser. Det kan resultere i, at institutionen ryger under det fastlagte solvenskrav fra BIS og i sidste ende medføre konkurs.

Det er altså vigtigt at kunne bestemme den kapital, en given finansiel institution skal opretholde, for at sikre sig mod store tab, og det er netop derfor, det er interessant at studere ekstreme hændelser, som sjældent sker. Inden for statistik kaldes dette område for ekstremværdi teori (EVT), hvor det på baggrund af den statistiske modellering af ekstreme værdier er muligt at finde relevante risikomål.

[1] https://www.finanstilsynet.dk/da/Leksikon/Individuelt-solvenskrav.aspx

Der findes flere forskellige former for risici, hvor fire af de mest benyttede kort er forklaret nedenfor.

- **Markedsrisiko** er risikoen for, at værdien af en portefølje ændrer sig som en konsekvens af ændringer på de finansielle markeder.
- **Kreditrisiko** er risikoen for at tabe penge, som konsekvens af at modparten ikke kan opfylde sine forpligtigelser. Et eksempel kunne være, hvis en virksomhed går konkurs, og ikke kan tilbagebetale eventuel gæld.
- **Likviditetsrisiko** beskriver risikoen forbundet med, at et værdipapir ikke kan omsættes på det tidspunkt, hvor man ønsker at sælge eller købe det. Med andre ord er det risikoen for, at prisen på ens aktiver falder.
- **Operationel risiko** er den risiko, en virksomhed kan være udsat for ud over de tre ovenstående former for risici. Denne risiko er dermed svær at måle og dækker over risici, som ikke er forventelige, som for eksempel risiko forbundet med nedbrud af it systemer og lignende.

Det ønskes i denne afhandling at studere forskellige estimationsmetoder inden for EVT'en og disse vil anvendes på en enkelt aktie fra det danske C20 indeks. Det er dermed markedsrisiko, der er fokus på, og for selve bestemmelsen af markedsrisikoen, er Value-at-Risk (VaR) et vigtigt værktøj. VaR er et risikomål, som beskriver det potentielle tab, en finansiel institution maksimalt kan tabe inden for en given periode med en given sandsynlighed. Det er vigtigt for en virksomhed at kunne kontrollere sin risiko, hvorfor det er fordelagtigt, på baggrund af korrekt statistisk modelleret data, at kunne beregne et risikomål såsom VaR. Risikomålet VaR er blevet en integreret del af reglementet sat af (BIS), hvilket antyder, at det er en brugbar og valid metode til at modellere risiko. Det er her relevant at nævne vigtigheden af den bagvedliggende statistiske metode, som der i denne afhandling sættes fokus på.

Bestemmelsen af VaR kræver, at man kan finde en korrekt fordelingsfunktion, så institutionen ikke står tilbage med et større tab end forventet. Det er dermed inden for risikostyring interessant at studere de ekstreme værdier, som ligger i halen af en sandsynlighedsfordeling.

I finansielle datasæt ligger udfordringen ofte i, at data består af flere ekstreme hændelser end normalt, og dermed har tendens til fede haler i forhold til normalfordelingen. Det betyder, at normalfordelingen ikke er en særlig god approksimation til finansielle afkast. Inden for EVT'en handler det om at bestemme, hvornår en observation er ekstrem og finde en fordeling, som beskriver disse ekstreme værdier bedst muligt.

1.2 Problemformulering

Afhandlingens hovedformål er at studere statistiske modeller til beskrivelse af ekstreme værdier med henblik på finansiel risikostyring. Metoderne vil blive illustreret på Vestas aktien.

Til dette studie er følgende delpunkter opstillet

- Hvornår kan en observation siges at være ekstrem, og hvordan udvælges ekstreme observationer?
- Hvilke statistiske fordelinger og estimationsmetoder er fordelagtige at benytte inden for ekstremværdi teori?
- Hvordan kan punktprocesser benyttes til at modellere ekstreme observationer?
- Analyse af ekstreme tab på Vestas aktien

1.3 Metode

Denne afhandling har en teoretisk tilgang, hvor problemformuleringen forsøges besvaret ved hjælp af litteratur og empiri. Strukturen kan beskrives som værende deduktiv og overordnet opdeles i en

teoretisk og empirisk del.

Del I udgør det teoretiske afsnit og giver en introduktion til EVT, beskrivelse af ekstremværdifordelinger, samt en beskrivelse af metoder til udvælgelse og estimation af ekstreme observationer. Del I indeholder også en udvidelse og dermed en anden tilgang til modellering af ekstreme observationer: Punktprocesser. Del II indeholder den empiriske del af afhandlingen, hvor teorien benyttes på data for Vestas aktien. Denne del inkluderer beskrivelse af data, analyse og resultater. For at give et overblik over de benyttede modeller og metoder er de ved hjælp af et Flowchart illustreret i figur (1).

Afhandlingen afrundes med en diskussion af resultater og estimationsmetoder, perspektivering til andet relevant litteratur, samt en konklusion som opsummerer alle relevante resultater til besvarelse af problemformuleringen.

Det er hensigten, at besvarelsen af afhandlingens problemformulering vil følge en deduktiv metode, hvor vi vil teste hvordan teori passer til data fra den virkelige verden. Denne metode er forskellig fra den induktive tilgang, hvor man ofte i den finansielle verden, stoler blindt på de allerede udviklede modeller. Afhandlingen vil overordnet været baseret på den positivistiske tankegang.

Datagrundlaget i analysen består af daglige aktiekurser for Vestas aktien i perioden fra den 10.05.2000 til den 28.01.2016. Vi har downloadet aktiekurserne fra http://finance.yahoo.com/, der anses som værende en valid kilde. Til at analysere de ekstreme observationer i datasættet og implementere de valgte estimationsmetoder, har vi benyttet programmeringsværktøjet `RStudio`, som er et integreret del af `R` statistik programmet. `R` er et frit programmeringsværktøj inden for statistik, som vi løbende igennem studiet har arbejdet i. Til selve databehandlingen i `R` har vi benyttet forskellige pakker, som overordnet set kan meget af det samme. Vi har i de følgende delpunkter beskrevet pakkerne, og hvad de i denne afhandling er brugt til.

- `evir`-pakken benyttes til at modellere ekstreme værdier, blandt andet til QQ-plots, 'Mean

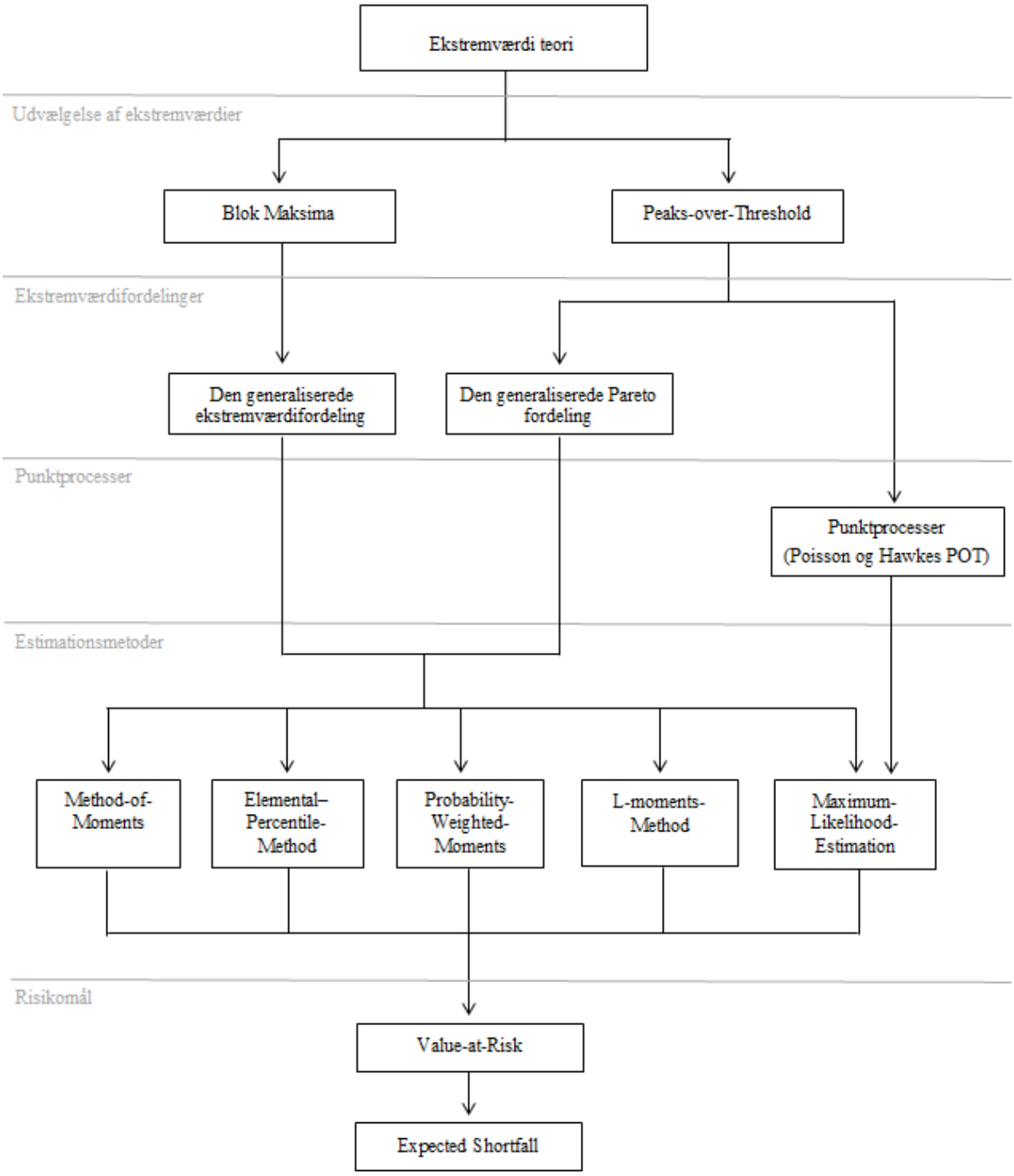
Excess' plot og Declustering af data.

- `evd`-pakken indeholder funktioner til modellering af ekstremværdifordelinger, og benyttes til at illustrere 'Mean Residual Life plot'.
- `QRM`-pakken indeholder funktioner til modellering af ekstremværdifordelinger, risikomål samt punktprocesser.
- `gPdtest`-pakken benyttes til at udføre en Bootstrap Goodness-of-Fit test for den generaliserede paretofordeling.
- `PerformanceAnalytics`-pakken er et økonometrisk redskab til risikoanalyser, og bruges til at illustrere histogrammer.
- `lmom`-pakken indeholder funktioner til beregning af momenter og estimater ved L-moment metoden.
- `ismev`-pakken benyttes til illustration af likelihoodfunktioner.
- `timeSeries` og `xts`-pakkerne benyttes til at opnå det korrekte tidsserie format.

Den benyttede `R` kode er vedlagt på en CD-ROM.

I del I er enkelte dele af teorien illustreret ud fra 'Danish Fire Insurance' data, som er et indbygget og dermed tilgængeligt datasæt i `R`. Datasættet er et meget benyttet datagrundlag i EVT'en, da det giver et pænt billede af, hvordan ekstreme værdier kan modelleres i praksis. I [12, Embrechts et al., 2012], som er en del af afhandlingens primære litteratur, er 'Danish Fire Insurance' også benyttet, og vi bruger i denne afhandling datasættet som en indikator for, hvordan den pågældende teori, når den implementeres på data med ekstreme observationer, kan se ud.

Hvis det har virket meningsforstyrrende at oversætte engelske teoretiske begreber, har vi valgt at bibeholde den engelske betegnelse. I bilag er der vedlagt en liste over de forkortelser, som er benyttet i afhandlingen.



Figur 1: Flowchart over de benyttede metoder og modeller i afhandlingen.

1.4 Afgrænsning

For at kunne være specifikke i besvarelsen af afhandlingens problemformulering, har vi foretaget følgende afgrænsninger.

Af hensyn til afhandlingens omfang har vi valgt at studere modellerne og metoderne illustreret i figur (1). Vi har altså valgt et udsnit af de tilgængelige statistiske metoder til beskrivelse af ekstreme værdier.

I teorien har vi valgt ikke at udlede alle modeller og ligninger. Derudover afgrænses teoriafsnittet om punktprocesser til ikke at indeholde den bagvedlæggende teori omkring mængdelære. Inden for punktprocesser studeres kun den mærkede Hawkes POT punktproces, og punktprocesser som leder til denne.

Den analytiske del af afhandlingen er til for at illustrere den gennemgåede teori, hvorfor vi har afgrænset os til kun at analysere på en enkelt aktie: Vestas aktien.

Det antages i denne afhandling, at læseren har en generel forståelse for sandsynlighedsregning, statistik og finansieringsteori.

Del I

Ekstremværdi teori

Ekstremværdi teori (EVT) er et område inden for sandsynlighedsregning, hvor fokus er på sandsynlighedsmassen, der befinder sig i halen af en sandsynlighedsfordeling. Det vil sige, man studerer de ekstreme hændelser, som sjældent sker. I denne afhandling vil de ekstreme hændelser være ekstreme tab på det danske aktiemarked. EVT'en er på baggrund af de sjældne hændelser, dermed meget forskellig fra den klassiske statistiske teori, hvor der primært fokuseres på data, som ligger i midten af en sandsynlighedsfordeling. I EVT'en er det ikke denne centrale del som studeres, men derimod observationer som ligger i halen af fordelingen. For at estimere parametre, og dermed fitte en fordeling til de ekstreme observationer kræves der en beskrivelse af, hvordan datapunkterne opfører sig. Det er derfor relevant at analysere data, og undersøge om man har uafhængige identisk fordelte (i.i.d.) stokastiske variable, eller om der er tendens til afhængighed i data omkring de ekstreme hændelser, som for eksempel klyngedannelse. Eventuel afhængighed skal tages i betragtning i den videre modellering af datasættet. En udfordring ved modellering af ekstremværdifordelinger kan være antallet af ekstreme observationer. Hvis data består af få ekstreme observationer, kan det være svært at opnå præcise estimater af parametrene i den pågældende fordeling, hvorfor det er vigtigt, at benytte en estimationsmetode, hvis parametre beskriver data bedst muligt.

Det kan være en udfordring at bestemme, hvornår en værdi er ekstrem, og der undersøges derfor i denne afhandling to metoder til at finde de ekstreme værdier. Data kan som nævnt være forskelligt, hvilket kan have betydning for udvælgelsen af de ekstreme værdier. I EVT'en er der to fundamentale tilgange til at identificere ekstreme hændelser: Blok Maksima (BM) og Peaks-over-Threshold (POT) metoden. BM metoden består i at opdele observationsperioden i ikke-overlappende tidsperioder af samme størrelse, og betragte den største observation i hver periode. Disse observationer udgør de ekstreme hændelser, og bliver kaldt for Blok Maksima. I POT metoden defineres de ekstreme

hændelser, som værende de observationer der overskrider en høj øvre fastlagt grænse u, også kaldet et threshold.

2 Maksima

2.1 Konvergens

I EVT'en er det altså interessant at undersøge, hvordan observationer i halen af en sandsynlighedsfordeling opfører sig. Det er derfor vigtigt at studere, hvordan stokastiske variable på forskellige måder konvergerer, da det kan give en bedre forståelse af fordelingens grænseværdier. I denne afhandling benyttes flere forskellige former for konvergens, som introduceres i dette afsnit.

De følgende underafsnit har reference til [12, Embrechts et al, 2012, Appendix A1], hvor der generelt for alle konvergens typer tages udgangspunkt i en sekvens af stokastiske variable X_n : $X_1, X_2, ..., X_n$.

2.1.1 Konvergens i fordelinger - Svag konvergens

Det siges, at X_n konvergerer i en fordeling, eller opfylder svag konvergens til en stokastisk variabel X, også skrevet ($X_n \xrightarrow{d} X$), hvis følgende relation gælder for alle begrænsede kontinuerte funktioner f:

$$E\left[f(X_n)\right] \to E\left[f(X)\right],\ n \to \infty. \tag{1}$$

Med andre ord vil den forventede værdi af funktionen $f(X_n)$ gå mod den forventede værdi af $f(X)$, for $n \to \infty$. Udtrykket 'Svag konvergens' benyttes, da det er forventninger der modelleres med og ikke reelle tal. Svag konvergens kan, på samme måde som i ligning (1), opskrives ved hjælp af fordelingsfunktionerne hørende til X_n og X. Det gælder at $X_n \xrightarrow{d} X$ hvis og kun hvis, følgende

relation holder for alle kontinuerte punkter y i fordelingsfunktionen $F_X(y)$:

$$F_{X_n}(y) \to F_X(y) \ , \ n \to \infty. \tag{2}$$

Det betyder, at svag konvergens er opfyldt, hvis fordelingsfunktionen $F_{X_n}(y)$ konvergerer mod $F_X(y)$ for alle mulige punkter y.

2.1.2 Konvergens i sandsynligheder

Når der tales om konvergens i sandsynligheder, handler det om, hvordan en sandsynlighed konvergerer, i stedet for at studere relationen mellem to funktioner som ved svag konvergens. Det undersøges her, hvordan sandsynligheden for et uventet resultat bliver mindre eller større, som konvergensen skrider frem.

At X_n konvergerer i sandsynlighed til en stokastisk variabel X, kan skrives som $(X_n \xrightarrow{P} X)$. Denne konvergens i sandsynlighed kan for alle positive konstanter ϵ udtrykkes ved

$$P\left(|X_n - X| > \epsilon\right) \to 0 \ , \ n \to \infty. \tag{3}$$

Det vil sige at sandsynligheden for, at X_n afviger mere fra X end værdien af ϵ, som definerer et meget lille tal, går mod nul, som n går mod uendelig. Altså vil sandsynligheden for, at for eksempel et statistisk estimat kommer tættere på sin 'sande' værdi, blive større for større værdi af n.

Konvergens i sandsynligheder er en stærkere form for konvergens end konvergens i fordelinger. Det gælder, hvis der haves konvergens i sandsynligheder, at der også er konvergens i fordelinger, men ikke vice versa. Derfor benyttes konvergens i sandsynligheder, når den er opfyldt, men det huskes stadig, at konvergens i fordelinger implicit er opfyldt.

2.1.3 Næsten sikker konvergens

Det antages at X_n 'næsten sikkert', eller med sandsynlighed en, konvergerer til den stokastiske variabel X, hvis følgende relation i sandsynlighed holder for næsten alle $\omega \in \Omega$:

$$X_n(\omega) \to X(\omega) \ , \ n \to \infty. \tag{4}$$

Næsten sikker konvergens er opfyldt, hvis den stokastiske variabel med sandsynlighed en konvergerer mod det 'sande' estimat for næsten alle ω i

$$P(X_n \to X) = P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1. \tag{5}$$

Relationen for næsten sikker konvergens kan opskrives som $X_n \overset{a.s}{\to} X$, hvor a.s står for 'almost surely' konvergens. I relation til konvergens i sandsynligheder kan udtrykket i ligning (5) opskrives som

$$\sup_{k \geq n} |X_k - X| \overset{P}{\to} 0. \tag{6}$$

Hvis differencen mellem de to stokastiske variable i ligning (6) i sandsynlighed konvergerer mod nul, vil sandsynligheden for at X_k konvergerer mod X være en. I 'næsten sikker' konvergens, skal relationen ikke gælde for alle X_n men kun for en subsekvens X_k af X_n. Hvis tilfældet var, at dette skulle gælde for alle X_n, ville der være tale om begrebet 'absolut konvergens', som opskrives ved

$$\lim_{n \to \infty} X_n(\omega) = X(\omega), \tag{7}$$

hvor de stokastiske variable for $n \to \infty$, vil være lig med hinanden for alle potentielle estimater ω.

Næsten sikker konvergens er stærkere end både konvergens i sandsynligheder og konvergens i fordelinger. Det gælder dog, at konvergens i sandsynligheder ikke nødvendigvis medfører 'næsten sikker' konvergens. Hvis der i næsten sikker konvergens ikke kun analyseres på en subsekvens, er det

derimod den stærkeste form for konvergens: Absolut konvergens som er opfyldt. Absolut konvergens indeholder alle former for konvergens beskrevet i dette afsnit.

2.1.4 Konvergens til typer af fordelinger

Hvis der findes to sæt af stokastiske variable X og Y med samme fordeling, kan deres relation opskrives som

$$X \stackrel{d}{=} Y. \tag{8}$$

I konvergens til typer af fordelinger, defineres først to konstanter $a \in \mathbb{R}$, og $b > 0$, hvilke bruges til at beskrive de to fordelinger ud fra hinanden. Det siges, at fordelingerne hørende til to sæt af stokastiske variable, X og Y, hører til samme type familie, eller er af samme type, hvis der eksisterer konstanter $a \in \mathbb{R}$, og $b > 0$, således at følgende lineære relation er opfyldt:

$$X \stackrel{d}{=} bY + a. \tag{9}$$

Med andre ord er de to sæt af stokastiske variable samme type hvis der findes en lineær sammenhæng mellem dem.

2.2 Egenskaber ved maksima

I EVT'en har de ekstreme observationer nogle specifikke egenskaber, som i dette afsnit vil studeres nærmere. Disse egenskaber spiller en vigtig rolle, når der senere skal findes og fittes fordelinger til de ekstreme observationer. Følgende underafsnit omkring maksima har reference til teorien beskrevet i [12, Embrechts et al.2012, 114-115].

Det antages, at $X_1, X_2, ..., X_n$ er en sekvens af i.i.d. ikke-degenererede stokastiske variable, med fordelingsfunktion F. En stokastisk variabel siges at være degenereret, hvis den er konstant med

sandsynlighed $P = 1$. Det vil sige, at den stokastiske variabel er degenereret hvis, for nogen $a \in \mathbb{R}$, $P(X = a) = 1$. Er denne relation ikke opfyldt, kaldes den stokastiske variabel for ikke-degenereret.

Det ønskes at se på disse stokastiske variable som værende maksima værdier, og det kan opskrives som

$$M_1 = X_1 \quad , \quad M_n = \max(X_1, ..., X_n), \tag{10}$$

hvor M_n nu udgør n antal ekstreme maksima værdier. Hvis man har med minima at gøre, er der en klar sammenhæng til maksima, hvor relationen kan opskrives som

$$\min(X_1, ..., X_n) = -\max(-X_1, ..., -X_n). \tag{11}$$

I denne afhandling vil fokus kun være på maksima værdier, men fremgangsmåden for modellering af minima værdier ville være den samme.

Det er fra generel sandsynligheds-og fordelingsteori velkendt, at fordelingen af stokastiske variable kan findes ved at beskrive sandsynligheden for, at de stokastiske variable ligger under en given fraktil x. Det betyder, at fordelingsfunktionen hørende til maksima kan opskrives som sandsynligheden for, at M_n ligger under en fraktil x:

$$\begin{aligned} P\{M_n \leq x\} &= P\{X_1 \leq x, ..., X_n \leq x\} \\ &= P\{X_1 \leq x\} \cdot ... \cdot P\{X_n \leq x\} \\ &= \{F(x)\}^n, \end{aligned} \tag{12}$$

hvor fordelingsfunktionen $\{F(x)\}^n$ endnu er ukendt. Det er netop denne fordelingsfunktion, som skal fastsættes til den videre modellering af de ekstreme observationer.

Idet sandsynligheden for ekstreme hændelser kan forklares ud fra halen af en sandsynlighedsfordeling, sættes M_n i relief til halen i en fordelingsfunktion F. Det er derfor vigtigt, at definere det

højre endepunkt i en fordeling som

$$x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}. \tag{13}$$

Det højre endepunkt x_F er den største værdi af x, hvor mængden af fordelingsfunktionen er mindre end 1. Med andre ord, det mindste x som er større end $F(x) < 1$. Det højre endepunkt kan benyttes til at studere grænserne i en fordeling F, hvilke kan skrives som

$$x \geq x_F \quad : \quad P(M_n \leq x) = 1 \tag{14}$$

$$x < x_F \quad : \quad P(M_n \leq x) = \underbrace{F(x)^n}_{<1} \to 0 \ , \ n \to \infty. \tag{15}$$

Udtrykket i ligning (14) er selvsagt, idet det angiver sandsynligheden for, at en stokastisk variabel ligger under en fraktil x, som er højere end det højre endepunkt. Eftersom x_F er defineret som værende den største værdi af x hvor $F(x) < 1$, vil sandsynligheden for, at en stokastisk variabel ligger under en fraktil, som er højere end denne, være lig med en. Konvergensen i udtryk (15) fremkommer af definitionen af fordelingsfunktionen for maksima i ligning (12). Det resulterer i et produkt af sandsynligheder, som er mindre end en, og som for $n \to \infty$ vil gå mod nul.

Udtrykkene i ligning (14) og (15) viser, at for $n \to \infty$ vil $M_n \xrightarrow{P} x_F$ for $x_F \leq \infty$. Det betyder, at uanset om en fordeling har et endeligt eller uendeligt højre endepunkt, vil den stokastiske variabel M_n, i sandsynlighed konvergere mod det højre endepunkt. Da sekvensen M_n er stigende i henhold til antallet af observationer n, kan det ydermere antages, at konvergensen er endnu kraftigere og opfylder 'næsten sikker' konvergens, hvilket kan skrives som

$$M_n \xrightarrow{a.s} x_F \ , \ n \to \infty. \tag{16}$$

Maksima værdierne M_n konvergerer altså mod det højre endepunkt, lige meget om dette er

endeligt eller uendeligt. Det vil sige, at M_n degenererer mod en punktmasse, eller en konstant værdi. For at afhjælpe dette, og opnå en ikke-degenereret fordeling studeres det hvordan data kan normaliseres.

For at forstå hvordan data inden for EVT kan normaliseres, undersøges forholdet mellem normalfordelingen og den centrale grænseværdisætning (CLT), da der her findes en sammenhæng. Derefter kan ideen påføres til maksima værdierne, og en normalisering i EVT'en kan opnås.

Med reference til [20, McNeil et al. 2015, 136] tages der udgangspunkt i n i.i.d. stokastiske variable $X_1, X_2, ..., X_n$ med endelig varians. CLT'en for tilnærmelsesvis normaliserede summer, benytter normaliseringskonstanterne $a_n = nE(X_1)$ og $b_n = \sqrt{nVar(X_1)}$. Hvis summen af de n første stokastiske variable angives som værende $S_n = X_1 + X_2 + ... + X_n$, vil de tilnærmelsesvis normaliserede summer ifølge CLT'en ud fra konstanterne a_n og b_n, konvergere mod en standard normalfordeling for $n \to \infty$. Dette kan opskrives som

$$\lim_{n\to\infty} P\left(\frac{S_n - a_n}{b_n} \leq x\right) = \phi(x) \quad , \quad x \in \mathbb{R}, \tag{17}$$

hvor $\phi(x)$ er fordelingsfunktionen hørende til standard normalfordelingen, som angiver sandsynlighedsmassen under en given fraktil x.

Normalfordelingen er ofte fordelagtig at benytte, da den udelukkende kan beskrives ud fra dens første to momenter. Det første moment er middelværdien, og det andet moment udtrykker variansen. Det tredje moment angiver skævheden af fordelingen, hvilket i normalfordelingen er nul, da fordelingen er symmetrisk. Det fjerde moment som er det interessante inden for ekstremværditeorien, angiver kurtosis eller halevægten. Det er velkendt, at finansielle afkast tilhører en fordeling med fede haler, som normalfordelingen ikke har. Normalfordelingen er altså ikke i stand til, at opfange sandsynlighedsmassen i halerne, hvor de ekstreme observationer ligger. Normalfordelingen er god at tage udgangspunkt i, hvis data har en central tendens, men ikke fordelagtig at benytte hvis det er

data, som ligger i halen af sandsynlighedsfordelingen der ønskes undersøgt.

Store tab på det finansielle marked sker ikke ofte, men når de sker, kan det have store konsekvenser. Det er dermed vigtigt at kunne beskrive ekstreme tab og finde en fordeling, som beskriver disse tab på den mest efficiente måde. Det er dermed nødvendigt at føre normaliseringsteorien i forhold til normalfordelingen videre, så den kan benyttes inden for EVT'en.

Fordelingsfunktionen hørende til maksima værdierne $F(x)^n$ er endnu ukendt, og det er vist, at de stokastiske maksima værdier M_n degenererer mod fordelingens højre endepunkt x_F. Denne degenerering kan afhjælpes ved at normalisere maksima værdierne ved brug af samme metode som for normalfordelte stokastiske variable. I ligning (17) erstattes S_n med M_n, og normaliseringen af maksima kan opskrives som

$$P\left(\frac{M_n - a_n}{b_n} \leq x\right). \tag{18}$$

Da ligning (18) tager udgangspunkt i maksima værdier M_n benyttes i stedet for CLT'en, Fisher-Tippets sætning som i [12, Embrechts et al. 2012, Theorem 3.2.3] er defineret ved

> *"Lad (X_n) være en sekvens af i.i.d. stokastiske variable. Hvis der eksisterer normaliseringskonstanter $\{c_n > 0\}$ og $\{d_n \in \mathbb{R}\}$ og en ikke-degenereret fordelingsfunktion H sådan at*
>
> $$c_n^{-1}(M_n - d_n) \xrightarrow{d} H, \tag{19}$$
>
> *da tilhører H en af de tre typer af fordelinger hørende til den generaliserede ekstremværdi familie".*

I den videre teori benyttes fortsat notationen a_n og b_n, og den generaliserede ekstremværdi (GEV) familie vil fremadrettet betegnes $G(x)$.

Fisher-Tippets sætning kan benyttes som analog til CLT'en, der for $n \to \infty$ går mod normalfordelingen, hvor den ukendte fordelingsfunktion F inden for EVT'en vil tilnærme sig GEV familien $G(x)$. Denne familie af fordelinger vil blive studeret nærmere i afsnit 3.2.

De normaliserede maksima værdier degenerer nu ikke længere mod en punktmasse, men konvergerer i fordeling til GEV familien, hvilket kan skrives som

$$\frac{M_n - a_n}{b_n} \xrightarrow{d} G(x). \tag{20}$$

Denne ikke-degenererede fordelingsfunktion leder videre til to vigtige egenskaber: Maks-stabilitet og Maximum Domain of Attraction.

2.2.1 Maks-stabile fordelinger

Ifølge [12, Embrechts et al., 2012, 120] siges en fordeling F at være maks-stabil, hvis der for alle $n \geq 2$ og en sekvens af i.i.d. stokastiske variable X_n eksisterer konstanter $b_n > 0$ og $a_n \in \mathbb{R}$, som opfylder

$$\frac{M_n - a_n}{b_n} \stackrel{d}{=} X. \tag{21}$$

Udtrykket i ligning (21) kan også opskrives som $M_n \stackrel{d}{=} b_n X + a_n$. Det betyder, at enhver maks-stabil fordeling er en grænsefordeling for normaliserede maksima af i.i.d. stokastiske variable, og som jævnfør udtrykket i ligning (20), vil tilhøre GEV familien. Det vil omvendt også gælde, at enhver ekstremværdifordeling er maks-stabil, og at de maks-stabile fordelinger er de eneste ikke-degenererede grænsefordelinger for normaliserede maksima af i.i.d. stokastiske variable.

2.2.2 Maximum Domain of Attraction

Med reference til [12, Embrechts et al., 2012, 128] siges en fordeling F af en sekvens af i.i.d. stokastiske variable X_n at være i Maximum Domain of Attraction (MDA) af en ekstremværdifordeling $G(x)$, hvis der eksisterer konstanter $b_n > 0$ og $a_n \in \mathbb{R}$, hvor følgende udtryk er opfyldt

$$\lim_{n\to\infty} n\overline{F}(b_n x + a_n) = -ln\, G(x), \tag{22}$$

hvor $\overline{F}$ er halefordelingen af F. Det vil med andre ord sige, at hvis de normaliserede maksima værdier konvergerer mod en ekstremværdifordeling for $b_n > 0$ og $a_n \in \mathbb{R}$, da gælder det, at F er i MDA af G, hvilket også kan skrives som $F \in MDA(G)$.

Ifølge [12, Embrechts et al., 2012, 116] gælder det ved hjælp af Poisson approksimation at $P(M_n \leq u_n) = F^n(u_n)$. Idet alle ekstremværdifordelinger er kontinuerte fordelingsfunktioner, kan udtrykket i ligning (22) på samme måde skrives på formen

$$\lim_{n\to\infty} P(M_n \leq b_n x + a_n) = \lim_{n\to\infty} F^n(b_n x + a_n) = G(x), \quad x \in \mathbb{R}. \tag{23}$$

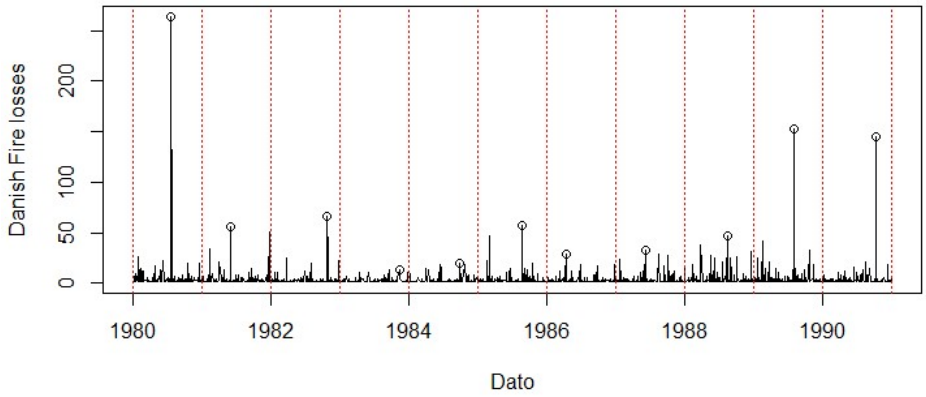
Den relevante teori bag maksima værdier, deres egenskaber og fordeling er nu gennemgået, og det vil i næste afsnit undersøges, hvordan ekstreme værdier kan udvælges og beskrives.

3 Ekstremværdifordelinger

3.1 Blok Maksima

Den første metode der i denne afhandling benyttes til udvælgelse af ekstreme værdier er Blok Maksima (BM) metoden. Denne metode tager som tidligere beskrevet udgangspunkt i at opdele perioden for observationerne i lige store tidsintervaller også kaldet blokke. De ekstreme værdier udgøres af den største observation i hver blok. I figur (2) illustreres, for den teoretiske forståelses skyld, de årlige

ekstreme observationer fra 'Danish Fire Insurance' datasættet fundet ved hjælp af BM metoden, hvilket giver elleve lige store perioder - en periode per år. Man ville dog højst sandsynligt i praksis have valgt mindre tidsintervaller, hvilket havde resulteret i flere observationer.



Figur 2: Ekstreme observationer i 'Danish Fire Insurance' datasættet fundet ved hjælp af BM-metoden med årlig opdeling.

En udfordring ved BM metoden, er valget af blok-størrelse. Store blokke vil generere færre BM værdier, og her kan metoden meget hurtigt komme til ikke at medtage alle relevante høje observationer. Denne udfordring skyldes i nogle tilfælde, at der kan forekomme flere ekstreme værdier inden for samme periode, som for eksempel ved klynger i datasættet. I perioder med generelt lave observationer, kan metoden komme til at definere en lav observation som en BM værdi, hvilket også er tilfældet i figur (2). BM metoden kan også være fordelagtig at benytte. Nogle af fordelene ved at bruge denne metode er jævnfør [13, Ferreira et al., 2014, 2] følgende.

1. BM metoden kan opfange ekstreme værdier, som ellers ville være kasseret ved brug af andre metoder, for eksempel POT metoden.

2. BM metoden er at foretrække, hvis data ikke er i.i.d. Det kan eksempelvis være data med sæsonudsving, hvor det kan være fordelagtigt at studere de ekstreme hændelser opdelt i for eksempel måneder eller årstider.

3. BM metoden er lettere at benytte, da opdelingen af blokke kan være naturlig i mange situationer.

3.2 Den generaliserede ekstremværdifordeling

Det er nu gennemgået, hvordan BM metoden kan benyttes til at finde ekstreme observationer. Disse fundne observationer kan under betingelserne for MDA beskrives ud fra en ekstremværdifordeling $G(x)$. Jævnfør teorien om maks-stabile fordelinger, kan maksima modelleres ud fra GEV familien. I dette afsnit vil GEV fordelingen studeres, og hvis ikke andet er angivet, har afsnittet reference til [20, McNeil et al., 2015, 136-137].

GEV fordelingen er en familie af kontinuerte sandsynlighedsfordelinger, hvis fælles kumulative fordelingsfunktion er givet ved

$$G_\xi(x) = \begin{cases} \exp\left\{-\left[1+\xi x\right]^{-1/\xi}\right\} & ,\xi \neq 0 \\ \exp\left\{-e^{-x}\right\} & ,\xi = 0, \end{cases} \tag{24}$$

hvor restriktionen $\xi = 0$ skal forstås som $\xi \to 0$. Fordelingsfunktionen $G_\xi(x)$ har kun en ukendt parameter: Formparameteren ξ, og en tre-parameter model kan opnås ved at definere $G_{\xi,\mu,\sigma}(x) = G_\xi((x-\mu)/\sigma)$ for lokationsparameter $\mu \in \mathbb{R}$, og skalaparameter $\sigma > 0$. Det betyder, at x i det ovenstående udtryk i ligning (24) erstattes med $(x-\mu)/\sigma$, og fordelingsfunktionen kan opskrives som

$$G_{\xi,\mu,\sigma}(x) = \begin{cases} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} & ,\xi \neq 0 \\ \exp\left(-e^{-\frac{x-\mu}{\sigma}}\right) & ,\xi = 0. \end{cases} \tag{25}$$

Det skal gælde at $1+\xi(x-\mu)/\sigma > 0$, og at parametrene opfylder følgende restriktioner: $-\infty < \mu < \infty$, $\sigma > 0$ og $-\infty < \xi < \infty$. Parameteren ξ er stadig fordelingens formparameter, som bestemmer

formen på fordelingen, og kan antage alle værdier i $\mathbb{R}$. Eksempler på denne parameter er det tredje og fjerde moment i en fordeling, som tidligere beskrevet bestemmer skævhed og kurtosis. Det er netop kurtosis, der er vigtig for at kunne beskrive ekstreme værdier, idet de befinder sig i halen af en fordeling. Grunden til at det er muligt at gå fra en fordeling med kun én parameter til en tre-parameter fordeling, er at det jævnfør teorien om konvergens til typer af fordelinger, er muligt ved hjælp af lokations-og skalaparameteren at udtrykke to typer af fordelinger ud fra hinanden.

Formparameteren ξ styrer halen af fordelingen, og værdien af parameteren indikerer typen af ekstremværdifordelingen. For $\xi = 0$, $\xi > 0$ og $\xi < 0$ er det muligt at opnå tre typer: Gumbel, Fréchet og Weibull fordelingen, hvis fordelingsfunktioner jævnfør [12, Embrechts et al., 2012, 121] kan skrives som i tabel (1).

Fordeling	ξ	Fordelingsfunktion
Gumbel $\Lambda(x)$	$\xi = 0$	$G_{\xi,\mu,\sigma}(x) = exp\left[-\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right], -\infty < x < \infty$
Fréchet $\Phi(x)$	$\xi > 0$	$G_{\xi,\mu,\sigma}(x) = \begin{cases} 0, & x \leq \mu \\ \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-\xi}\right\}, & x > \mu \end{cases}$
Weibull $\Psi(x)$	$\xi < 0$	$G_{\xi,\mu,\sigma}(x) = \begin{cases} \exp\left\{-\left[-\left(\frac{x-\mu}{\sigma}\right)^{\xi}\right]\right\}, & x < \mu \\ 1, & x \geq \mu \end{cases}$

Tabel 1: Ekstremværdifordelinger for henholdsvis $\xi = 0$, $\xi > 0$ og $\xi < 0$.

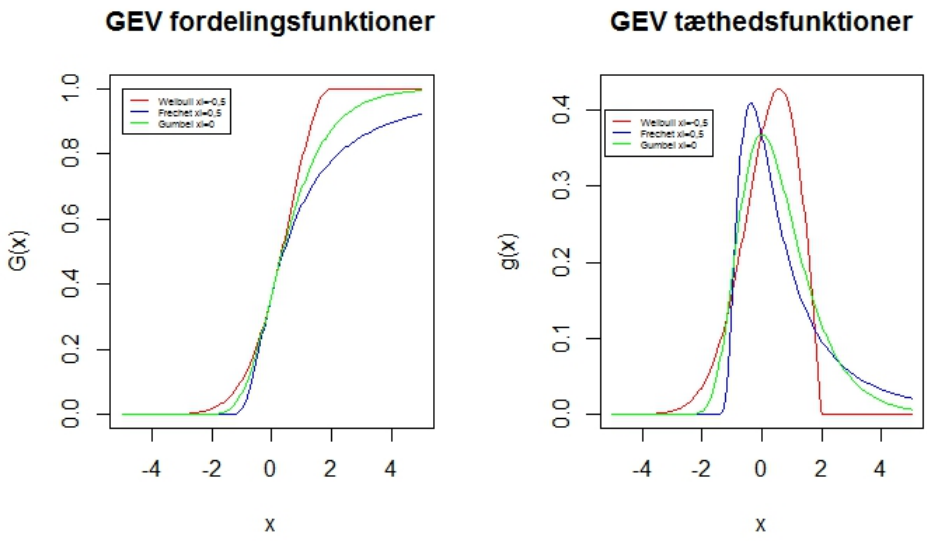
GEV fordelingsfunktionen i ligning (25) kan siges at være generaliseret i den forstand, at den alt efter værdien af formparameteren er en kombination af de tre ovenstående fordelinger i tabel (1).

Tæthedsfunktionen hørende til GEV fordelingen kan findes ved at integrere fordelingsfunktionen i ligning (25), og kan med reference til [19, Markose et al., 2005, 6] skrives som

$$g_{\xi,\mu,\sigma}(x) = \left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}-1} \cdot \exp\left(-\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right)\right), \ \xi \neq 0. \tag{26}$$

Haleadfærden i GEV fordelingen kan nu studeres nærmere ved at illustrere fordelings-og tætheds-

funktionen grafisk, for de tre forskellige værdier af ξ. Funktionerne er illustreret i figur (3).



Figur 3: Fordelings-og tæthedsfunktioner for Gumbel ($\xi = 0$), Fréchet ($\xi = 0,5$) og Weibull ($\xi = -0,5$) fordelingen.

De tre underliggende ekstremværdifordelingers haleadfærd kan studeres ved at betragte tæthedsfunktionernes højre endepunkt x_F. Tæthedsfunktionen for Weibull fordelingen i figur (3), hvor ξ i dette tilfælde antager værdien $-0,5$, har tynde haler og et såkaldt endeligt højre endepunkt. Gumbel fordelingen, hvor ξ antager værdien 0, har et uendeligt højre endepunkt $x_F = \infty$, og det samme er gældende for Fréchet fordelingen, der i dette tilfælde har $\xi = 0,5$. Det bemærkes, at halen i Fréchet fordelingen henfalder langsommere end i Gumbel fordelingen, hvilket indikerer, at de tre typer af fordelinger giver et meget forskelligt billede af haleadfærden af ekstreme værdier.

3.2.1 Udledning af fordelinger

Dette afsnit indeholder en teoretisk gennemgang af de tre underliggende ekstremværdifordelinger: Gumbel, Fréchet og Weibull. For at få en bedre forståelse af de tre typer af ekstremværdifordelinger,

udledes en-parameter fordelingerne ud fra allerede velkendte fordelinger.

Til at starte med studeres Gumbel fordelingen $\Lambda(x)$, hvor det kan vises, at fordelingen kan fremkomme med udgangspunkt i forskriften for en eksponentialfordeling.

Ifølge [12, Embrechts et al., 2012, 125] antages det, at den underliggende fordelingsfunktion er en eksponentialfordeling på formen $F(x) = 1 - e^{-\lambda x}$, som for $n \to \infty$ konvergerer mod en Gumbel fordeling. Hvis det antages, at X_n er en sekvens af i.i.d. stokastiske variable, som følger en eksponentialfordeling med rateparameter $\lambda = 1$, kan fordelingen af maksima opskrives som

$$\begin{aligned} P(M_n \leq x) &= (1 - e^{-x})^n \\ P(M_n - ln(n) \leq x) &= P(M_n \leq x + ln(n))^n = (1 - e^{-(x+\ln(n))})^n \\ &= \left(1 - \tfrac{e}{n}^{-x}\right)^n \\ &\underset{n\to\infty}{\to} e^{-e^{-x}}. \end{aligned} \tag{27}$$

Det vides ud fra teori omkring grænseværdier at $\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x$, hvilket betyder at det ovenstående udtryk i ligning (27) for $n \to \infty$, vil konvergere mod Gumbel fordelingen $\Lambda(x) = e^{-e^{-x}}$.

Fréchet fordelingen $\Phi(x)$ kan ligesom Gumbel fordelingen også beskrives ud fra en anden velkendt fordeling, nemlig Cauchy fordelingen. Med reference til [12, Embrechts et al., 2012, 125] antages igen en sekvens X_n af i.i.d. stokastiske variable, som nu er standard Cauchy fordelte med tæthedsfunktionen

$$f(x) = (\pi(1 + x^2))^{-1} \quad x \in \mathbb{R}.$$

Det ønskes nu at finde overskridelsesfordelingsfunktionen, som kan defineres ved $\overline{F}(x) = 1 - F(x)$. $\overline{F}(x)$ kan også udtrykkes ved hjælp af integralet af tæthedsfunktionen $f(y)$, hvor der integreres fra

en grænse x til ∞:

$$\overline{F}(x) = \int_x^\infty f(y)dy = \frac{1}{\pi}\int_x^\infty \frac{1}{1+y^2}dy \cong \frac{1}{\pi}\int_x^\infty \frac{1}{y^2}\cdot \underbrace{\frac{y^2}{1+y^2}}_{L} dy. \tag{28}$$

Det sidste led i integralet kan tolkes som værende en 'Slowly Varying' funktion L. Ifølge [20, McNeil et al., 2015, 139] er en funktion L i intervallet $(0,\infty)$ 'Slowly Varying' hvis $\lim_{n\to\infty} \frac{L(tx)}{L(x)} = 1,\ t > 0$.

Ud fra Karamata's sætning som er beskrevet i [12, Embrechts et al., 2012, 567], kan integralet $\int_x^\infty t^\alpha L(t)dt$ for $x \to \infty$ hvis L er 'Slowly Varying' og $\alpha > -1$ udtrykkes som

$$\int_x^\infty t^\alpha L(t)dt \backsim (\alpha+1)^{-1}x^{\alpha+1}L(x), \quad x \to \infty. \tag{29}$$

Det betyder at overskridelsesfordelingsfunktionen $\overline{F}(x)$ i ligning (28) i dette tilfælde kan udtrykkes som

$$\overline{F}(x) = \frac{1}{\pi}\int_x^\infty y^{-2}L(y)dy \ \backsim \frac{1}{\pi}(-2+1)^{-1}x^{-2+1}L(x). \tag{30}$$

Ud fra definitionen af en 'Slowly Varying' funktion går grænseværdien for $L(x)$ mod en, og ovenstående udtryk i ligning (30) kan skrives som $\overline{F}(x) \backsim (\pi x)^{-1}$. Fordelingen af maksima værdierne kan dermed opskrives som følgende

$$\begin{aligned} P(M_n \le x_n) &= (1-\overline{F}(x_n))^n \\ P(M_n \le \tfrac{nx}{\pi}) &= (1-\overline{F}(\tfrac{nx}{\pi}))^n \\ &= (1-\tfrac{1}{n}(\tfrac{1}{x}+o(1))^n \\ &\underset{n\to\infty}{\to} e^{-1/x} = \Phi_1(x) \quad x>0, \end{aligned} \tag{31}$$

hvor x_n sættes lig med nx/π. Hvis $n \to \infty$ konvergerer fordelingen for maksima hvor $\xi = 1$ mod Fréchet fordelingen $\Phi_1(x)$, for alle positive værdier af x.

Der findes ligesom for Gumbel og Fréchet fordelingen også en fordeling, som leder til den sid-

ste af de tre ekstremværdifordelinger: Weibull fordelingen. Med inspiration fra [7, Coles, 2001, 52] antages X_n i.i.d. stokastiske variable, men denne gang følger de en uniform fordeling $\mathcal{U}(a,b)$, hvor fordelingsfunktionen er givet ved

$$F(x) = \begin{cases} 0 & for\ x \leq a \\ \frac{x-a}{b-a} & for\ x \in (a,b) \\ 1 & for\ x \geq b. \end{cases} \tag{32}$$

I denne udledning studeres $\mathcal{U}(0,1)$, og fordelingen af maksima kan opskrives som

$$\begin{aligned} P(M_n \leq x) &= x^n \\ P(n(M_n - 1) \leq x) &= (1 + \tfrac{x}{n})^n \underset{(n\to\infty)}{\to} e^x \end{aligned} \tag{33}$$

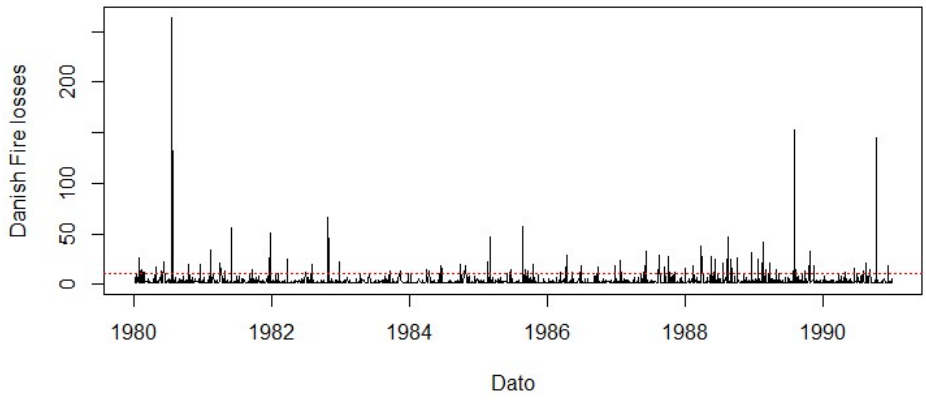
Det første udtryk i ligningssystemet er fordelingen af maksima for $x \in (0,1)$, som findes ved at substituere værdierne for a og b ind i forskriften for $F(x)$ i ligning (32). I det andet udtryk er normaliseringskonstanterne, $a_n = 1$ og $b_n = 1/n$, substitueret ind, og for $n \to \infty$ fås forskriften for den sidste af de tre ekstremværdifordelinger, Weibull fordelingen $\psi_1(x)$.

Vi har nu gennemgået fordelingerne hørende til de ekstreme observationer fundet ud fra BM metoden. Som beskrevet er BM metoden ikke den eneste metode til at udvælge ekstreme observationer, hvorfor der i næste afsnit studeres en anden metode: Peaks-over-Threshold.

3.3 Peaks-over-Threshold metoden

I Peaks-over-Threshold (POT) metoden udvælges, jævnfør [13, Ferreira et al., 2014, 1-3], de observationer, som er større end et givet threshold u, og de udgør de ekstreme observationer. De ekstreme observationer udvalgt ved POT metoden er illustreret i figur (4), hvor der i 'Danish Fire Insurance' datasættet er fastsat et threshold på $u = 10$. De observationer, som ligger over denne thresholdværdi,

udgør de ekstreme observationer i datasættet.



Figur 4: Ekstreme observationer i 'Danish Fire Insurance' datasættet fundet ved hjælp af POT-metoden.

Der kan for observationer fundet ved hjælp af POT metoden fittes en fordeling af haleobservationerne, som kan approksimeres ved hjælp af den generaliserede paretofordeling (GPD). Modsat BM metoden er det altså ikke maksima observationen i en givet blok som studeres, men haleobservationerne over en given thresholdværdi. Det betyder, at POT metoden kan have udfordringer forbundet med data med for eksempel sæsonudsving, da grænsen u er konstant. Det kan derfor være nyttigt at tage højde for eventuel sæsonkorrigering af data, inden denne metode benyttes. Har datasættet tendens til klyngedannelse, vil POT metoden være at foretrække frem for BM metoden, idet den har mulighed for at opfange flere ekstreme observationer. Ved brug af POT metoden i praksis er det dog vigtigt at overveje sit valg af threshold u grundigt, hvilket vil blive gennemgået senere i dette teori afsnit. Først gennemgås teorien omkring GPD'en.

3.4 Den generaliserede paretofordeling

Vi har indtil nu gennemgået BM metoden, hvor de udvalgte ekstreme observationer fittes til en GEV fordeling. Ekstreme observationer er som tidligere nævnt fordelt på forskellige måder, alt efter hvordan de udvælges. Studeres data som er bestemt ud fra POT metoden, er GPD'en den mest

korrekte fordeling at fitte. GPD'en vil dermed i dette afsnit blive gennemgået, og er skrevet med inspiration fra [12, Embrechts et al., 2012, 6.5.1] og [20, McNeil et al., 2015, 146-149].

Til at starte med studeres den ukendte fordelingsfunktion F, som udgøres af en sekvens af i.i.d. stokastiske variable $X_1, X_2, .., X_n$, hvor det nu ønskes at estimere overskridelsesfordelingsfunktionen F_u for værdier af x, som ligger over et threshold u. Denne overskridelsesfordelingsfunktion F_u kaldes også for den betingede excess fordeling, og er defineret ved

$$F_u(x) = P(X - u \leq x \mid X > u), \quad 0 \leq x \leq x_F - u, \tag{34}$$

hvor $x_F \leq \infty$ betegner det højre endepunkt af en fordeling F. $F_u(x)$ kan altså beskrives som sandsynligheden for, at et ekstremt tab af størrelsen $(X - u)$ er mindre end eller lig med en fraktil x, givet at threshold værdien u er overskredet.

En anden måde hvorpå F_u kan udtrykkes, som er fordelagtig til videre brug, er at opskrive F_u ved hjælp af en fordelingsfunktion F:

$$F_u(x) = \frac{F(u + x) - F(u)}{1 - F(u)}, \quad y > 0. \tag{35}$$

hvor $F(u+x)$ angiver sandsynlighedsmassen som ligger mellem thresholdværdien og en given fraktil x.

I fordelingsteori studeres ofte stokastiske variable X, som ligger centreret i en sandsynlighedsfordeling, hvor der ikke vil være udfordringer forbundet med at estimere en fordeling F. Det vanskelige er at estimere F_u, som kun består af ekstreme værdier over et threshold u, da den udgør en anden fordeling end den oprindelige dog med færre observationer.

Hvis fordelingen F er kendt, skulle man tro, at F_u direkte kunne findes derud fra, men det er ikke tilfældet, da det er en grænsefordeling der ønskes estimeret. Det svarer til at anvende GEV fordelingen, som en tilnærmelse til fordelingen af observationer fundet ved hjælp af BM metoden,

når fordelingen F er ukendt.

Når POT metoden benyttes til udvælgelse af ekstreme observationer, fittes disse haleobservationer til GPD'en. I denne afhandling er GPD'en et udtryk for en to-parameter fordeling, bestående af en formparameter ξ, og en skalaparameter σ. Den tilhørende fordelingsfunktion kan opskrives som

$$G_{\xi,\sigma}(x) = \begin{cases} 1 - (1 + \frac{\xi x}{\sigma})^{1/\xi} & ,\xi \neq 0 \\ 1 - \exp(-\frac{x}{\sigma}) & ,\xi = 0 \ , \end{cases} \tag{36}$$

hvor $\sigma > 0$, $x \geq 0$ når $\xi \geq 0$ og $0 \leq x < -\frac{\sigma}{\xi}$ når $\xi < 0$. Restriktionen $\xi = 0$ skal ligesom i GEV fordelingen forstås som $\xi \to 0$. Differentieres udtrykket i ligning (36) opnås den tilhørende tæthedsfunktion, som med reference til [8, Marcelo et al., 2015, 848] kan udtrykkes som

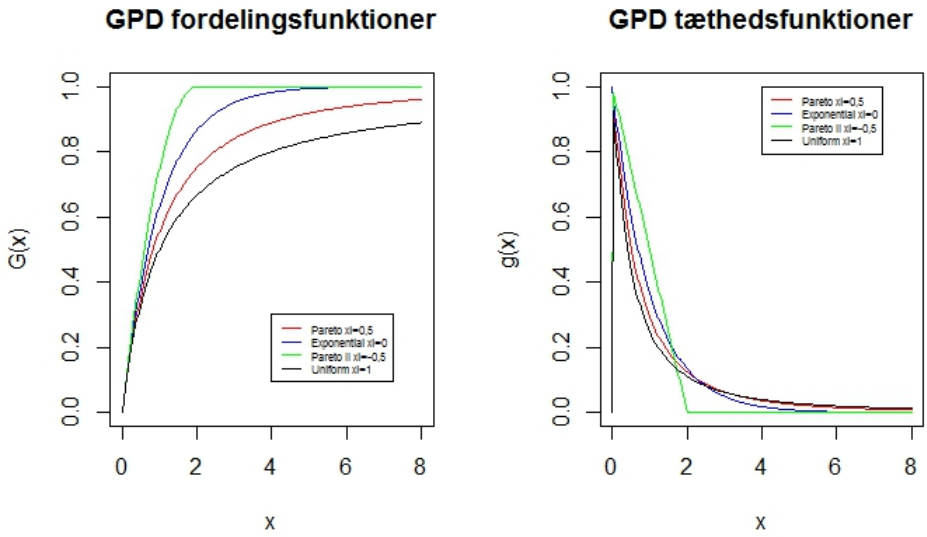
$$g_{\xi,\sigma}(x) = \begin{cases} \frac{1}{\sigma}\left(1 - \frac{\xi x}{\sigma}\right)^{1/\xi - 1} & ,\xi \neq 0 \\ \frac{1}{\sigma}\exp\left(-\frac{x}{\sigma}\right) & ,\xi = 0. \end{cases} \tag{37}$$

På samme måde som GEV fordelingen er GPD'en generaliseret i den forstand, at den afhængigt af værdien af formparameteren ξ, består af flere seperate sandsynlighedsfordelinger, som tilhører GPD familien:

- $\xi = 0$: Eksponentialfordeling med middelværdi σ.
- $\xi = 1$: Uniform fordeling $U[0, \sigma]$.
- $\xi > 0$: Ordinær paretofordeling med $\alpha = 1/\xi$ og $\kappa = \sigma/\xi$.
- $\xi < 0$: Pareto type II fordeling med endeligt højre endepunkt og parameter ξ.

I figur (5) er fordelings-og tæthedsfunktionen hørende til GPD'en for forskellige værdier af formparameteren ξ illustreret, hvor skalaparameteren σ holdes konstant, og antager værdien en. Det

observeres, at fordelingen for negative værdier af ξ har et endeligt højre endepunkt, og jo højere værdier parameteren ξ antager, desto federe haler har GPD'en.



Figur 5: Fordelings-og tæthedsfunktioner for GPD'en for $\xi = -0,5,\ \xi = 0,\ \xi = 0,5$ og $\xi = 1$.

Disse sandsynlighedsfordelinger i GPD familien er de eneste kontinuerte fordelingsfunktioner, der er stabile i forhold til modellering af overskridelsesobservationer i EVT'en, det vil sige fordelingerne er POT-stabile. Det, at GPD'en er POT-stabil, svarer til, at GEV fordelingen på baggrund af maksima værdier, er maks-stabil.

I forhold til MDA egenskaben er GPD'en i MDA af GEV fordelingen, hvilket udtrykkes som

$$G_{\xi,\sigma} \in MDA(G_\xi)\ \forall\ \xi \in \mathbb{R}. \tag{38}$$

Det betyder, at det for alle værdier af ξ er muligt at udtrykke GEV fordelingen og GPD'en ud fra hinanden.

Argumentet for at overskridelsesobservationerne netop skal fittes til GPD'en, er defineret ud fra

Pickands-Balkema-de-Haans sætning i [20, McNeil et al.,2015, 149], som siger

For ethvert $\xi \in \mathbb{R}$ da er $F \in MDA(G_\xi)$ hvis og kun hvis

$$\lim_{u \to x_F} \sup_{0 \leq x < x_F - u} | F_u(x) - G_{\xi,\sigma(u)}(x) | = 0, \tag{39}$$

for nogle positive funktioner σ.

Udtrykket i ligning (39) siger, at når u går mod det højre endepunkt, skal det gælde, at for det første x efter u skal forskellen mellem overskridelsesfordelingen og GPD'en tilnærmelsesvis være nul.

Ud fra det generelle udtryk for overskridelsesfordelingen F_u i ligning (35) og fordelingsfunktionen for GPD'en i ligning (36) kan det vises, at GPD'en er den korrekte fordeling at benytte til at modellere ekstreme observationer over et threshold u. Indsættes fordelingsfunktionen for henholdsvis $F(u+x)$ og $F(u)$ i udtrykket for overskridelsesfordelingen i ligning (35) fås

$$\begin{aligned} F_u(x) &= \frac{F(x+u)-F(u)}{1-F(u)} &&= \frac{\left(1-(1+\xi(x+u)/\sigma)^{-1/\xi}\right)-\left(1-(1+\xi u/\sigma)^{-1/\xi}\right)}{1-\left(1-(1+\xi u/\sigma)^{-1/\xi}\right)} \\ &= 1-\left(\frac{1+\xi x/\sigma+\xi u/\sigma}{1+\xi u/\sigma}\right)^{-1/\xi} &&= 1-\left(1+\frac{\xi x}{\sigma+\xi u}\right)^{-1/\xi} \\ &= G_{\xi,\sigma(u)}(x)\,. \end{aligned} \tag{40}$$

hvor $\sigma(u) = \sigma + \xi u$. Det er herved vist at $F_u(x) = G_{\xi,\sigma(u)}(x)$, hvormed Pickands-Balkema-de-Haans sætning i ligning (39) er opfyldt, og GPD'en vil fremadrettet benyttes til modellering af overskridelsesobservationer.

3.5 Bestemmelse af threshold

Der kan være udfordringer forbundet med at finde et passende threshold, når POT metoden benyttes. For kun at opfange ekstreme observationer ønskes det at sætte grænsen u så højt som muligt, men hvis grænsen sættes for højt vil antallet af overskridelser være få, og estimaterne af parametrene i

fordelingen kan ende med at have en høj varians. Hvis thresholdværdien derimod sættes for lavt, vil udfordringen med for få observationer være løst, men det kan betyde, at nogle af observationerne vil være for lave til at kunne betegnes som ekstremværdier. Det vil sige, jo flere ikke ekstreme observationer der medtages, jo højere bias for parameterestimaterne. Parameterestimaterne afviger altså fra de 'sande' parametre.

Ifølge [12, Embrechts et al., 2012, 356] kan det ikke forventes, at der findes en unik løsning til hvor grænsen skal sættes, og dermed heller ikke et unikt valg af et threshold u. Det foreslås at benytte forskellige plots, og være kritisk overfor data, samt at bruge sin sunde fornuft. I denne afhandling tages der udgangspunkt i et plot til fastsættelse af threshold værdien u: Mean Excess (ME) plottet. Derefter benyttes en Goodness-of-Fit test, hvor model fits for de valgte thresholdværdier sammenlignes og valget af thresholdværdien, ud fra ME plottet, kan valideres.

Det nedenstående afsnit omhandlende ME plottet er skrevet med inspiration fra [12, Embrechts et al., 2012, 355].

3.5.1 Mean Excess Plot

Det kan antages at data, der overskrider en given thresholdværdi er GP fordelt, hvorfor ME plottet tager udgangspunkt i ME funktionen for GPD'en. ME funktionen kan opskrives ud fra middelværdien af de GP fordelte variable, som jævnfør [12, Embrechts et al., 2012] kan skrives på følgende måde

$$e(u) = E(X - u|X > u) = \frac{\sigma(u)}{1-\xi} = \frac{\sigma + \xi u}{1-\xi}. \tag{41}$$

Det vil sige, at der tages udgangspunkt i middelværdien af en standard GPD, altså middelværdien af de observationer, der har overskredet et threshold u. Den betingede excess fordeling F_u forbliver dermed en GPD med den samme formparameter ξ, men med en skalaparameter der vokser lineært med thresholdværdien u: $\sigma(u) = \sigma + \xi u$. Funktionen $e(u)$ vil på grund af skalaparameteren dermed også være lineær.

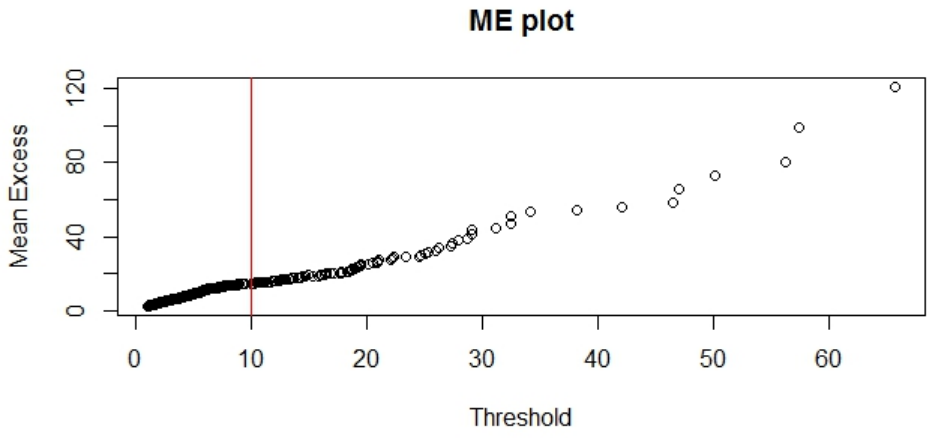
ME funktionen er lineær hvis og kun hvis, observationerne er GP fordelte med parametre $\sigma(u)$ og ξ, og med denne antagelse benyttes den empiriske ME funktion. Da det er tilfældet, at overskridelsesobservationerne kan antages at være GP fordelte, opskrives den empiriske ME funktion som

$$e_n(u) = \frac{\sum_{i=1}^{n}(X_i - u) \cdot 1_{\{X_i > u\}}}{\sum_{i=1}^{n} 1_{\{X_i > u\}}} = \frac{1}{N_u} \sum_{i \in (X_i > u)} (X_i - u). \tag{42}$$

Udtrykket i ligning (42) siger, at summen af overskridelsesværdierne divideret med antallet af observationer N_u som ligger over thresholdværdien u, også kan udtrykkes som middelværdien af overskridelserne. ME plottet kan nu fremkomme ved at plotte den empiriske ME funktion i forhold til forskellige thresholdværdier u, og kan jævnfør [7, Coles, 2001, 78] skrives som

$$\{(u, e_n(u)), : u < X_{\max}\}, \tag{43}$$

hvor $X_{\max}$ er den største værdi af de X_i observationer.

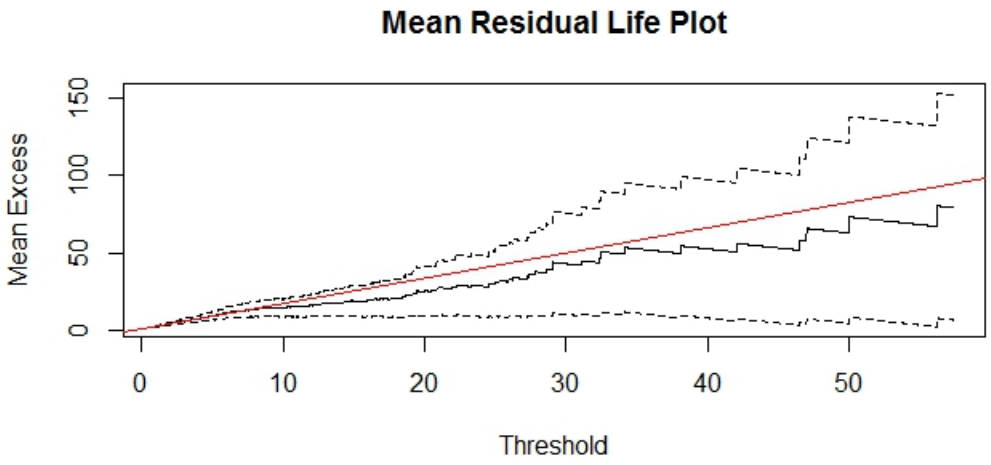


Figur 6: ME plot for 'Danish Fire Insurance' data, med $u = 10$.

ME plottet er som beskrevet tilnærmelsesvis lineært, hvis observationerne er GP fordelte og værdien af formparameteren ξ afhænger af, hvorvidt trenden i plottet er opadgående, nedadgående eller horisontal. Med reference til [20, McNeil et al., 2015, 151] vil et plot med en lineær opadgående eller nedadgående trend have en henholdsvis positiv og negativ formparameter ξ. Har plottet en horisontal trend, vil formparameteren være tilnærmelsesvis lig med nul. Da plottet sjældent er perfekt lineært, ligger udfordringen i at aflæse ME plottet korrekt, og det gør selve bestemmelsen af et treshold forholdsvis svært. Det er ofte i den højre side af plottet, hvor gennemsnittet bygger på et mindre antal høje overskridelser, at den lineære trend ikke er så synlig. Det kan derfor være fordelagtigt at fjerne nogle af disse observationer, så man kan tydeliggøre plottet, som thresholdværdien skal vælges på baggrund af.

I [12, Embrechts et al., 2012, 355] er valget af threshold formuleret som det $u > 0$, hvor ME plottet er tilnærmelsesvis lineært for værdier af $x \geq u$. Det betyder, at vi skal vælge et threshold u, som den værdi, der ligger i begyndelsen af den lineære del af plottet. I figur (6), vises ME plottet for 'Danish Fire Insurance' datasættet, og der anes en lineær tendens igennem hele plottet. Denne opadgående lineære trend giver en forventning om, at datasættet kan fittes til en GPD med positiv formparameter ξ. Der er et lille knæk i plottet ved $u = 10$, hvilket indikerer, at et threshold kan sættes til denne værdi, og de observationer som ligger over dette threshold, vil udgøre de ekstreme observationer. I figur (6) observeres det, at der er få observationer i den ydre højre side af plottet, hvorved det her kunne være rimeligt at fjerne nogle af punkterne og opnå et tydeligere og mere let aflæseligt plot.

Mean Residual Life (MRL) plottet som er illustreret i figur (7), kan, med reference til [10, De Silva,2006,26], også benyttes til valg af threshold. I plottet fittes en lineær regressions linje for det valgte threshold til plottet, og hvis linjen ligger inden for 95 % konfidensintervallet, kan det konkluderes, at den pågældende thresholdværdi på et 95% niveau er et rimeligt valg.



Figur 7: MRL plot for 'Danish Fire Insurance' data med regressionslinje for $u = 10$.

MRL plottet er teoretisk det samme som ME plottet, men grafisk ser det lidt anderledes ud. ME plottet er et mere simpelt og lettere aflæseligt plot, hvorimod MRL plottet viser, hvorvidt det valgte threshold er et muligt valg.

Det er dog stadig ikke let at aflæse disse plots, da der ikke direkte er nogen værdier eller resultater der understøtter valget. Der vil derfor i analyse afsnittet benyttes en Bootstrap Goodness-of-Fit test, til at understøtte eller justere valget af thresholdværdien u. Denne metode er benyttet med udgangspunkt i [3, Alva et al., 2009], hvor der testes for, hvor godt overskridelsesobservationerne fitter GPD'en.

4 Estimation

Vi har nu gennemgået teorien bag ekstreme observationer samt to forskellige ekstremværdifordelinger, og ønsker i dette afsnit at gennemgå estimationen af GPD'ens ukendte parametre, form-og skalaparameteren (ξ ,σ) samt de tilhørende risikomål. Dette afsnit er, hvis ikke andet er angivet, skrevet med udgangspunkt i artiklen [4, Castillo et al., 1997].

I denne afhandling vil følgende fem estimationsmetoder blive studeret: Method-of-Moments (MOM), Probability-Weighted-Moments (PWM), Elemental-Percentile-Method (EPM), L-Moments-Method (LMOM) og Maksimum Likelihood Estimation (MLE) som hver især tager udgangspunkt i forskellige matematiske metoder.

4.1 Method-of-Moments

Method-of-Moments (MOM) estimationsmetoden tager udgangspunkt i de såkaldte momentbetingelser, hvor ideen er, at de teoretiske momenter i fordelingen sættes lig med sample momenterne. De teoretiske momenter udtrykkes, som det forventede moment og angiver derfor det 'sande' moment, mens samplemomenterne kan udtrykkes som det empiriske moment. Udledningen af MOM estimaterne har reference til [1, PennState Eberly College of Science].

Idet GPD'en er en to-parameter fordeling, er det kun nødvendigt at studere det første og andet moment, som er et udtryk for henholdsvis middelværdi og varians. Det er netop disse momenter som form-og skalaparameteren bliver fundet ud fra. Det antages, at $X_1, X_2, ..., X_n$ er GP fordelte stokastiske variable med parametre ξ og σ, og at middelværdien for GPD'en ud fra artiklen [11, Bermudez et al., 2009, 1355] kan opskrives som

$$E(X) = \frac{\sigma}{1+\xi} \ , \ \xi > -1. \tag{44}$$

Det første sample moment $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$, findes ud fra sekvensen af de stokastiske variable, og sættes lig med det første teoretiske moment $E(X)$

$$E(X) = \sigma/(1+\xi) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}. \tag{45}$$

Dernæst sættes det andet sample moment $M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i - X_i^2$ på samme måde lig med det tilsvarende teoretiske moment $E(X_i - \mu)^2$:

$$E(X_i - \mu)^2 = \frac{\sigma^2}{(1+\xi)^2(1+2\xi)} = \frac{1}{2}\sum_{i=1}^{n}(X_i - \overline{X})^2, \tag{46}$$

som også er et udtryk for variansen af de stokastiske variable. Udtrykkene i ligning (45) og (46) kaldes for momentbetingelserne, og ved at løse disse ligninger med hensyn til form-og skalaparameteren kan der findes et udtryk for parameterestimaterne. Isoleres skalaparameteren σ i ligning (45), fås $\sigma = \overline{X} \cdot (1+\xi)$, og MOM estimatet for formparameteren ξ findes ved at substituere udtrykket for σ ind i udtrykket for det andet moment i ligning (46)

$$\begin{array}{rcccl} E(X_i-\mu)^2 & = & \frac{(\overline{X}\cdot(1+\xi))^2}{(1+\xi)^2(1+2\xi)} & \Leftrightarrow & \frac{\overline{X}^2}{(1+2\xi)} \\ & \Rightarrow & s^2(1+2\xi) = \overline{X}^2 & \Leftrightarrow & s^2 2\xi = \overline{X}^2 - s^2 \\ & \Leftrightarrow & \xi = \frac{\overline{X}^2 - s^2}{2s^2} & \Rightarrow & \xi = \frac{1}{2}\left(\frac{\overline{X}^2}{s^2} - 1\right). \end{array} \tag{47}$$

MOM estimatet for skalaparameteren σ som kun er udtrykt ved sample middelværdien og variansen, findes ved at substituere det fundne ξ i ligning (47) ind i udtrykket for det første moment i ligning (45):

$$\begin{array}{rcccl} \sigma = \overline{X}\cdot\left(1+\frac{1}{2}\left(\frac{\overline{X}^2}{s^2}-1\right)\right) & \Leftrightarrow & \overline{X} + \frac{1}{2}\overline{X}\left(\frac{\overline{X}^2}{s^2}-1\right) & \Leftrightarrow & \overline{X} + \frac{1}{2}\overline{X}\frac{\overline{X}^2}{s^2} - \frac{1}{2}\overline{X} \\ & & & \Rightarrow & \sigma = \frac{1}{2}\overline{X}\left(\frac{\overline{X}^2}{s^2}+1\right). \end{array} \tag{48}$$

Ved hjælp af MOM estimationsmetoden kan parameterestimaterne i GPD'en dermed beregnes ud fra følgende lukkede formler

$$\xi_{MOM} = \frac{1}{2}\left(\frac{\overline{X}^2}{s^2} - 1\right) \quad \text{og} \quad \sigma_{MOM} = \frac{1}{2}\overline{X}\left(\frac{\overline{X}^2}{s^2}+1\right), \tag{49}$$

hvor $\overline{X}$ og s^2 er henholdsvis middelværdien og variansen af samplen. Den næste estimationsmetode der vil blive gennemgået, er Elemental-Percentile-Method (EPM).

4.2 Elemental-Percentile-Method

Elemental-Percentile-Method (EPM) er en anden estimationsmetode til at fitte data til GPD'en, hvor metoden modsat MOM, som benytter lukkede formler, bruger en numerisk tilgang. Ideen bag EPM metoden er at matche den teoretiske fordeling med den empiriske fordeling. Parametrene i EPM'en opnås ved at starte med at lave en reparametrisering af GPD'en, hvor δ substitueres ind på σ/ξ plads i fordelingsfunktionen. Det betyder, at GPD'ens kumulative fordelingsfunktion for $\xi \neq 0$ kan skrives på formen

$$F(x) = 1 - (1 - x/\delta)^{1/\xi}, \qquad \xi \neq 0 \quad \text{og} \quad \delta\xi > 0. \tag{50}$$

Metodens procedure er inddelt i to trin, hvor den starter med, ved hjælp af en algoritme, at beregne et antal initial estimater for parametrene ξ og σ, og herefter findes de endelige parameterestimater ud fra disse initial estimater.

I det følgende gennemgås først en teoretisk udledning af den generelle procedure, hvorefter vi har illustreret selve algoritmen, som i analysen implementeres i `R`.

I det første trin sættes den fundne fordelingsfunktion $F(x)$ fra ligning (50) lig med den empiriske fordeling p:

$$F(x_{i:n}) = p_{i:n} \quad \text{og} \quad F(x_{j:n}) = p_{j:n}, \tag{51}$$

hvor $x_{i:n}$ og $x_{j:n}$ er den i'te og j'te observation i et sorteret datasæt med størrelsen n, og $p_{i:n} = i/(n+1)$. Ved at substituere ligning (50) ind i udtrykket for $F(x_{i:n})$ i ligning (51), fås

$$F(x_{i;n}) = p_{i:n} \Rightarrow 1 - (1 - x_{i:n}/\delta)^{1/\xi} = p_{i:n} \Rightarrow 1 - x_{i:n}/\delta = (1 - p_{i:n})^{\xi}, \tag{52}$$

hvor det samme gør sig gældende for $F(x_{j:n})$. Tages logaritmen på begge sider af lighedstegnet, opnås følgende udtryk for både det i'te og j'te element af x

$$\ln(1 - x_{i:n}/\delta) = \xi \cdot C_i \quad \text{og} \quad \ln(1 - x_{j:n}/\delta) = \xi \cdot C_j, \tag{53}$$

hvor $C_i = \ln(1 - p_{i:n})$ og $C_j = \ln(1 - p_{j:n})$.

Det næste trin er nu at løse ligningerne i ligning (53) for parametrene ξ og δ. Først isoleres ξ i ligningerne, og udtrykkene sættes lig med hinanden:

$$\begin{aligned} \ln(1 - x_{i:n}/\delta)/C_i &= \ln(1 - x_{j:n}/\delta)/C_j \quad \Rightarrow \\ C_j \cdot \ln(1 - x_{i:n}/\delta) &= C_i \cdot \ln(1 - x_{j:n}/\delta). \end{aligned} \tag{54}$$

Dernæst isoleres δ i ligningerne i (53), og de sættes lig med hinanden :

$$\begin{aligned} [1 - (1 - p_{i:n})^{\xi}]/x_{i:n} &= [1 - (1 - p_{i:n})^{\xi}]/x_{j:n} \quad \Rightarrow \\ x_{j:n} \cdot [1 - (1 - p_{i:n})^{\xi}] &= x_{i:n} \cdot [1 - (1 - p_{i:n})^{\xi}]. \end{aligned} \tag{55}$$

Det ønskes at finde en løsning til ligning (54) og (55), og jævnfør artikel [4, Castillo et al., 1997, 1611-1612] følger det, at ligning (54), som er en funktion af kun en variabel, δ, har en endelig løsning $\hat{\delta}(i, j)$, som kan findes ved brug af Bisection metoden[2]. Denne metode er numerisk, og benyttes til at finde en rod for en given funktion i et givet interval. Den endelige løsning for ligning (54) i intervallet $(\delta_0, 0)$ for $x_{i:n} < C_i x_{j:n}/C_j$ eller i intervallet $(x_{j:n}, \delta_0)$ for $x_{i:n} > C_i x_{j:n}/C_j$ er dermed givet ved

$$\delta_0 = \frac{x_{i:n} x_{j:n}(C_j - C_i)}{C_j x_{i:n} - C_i x_{j:n}}. \tag{56}$$

Løsningen med hensyn til δ i ligning (54) kan også skrives som $\hat{\delta}(i, j)$. Substitueres dette estimatet for $\hat{\delta}(i, j)$ ind i ligning (53), kan det tilsvarende estimat for ξ, der skrives som $\hat{\xi}(i, j)$, findes til at være

[2]Hvis vi har en funktion $f(x)$, som er kontinuert over et interval $[a, b]$ og $f(a) \neq f(b)$, findes der en værdi $c \in [a, b]$ så $f(c) = 0$, og det betyder, at c er en rod i intervallet $[a, b]$.

$$\ln(1 - x_{i:n}/\delta) = \xi \cdot C_i \quad \Rightarrow \quad \hat{\xi}(i,j) = \frac{\ln(1 - x_{i:n}/\hat{\delta}(i,j))}{C_i}. \tag{57}$$

Der er nu fundet et estimat for formparameteren ξ i GPD'en. Estimatet hørende til skalaparameteren σ, kan i EPM metoden findes som produktet af parameterestimaterne for δ og ξ:

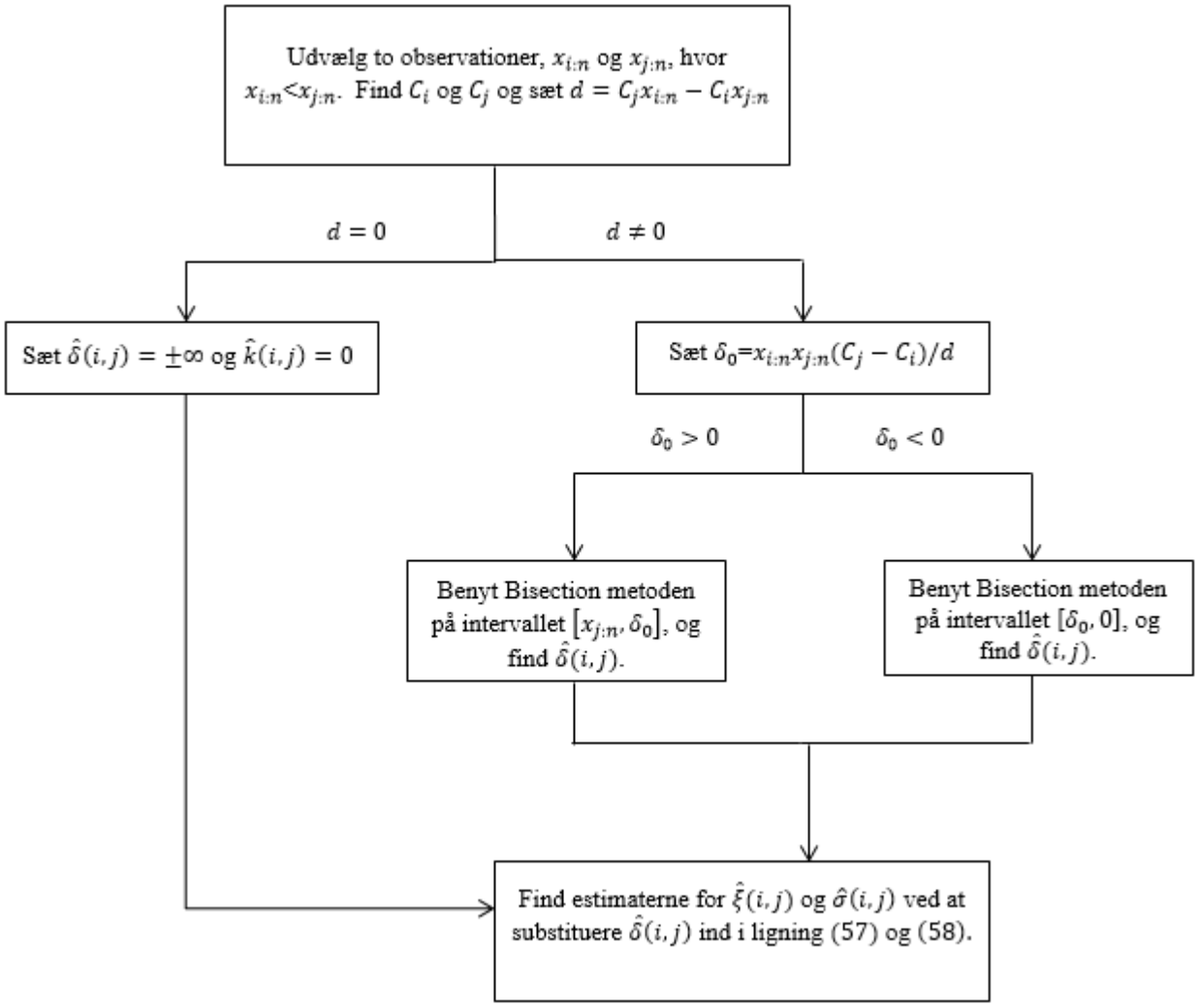
$$\hat{\sigma}(i,j) = \hat{\delta}(i,j)\hat{\xi}(i,j). \tag{58}$$

EPM metoden benytter en numerisk tilgang, hvor selve beregningen af de initiale parameterestimater følger en fem-trins algoritme, som er beskrevet i [4, Castillo et al., 1997, 1612, Algorithm 1]. Denne algoritme kan nemt virke uoverskuelig, og vi har derfor, for overskuelighedens og læserens skyld, illustreret den grafisk i figur (8).

Ud fra algoritmen er der nu fundet parameterestimater, men disse fundne estimater $\hat{\xi}(i,j)$ og $\hat{\sigma}(i,j)$ er kun baseret på to ordnede observationer $x_{i:n}$ og $x_{j:n}$. Det næste trin er derfor at beregne estimaterne på baggrund af alle mulige værdier af i og j, hvorudfra de endelige estimater kan findes som medianen af estimaterne af alle mulige kombinationer af i og j:

$$\begin{aligned} \hat{\xi}_{EPM} &= \text{median}\left(\hat{\xi}(1,2), \hat{\xi}(1,3), ..., \hat{\xi}(n-1,n)\right) \\ \hat{\sigma}_{EPM} &= \text{median}\left(\hat{\sigma}(1,2), \hat{\sigma}(1,3), ..., \hat{\sigma}(n-1)\right). \end{aligned} \tag{59}$$

En udfordring ved denne estimationsmetode er antallet observationer, som er påvirket af det specifikke datasæt. Det kan risikeres, at n er meget stor, hvilket vil medføre et stort antal par (i,j), og metoden kan dermed hurtigt blive beregningstung. En mulig løsning kan være at antage at $i = 1, 2, .., n-1$ og $j = n$ fremfor at tage udgangspunkt i alle mulige par (i,j).



Figur 8: Træ over EPM metodens algoritme til beregning af estimater.

4.3 Probability-Weighted-Moments

Den følgende teori omkring Probability-Weighted-Moments (PWM) metoden har gennemgående reference til [11, Bermudez et al., 2009, 1357].

PWM estimationsmetoden tager udgangspunkt i momenter af typen $E[X(1-F)^s]$, eller alternativt $E[XF^r]$, hvor s og r, alt efter hvilken type af momenter der benyttes, antager værdierne 0, 1, 2 eller større, afhængigt af antallet af parametre, der ønskes estimeret. I PWM metoden kan man, ud fra momenter og den pågældende fordelingsfunktion, danne antallet af ligninger, hørende til antallet af parametre der ønskes estimeret, hvor der i dette tilfælde tages udgangspunkt i GPD'en. Ved løsning af ligningssystemet opnås et lukket udtryk for hver parameter.

Momenterne i PWM metoden kan generelt opskrives som

$$M_{p,r,s} = E[X^p F^r (1-F)^s], \tag{60}$$

hvor X er en stokastisk variabel, og $F = P(X \leq x)$ er den tilhørende kumulative fordelingsfunktion. Der vil i fordelingsfunktionen F findes k antal ukendte parametre $\phi_1, \phi_2, ..., \phi_k$, som ønskes estimeret. I ligning (60) angiver p, r og s reelle tal, hvor der jævnfør artikel [11, Bermudez et al., 2009, 1357] i denne afhandling benyttes $p = 1$. Denne værdi af p medfører, at PWM'en afhænger direkte af observationerne X, lige meget hvilken type moment man vælger at benytte. De mest benyttede PWM'er er $M_{1,0,s}$ eller $M_{1,r,0}$, hvilke kan udtrykkes som α_s og β_r på følgende måde

$$\alpha_s = M_{1,0,s} = E[X^p(1-F)^s] \qquad \text{og} \qquad \beta_r = M_{1,r,0} = E[X^p F^r]. \tag{61}$$

Det kan til videre brug, i LMOM metoden, være fordelagtigt ikke at udtrykke momenterne ud fra den forventede værdi, men opskrive dem ved hjælp af fraktilfunktionen $x(F)$ og fordelingsfunktionen F. I stedet for at benytte forventningen som i ligning (61) udtrykkes momenterne ved hjælp af et integrale, hvilket kan skrives som

$$\alpha_s = \int_0^1 x(F)^p(1-F)^s dF \qquad \text{og} \qquad \beta_r = \int_0^1 x(F)^p F^r dF. \tag{62}$$

Momenterne $\{\alpha_s : s = 0, 1, 2, ...\}$ og $\{\beta_r : r = 0, 1, 2...\}$ kan opskrives som linearkombinationer af hinanden, og afhænger derfor af hinanden ud fra følgende relationer

$$M_{p,0,s} = \sum_{r=0}^{s} \begin{pmatrix} s \\ r \end{pmatrix} (-1)^r M_{p,r,0} \qquad \text{og} \qquad M_{p,r,0} = \sum_{s=0}^{r} \begin{pmatrix} r \\ s \end{pmatrix} (-1)^s M_{p,0,s}, \tag{63}$$

Idet momenterne er linearkombinationer af hinanden, benyttes kun en af relationerne i ligning (63), nemlig α_s, men samme metode og beregninger ville være gældende for β_r.

Fordelingsfunktionen hørende til GPD'en med formparameter $\xi \neq 0$, og skalaparameter σ er givet ved

$$G_{\xi,\sigma}(x) = 1 - \left(1 - \frac{\xi x}{\sigma}\right)^{1/\xi}, \tag{64}$$

som substitueres ind i udtrykket for momentet i ligning (61). Ved denne substitution fås jævnfør artikel [11, Bermudez et al., 2009, 1357] følgende udtryk α_s for GPD'en

$$\alpha_s = E[X^p(1-F)^s] = \frac{\sigma}{(s+1)(s+1+\xi)}. \tag{65}$$

Momentet i ligning (65) indeholder de to ukendte parametre: Formparameteren ξ, og skalaparameteren σ, som ønskes estimeret. Da konstanten s angiver antallet af parametre, opstilles i tilfældet med en to-parameter GPD to ligninger for $s = 0$ og $s = 1$:

$$\begin{aligned} \alpha_0 &= \tfrac{\sigma}{1\cdot(1+\xi)} = \tfrac{\sigma}{1+\xi} \\ \alpha_1 &= \tfrac{\sigma}{2\cdot(2+\xi)}. \end{aligned} \tag{66}$$

Løses de to ovenstående ligninger i forhold til de to ukendte parametre, kan følgende lukkede udtryk

for ξ og σ opnås

$$\xi = \frac{\alpha_0}{\alpha_0 - 2\alpha_1} - 2 \qquad \text{og} \qquad \sigma = \frac{2\alpha_0\alpha_1}{\alpha_0 - 2\alpha_1}. \tag{67}$$

Estimaterne i ligning (67) afhænger af α_0 og α_1, som begge er udtrykt ved både ξ og σ. Det betyder, at parameterestimaterne implicit er udtrykt ved sig selv, hvilket ønskes afhjulpet. I stedet for at tage udgangspunkt i den forventede værdi af $X^p(1-F)^s$, opstilles i stedet et tilsvarende udtryk for middelværdien af data. Momentet α_s opskrives som et sample estimat på formen

$$a_s = \frac{1}{n}\sum_{i=1}^{n} x_{i:n}(1 - p_{i:n})^s, \tag{68}$$

hvor $p_{i:n}$, ligesom fordelingsfunktionen F, angiver sandsynligheden for, at data ligger under en given fraktil. Dermed angiver udtrykket $1 - p_{i:n}$ sandsynligheden for at fraktilen overskrides, nemlig overskridelsessandsynligheden, som forklarer halen i den pågældende fordeling. Der er i artiklen [25, Whalen et al., 2003, 222] foretaget studier, hvor det vises at det er fordelagtigt at benytte $p_{i:n} = i - 0,35/n$.

Ud fra ovenstående, og parameterestimaterne i ligning (67), kan der nu opskrives to udtryk for parametrene ξ og σ, hvor alle elementer er kendte:

$$\xi_{PWM} = \frac{a_0}{a_0 - 2a_1} - 2 \qquad \text{og} \qquad \sigma_{PWM} = \frac{2a_0a_1}{a_0 - 2a_1}. \tag{69}$$

4.4 L-Moments-Method

L-Moments-Method (LMOM) metoden er beskrevet i bogen [16, Hosking et al., 1997, 18-22] og hvis ikke andet er angivet, har det følgende afsnit reference til denne. LMOM metoden er tæt relateret til PWM metoden, da momenterne i LMOM er linearkombinationer af momenterne i PWM'en. Den mest simple tilgang til at beskrive LMOM er derfor at tage udgangspunkt i definitionen af PWM

momenterne, som er gennemgået i forrige afsnit.

I PWM metoden kan det r'te moment β_r ud fra ligning (62) udtrykkes ved

$$\beta_r = \int_0^1 x(F)F^r dF. \tag{70}$$

Det er netop disse PWM momenter, som danner grundlaget for LMOM, idet L-momenterne som beskrevet kan opskrives som linearkombinationer af β_r. Momenterne i metoderne har følgende relation

$$\lambda_{r+1} = \sum_{k=0}^{r} p^*_{r,k}\beta_k \qquad \text{hvor} \qquad p^*_{r,k} = (-1)^{r-k}\begin{pmatrix} r \\ k \end{pmatrix}\begin{pmatrix} r+k \\ k \end{pmatrix} = \frac{(-1)^{r-k}(r+k)}{(k!)^2(r-k)}, \tag{71}$$

hvor $p^*_{r,k}$ angiver koefficienterne i et 'Shifted Legendre' polynomium [3]. De fire første L-momenter for GPD'en kan ved hjælp af β_r momenterne i ligning (70) og (71) opskrives som

$$\begin{aligned} \lambda_1 &= \beta_0 \\ \lambda_2 &= 2\beta_1 - \beta_0 \\ \lambda_3 &= 6\beta_2 - 6\beta_1 + \beta_0 \\ \lambda_4 &= 20\beta_3 - 30\beta_2 + 12\beta_1 + \beta_0 \end{aligned} \quad . \tag{72}$$

Det første L-moment, λ_1, angiver L-lokationen, eller middelværdien af fordelingen, og andet moment λ_2 er L-skala, som er et udtrykt for variansen. Udtrykkene for λ_3 og λ_4 benyttes til at bestemme forholdet mellem L-momenterne, som kan defineres ud fra formlen $\tau_r = \lambda_r/\lambda_2$, hvor τ_3 og τ_4 angiver henholdsvis L-skævheden og L-kurtosis.

Der findes to tilgange til LMOM metoden, hvor den første og nu gennemgåede er LMOM metoden med udgangspunkt i en sandsynlighedsfordeling. Denne kan, når parameterestimaterne er kendte,

[3] Jævnfør artikel [17, Hosking, 1990, 107]er $P^*_r(F)$ det r'te 'Shifted Legendre' polynomium, som kan relateres til standard Legendre polynomiet $P_r(u)$ ud fra $P^*_r(u) = P_r(2u-1)$.

benyttes til at beskrive formen af en fordeling. Den anden tilgang til LMOM metoden er at tage udgangspunkt i en sample af størrelsen n, hvorudfra de ukendte parametre kan estimeres. Sample momenterne for PWM metoden kan for enhver given fordeling, hvor observationerne er arrangeret som $x_{1:n} \leq x_{2:n} \leq ... \leq x_{n:n}$, findes ud fra følgende formel

$$b_r = n^{-1} \begin{pmatrix} n-1 \\ r \end{pmatrix}^{-1} \sum_{j=r+1}^{n} \begin{pmatrix} j-1 \\ r \end{pmatrix} x_{j:n}, \qquad r = 0, 1, 2..., \tag{73}$$

hvor b_r er en unbiased estimator af β_r. De fire første PWM momenter kan altså skrives som

$$\begin{array}{rcl} b_0 & = & n^{-1} \sum_{j=1}^{n} x_{j:n} \\ b_1 & = & n^{-1} \sum_{j=2}^{n} \frac{(j-1)}{(n-1)} x_{j:n} \\ b_2 & = & n^{-1} \sum_{j=3}^{n} \frac{(j-1)(j-2)}{(n-1)(n-2)} x_{j:n} \\ b_3 & = & n^{-1} \sum_{j=4}^{n} \frac{(j-1)(j-2)(j-3)}{(n-1)(n-2)(n-3)} x_{j:n} \end{array} . \tag{74}$$

Sample L-momenterne kan opskrives på samme måde som L-momenterne for sandsynlighedsfordelingen i ligning (71), hvor momentet β_r erstattes af b_r

$$l_1 = b_0 \qquad l_2 = 2b_1 - b_0 \qquad l_3 = 6b_2 - 6b_1 + b_0 \qquad l_4 = 20b_3 - 30b_2 + 12b_1 + b_0. \tag{75}$$

Fordelen ved sample LMOM metoden er, at sampleestimaterne l_r er unbiased estimater af λ_r. Parameterestimaterne for GPD'ens form-og skalaparameter ξ og σ kan jævnfør artikel [22, Pandey et al., 2001,183] findes ud fra samplemomenterne ved hjælp af følgende formler

$$\xi = \frac{(3\tau_3 - 1)}{(\tau_3 + 1)} \quad \text{og} \quad \sigma = (1 - \xi)(2 - \xi) l_2, \tag{76}$$

hvor $\tau_3 = l_3 / l_2$.

4.5 Maksimum-Likelihood-Estimation

Maksimum-Likelihood-Estimation (MLE) metoden er en af de mest benyttede estimationsmetoder, da den er nem at benytte, hvis man kender fordelingen af data. Grundprincippet i denne estimationsmetode er at opstille en likelihoodfunktion, som er en fælles tæthedsfunktion for hele datasættet, og at maksimere den med hensyn til de parametre der ønskes estimeret. Det vil med andre ord sige, at i MLE metoden maksimeres sandsynligheden for, at det observerede data fitter den estimerede værdi af den ukendte parameter bedst muligt.

Det kan, med reference til [2, PennState Eberly College of Science], antages at der generelt for et i.i.d. datasæt kan opskrives en fælles tæthedsfunktion ved at multiplicere tæthedsfunktionerne hørende til n antal observationer på følgende måde

$$f(X_1, X_2, ..., X_n) = f(X_1|\theta) \cdot f(X_2|\theta) \cdot ... \cdot f(X_n|\theta). \tag{77}$$

De enkelte tæthedsfunktioner i ligning (77) afhænger alle af en vektor af parametre θ, som antager samme værdi, og det er parametrene i denne, som ønskes estimeret. Studeres tæthedsfunktionen i ligning (77), hvor observationerne $X_1, X_2, ..., X_n$, er faste værdier og θ er variabel, kan likelihoodfunktionen opskrives som

$$\mathcal{L}(\theta; X_1, ..., X_n) = f(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} f(X_i|\theta). \tag{78}$$

For at gøre udregningerne mere overskuelige, tages logaritmen til likelihoodfunktionen. Dette har ingen betydning for outputtet af metoden, da både likelihoodfunktionen og log-likelihoodfunktionen vil have deres maksimum for samme værdi af de ukendte parametre i θ. Ved at tage logaritmen til likelihoodfunktionen sikres det, at funktionen altid er positiv, og det er muligt at addere leddene sammen i stedet for at multiplicere dem. Log-likelihoodfunktionen kan skrives på formen

$$\log L(\theta; X_1, X_2, ..., X_n) = \sum_{i=1}^{n} \log f(X_i|\theta). \tag{79}$$

For GPD'en med to parametre, ξ og σ, er den tilhørende fordelingsfunktion velkendt, og tæthedsfunktionen kan som i ligning (37) findes ved hjælp af differentiation, som for $\xi \neq 0$ kan opskrives som

$$g_{\xi,\sigma}(x) = \frac{1}{\sigma}\left(1 - \frac{\xi \cdot x}{\sigma}\right)^{\frac{1}{\xi - 1}}. \tag{80}$$

Indsættes denne tæthedsfunktion i ligning (79), findes log-likelihoodfunktionen for alle n i GPD'en som værende

$$L(\xi, \sigma) = -n \cdot \log(\sigma) + \left(\frac{1}{\xi} - 1\right) \frac{1}{n} \sum_{i=1}^{n} \log\left(1 - \frac{\xi \cdot X_i}{\sigma}\right) \ , \text{ for} \, \xi \neq 0. \tag{81}$$

Det ønskes i denne afhandling kun at estimere parametre hørende til de ekstreme værdier, og dermed kun en sekvens af datasættet. Da den generelle MLE metode ikke kan estimere parametre ud fra en sekvens af observationer, er det nødvendigt at benytte en alternativ tilgang.

Ifølge [12, Embrechts et al. 2012, 357] er det på grund af relationen $\overline{F}_u(y) \approx \overline{G}_{\xi,\sigma(u)}(u)$ mere realistisk, at antage en GPD for overskridelsesdata $Y_1, ..., Y_N$, hvor $N = N_u$ er uafhængig af Y_i. Det betyder, at antallet af overskridelser er uafhængig af værdien af den enkelte overskridelse Y_i, sammenlignet med det fulde datasæt $X_1, ..., X_n$ hvor antallet af overskridelser afhænger af størrelsen af den stokastiske variabel X_i. Observationen X_i er altså først en overskridelsesobservation, hvis dens værdi er større end et threshold u.

I tilfældet med overskridelsesdata benyttes der jævnfør [12, Embrechts et al. 2012, 357] en reparametrisering af parametrene $(\xi, \sigma) \to (\xi, \tau)$, hvor $\tau = -\xi/\sigma$, som leder frem til et estimat for ξ, som nu implicit afhænger af σ:

$$\hat{\xi} = \hat{\xi}(\tau) = N^{-1} \sum_{i=1}^{N} \log(1 - \tau Y_i), \tag{82}$$

hvor τ opfylder at

$$h(\tau) = \frac{1}{\tau} + \frac{1}{N} \left(\frac{1}{\hat{\xi}(\tau)} + 1 \right) \sum_{i=1}^{N} \frac{Y_i}{1 - \tau Y_i} = 0. \tag{83}$$

De fem ovenstående gennemgåede estimationsmetoder for modellering af ekstreme observationer: MOM, EPM, PWM, LMOM og MLE har nogle egenskaber og restriktioner, som skal være opfyldt. MOM er en let anvendelig metode idet parameterestimaterne, for ξ og σ, kun afhænger af samplens middelværdi og varians. Ifølge artikel [4, Castillo et al. 1997, 1610] skal det dog for MOM og PWM metoden gælde at $\xi > -0.5$, for ellers opnår metoderne ikke mulige parameterestimater, da variansen ikke vil være defineret. Findes parameterestimaterne ud fra MOM eller PWM metoden, kan der dog stadig opstå udfordringer. Parameterestimaterne er nemlig ikke nødvendigvis konsistente med de observerede sample værdier. Udfordringen kan opstå, hvis forholdet mellem skala-og formparameteren er mindre end den største ordnede observation i en sample af størrelsen n. Det vil sige hvis $\sigma/\xi < x_{n:n}$. EPM metoden har modsat MOM og PWM ikke nogle parameter restriktioner, hvilket betyder, at estimaterne eksisterer for alle værdier af ξ og σ, men metoden kan beregningsmæssigt blive tung. I LMOM metoden er L-momenterne linearkombinationer af momenterne fra PWM metoden, hvormed parameterestimaterne forventes at ligge tæt på hinanden. I MLE metoden maksimeres likelihoodfunktionen, hvormed antallet af observationer kan have en indflydelse på variansen af parameterestimaterne, jo flere observationer jo lavere varians.

4.6 Estimation af risikomål

Inden for den finansielle verden er det vigtigt for institutioner at kende til risikoen forbundet med for eksempel investeringer og mere specifikt at kunne estimere risikoen på markedet. Et par velkendte eksempler på dette er Value-at-Risk (VaR) og Expected Shortfall (ES). VaR er et risikomål, som beskriver det maksimale tab af markedsværdien over en given periode med en given sandsynlighed. Da det er vigtigt at minimere risiko, og derved prøve at reducere det maksimale tab, ønskes det at estimere ekstreme hale-fraktiler. VaR målet tager ikke højde for størrelsen af tabet, hvorfor ES målet benyttes, hvor størrelsen af det forventede tab, betinget af at tabet overstiger VaR målet, estimeres.

I dette afsnit, med reference til artikel [14, Gilli et al., 2006], vises den teoretiske estimation af de to risikomål for GPD'en.

VaR målet VaR_p kan defineres som den p'te fraktil af fordelingen F:

$$\text{VaR}_p = F^{-1}(p) = \inf\{x \in \mathbb{R} : \ F(x) \geq p\}, \qquad 0 < p < 1, \tag{84}$$

hvor $F^{-1}(p)$ er fraktilfunktionen, som angiver sandsynlighedsmassen, der ligger over en fraktil p. Det er derfor relevant at studere overskridelsesfordelingen F_u, som med udgangspunkt i ligning (35) for $(x-u)$ kan udtrykkes ved

$$F_u(x-u) = \frac{F(x) - F(u)}{1 - F(u)}. \tag{85}$$

Da overskridelsesfordelingen for $u \to \infty$ går mod en GPD, benyttes udtrykket i ligning (85), hvorudfra $F(x)$ kan isoleres og opskrives som

$$F(x) = (1 - F(u))F_u(x-u) + F(u). \tag{86}$$

Erstattes $F_u(x-u)$ med forskriften for GPD'en, hvor x sættes lig med $(x-u)$, og $F(u)$ med $(n-N_u)/n$ opnås følgende udtryk

$$\hat{F}(x) = \frac{N_u}{n}\left(1 - \left(1 + \frac{\hat{\xi}}{\hat{\sigma}}(x-u)\right)^{-1/\hat{\xi}}\right) + \left(1 - \frac{N_u}{n}\right) = 1 - \frac{N_u}{n}\left(1 + \frac{\hat{\xi}}{\hat{\sigma}}(x-u)\right)^{-1/\hat{\xi}}, \quad (87)$$

hvor N_u er antallet af overskridelsesobservationer over et givet threshold u, og n er det totale antal observationer i datasættet. Udtrykket i ligning (87) inverteres for en given sandsynlighed p, og VaR målet kan med reference til [14, Gilli et al., 2006, 214] skrives som

$$\mathrm{VaR}_p = u + \frac{\hat{\sigma}}{\hat{\xi}}\left(\left(\frac{n}{N_u}p\right)^{-\hat{\xi}} - 1\right). \quad (88)$$

Et andet risikomål er Expected Shortfall (ES), som tager udgangspunkt i VaR målet, og er defineret som den forventede værdi af et tab, som er større end VaR_p. ES målet kan skrives som

$$\mathrm{ES}_p = \mathrm{VaR}_p + E(X - \mathrm{VaR}_p | X > \mathrm{VaR}_p). \quad (89)$$

Det første led på højre side af lighedstegnet er det fundne VaR_p mål, mens det andet led er den forventede værdi af overskridelserne over VaR målet. Det huskes at ME funktionen for GPD'en kan skrives som i ligning (41):

$$e(z) = E(X - z | X > z) = \frac{\sigma + \xi z}{1 - \xi},\ \sigma + \xi z > 0. \quad (90)$$

Da det ikke er hele overskridelsesfordelingen som observeres, men kun fordelingen op til fraktilen p, sættes $z = \mathrm{VaR}_p - u$, og ud fra definitionen af ES i ligning (89) kan risikomålet skrives som

$$\mathrm{ES}_p = \mathrm{VaR}_p + \frac{\hat{\sigma} + \hat{\xi}\left(\mathrm{VaR}_p - u\right)}{1 - \hat{\xi}} = \frac{\mathrm{VaR}_p}{1 - \hat{\xi}} + \frac{\hat{\sigma} - \hat{\xi}u}{1 - \hat{\xi}}. \quad (91)$$

De ovenstående VaR og ES risikomål vil benyttes i analyse afsnittet, hvor der analyseres på risiko-

målene fundet i forhold de forskellige gennemgåede estimationsmetoder. Det ønskes mere præcist at studere haleadfærden for en ekstremværdifordeling for høje fraktiler.

4.6.1 Kritik af VaR

Finansielle institutioner benytter VaR målet til at måle risiko, og det er samtidig det mål, som BIS udsteder de lovmæssige krav til institutionerne ud fra. Da det er et så udbredt risikomål, må man antage, at det er validt.

Det er dog ikke alle, der er enige i dette. Hans Rau-Bredow har i [23, Rau-Bredow, 2004, k.5] kritiseret VaR målet, og kritikken bygger på antagelsen om en normalfordeling, samt konveksitet og subadditivitet af risikomålet.

Er datagrundlaget perfekt normalfordelt, kan VaR målet findes direkte ud fra standardafvigelsen. Da data ikke altid kan antages at være normalfordelt, er det sjældent, at denne relation mellem VaR målet og standardafvigelsen kan benyttes. Teoretisk er det ikke korrekt at antage, at finansielle afkast er normalfordelte, hvilket der ikke altid bliver taget højde for i praksis. I denne afhandling antages data at være GP fordelt, hvilket er implementeret i VaR og ES risikomålene. Implementeringen medfører mere komplicerede udregninger end ved antagelsen om en normalfordeling, men vil også give mere korrekte resultater.

En anden udfordring ved VaR målet, der ikke på samme måde kan tages højde for, er kravet om konveksitet, som medfører subadditivitet. Kravet er opfyldt for alle såkaldte elliptiske fordelinger, med andre ord alle fordelinger som generaliserer multivariate normalfordelinger, herunder lognormalfordelingen, som man ofte antager, at finansielle afkast følger. I tilfældet med ekstreme værdier kan det ikke antages, at data er lognormalfordelt, og det kan dermed ikke nødvendigvis antages, at kravet om subadditivitet er opfyldt. Dette vil dog ikke have den store konsekvens, da der i denne afhandling kun studeres en enkelt akties afkast, og ikke en portefølje. Hvis fokus havde været på mere end en aktie, kunne konsekvensen af manglende subadditivitet have været stor. Værdien af

VaR af en samlet portefølje kan være større end værdien af summen af de enkelte VaR værdier, hvilket strider imod ideen om at investere i for eksempel en portefølje af aktier, i stedet for en enkelt aktie. Ved at diversificere sin portefølje kan man, i og med at aktierne er korrelerede, opnå en mindre risiko.

Ifølge [23, Rau-Bredow, 2004, k.5] kan man ved hjælp af den første og anden afledte af både VaR og ES målet, vise om de opfylder subadditivitet eller ej, hvilket vi ikke har valgt at gå i dybden med i denne afhandling. Resultaterne viser, som skrevet ovenfor, at VaR målet ikke altid opfylder subadditivitet for ikke elliptiske fordelinger, men det gør ES målet derimod, hvorfor det også er relevant at studere dette risikomål.

4.7 Test af estimater

Der er nu gennemgået fem forskellige metoder til at estimere de ukendte parametre, ξ og σ samt høje fraktiler i en GPD, hvilket resulterer i forskellige parameterestimater. Det næste er at undersøge, hvordan der kan skelnes mellem estimaterne, og finde en metode, som kan indikere hvilken estimationsmetode, der er bedst at benytte. Det optimale er at finde et estimat, som er så tæt som muligt på den 'sande' værdi θ^*. Inden for statistik findes der to former for data: Populations data, hvorudfra man kan angive de parametre og informationer, som er korrekte for den pågældende population. Det vil sige de 'sande' værdier. Derudover er der sample data, som ikke er en komplet mængde af data, men en stikprøve, hvorfor estimaterne kan variere fra de 'sande' parametre. I denne afhandling findes den 'sande' værdi ud fra Bootstrapping, hvor den 'sande' værdi er middelværdien af estimaterne beregnet på baggrund af simulerede datasæt. Denne Bootstrap metode er fordelagtig at benytte til at bestemme præcisionen af de estimerede parametre, da der ud fra metoden kan beregnes Standard Errors og konfidensintervaller. Teorien i dette afsnit er skrevet med reference til [24, Ruppert, 2011, 133].

4.7.1 Standard Error og konfidensinterval

Standard Error (SE) benyttes til at undersøge præcisionen af et parameterestimat, og angiver standardafvigelsen af mindste kvadraters estimat. Det er ikke nok kun at analysere ud fra standardafvigelsen i sig selv, da den beskriver variationen i en population og ikke en udvalgt sample. Det kan hermed siges, at standardafvigelsen er den 'sande' SE, som for eksempel kan benyttes til bestemmelse af konfidensintervaller.

SE er et mål for samplingsfejl, og kan beregnes ud fra følgende formel

$$SE_{\hat{\theta}} = \sqrt{\frac{\sum_{n=1}^{N}(\theta_n^* - \hat{\theta})^2}{N}}, \tag{92}$$

hvor θ_n^* er det n'te estimat beregnet ud fra Bootstrap metoden, og middelværdien af disse n estimater angiver den 'sande' parameter. Ud fra det stokastiske parameterestimat $\hat{\theta}$, θ_n^* og antallet af observationer, kan SE hørende til de fundne estimater beregnes. SE er dermed et mål for hvor præcist det fundne parameterestimat er. Desto mindre SE værdien er, jo bedre er estimatet i forhold til den 'sande' parameter.

SE kan benyttes til at beregne konfidensintervaller for parameterestimaterne. Helt generelt findes konfidensintervaller for en fordeling ud fra den 'sande' parameter, en fraktilværdi som angiver konfidensniveauet samt SE. For normalfordelingen beregnes konfidensintervaller for de fundne parametre på et 95%-niveau ud fra formlen

$$\theta^* \pm 1,96 \cdot SE_{\hat{\theta}}. \tag{93}$$

I tilfældet hvor der modelleres med ekstreme værdier, og fordelingen ikke kan antages at være normal, benyttes der i stedet Bootstrap konfidensintervaller. Denne metode bygger på Bootstrapping, hvor den nedre og øvre grænse i 95% konfidensintervallet $[\theta_n; \theta_\emptyset]$, findes ud fra de sorterede Bootstrapværdier. Den nedre grænse er værdien liggende ved $2,5\%$, og den øvre grænse er værdien

liggende ved $97,5\%$.

Konfidensintervallet benyttes til at teste om det fundne estimat ligger inden for, i dette tilfælde, et 95% konfidensinterval af den 'sande' parameter.

4.7.2 Bias

Bias af et estimat angiver hvor stor en forskel, der er på den estimerede værdi og den 'sande' værdi. Bias for et parameterestimat findes ved at se på distancen til den 'sande' parameter:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta^*, \tag{94}$$

hvor $E(\hat{\theta})$ er den forventede værdi af den estimerede parameter, og θ^* igen angiver den 'sande' populations parameter. Hvis $B(\hat{\theta}) = 0$ er estimatet $\hat{\theta}$ unbiased, hvilket vil sige, at den estimerede parameter er lig med den 'sande' parameter. Der ønskes derfor et så lille bias som muligt.

4.7.3 Mean Sqared Error

Det kan ske, at den estimerede parameter ikke er lig med den 'sande' parameter, hvilket kan skyldes andre faktorer som for eksempel den naturlige varians for stokastiske variable. Derfor er det ikke kun nok at analysere parameterestimaterne ud fra Bias. Det er i stedet fordelagtigt at bestemme den forventede afvigelse mellem parameterestimatet og dens 'sande' værdi. Metoden til dette kaldes for Mean Squared Error (MSE), og kan opskrives som

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta^*)^2]. \tag{95}$$

I udtrykket for MSE er der en betydelig relation mellem varians og bias, i og med at man ud fra den forventede afvigelse også kan skrive

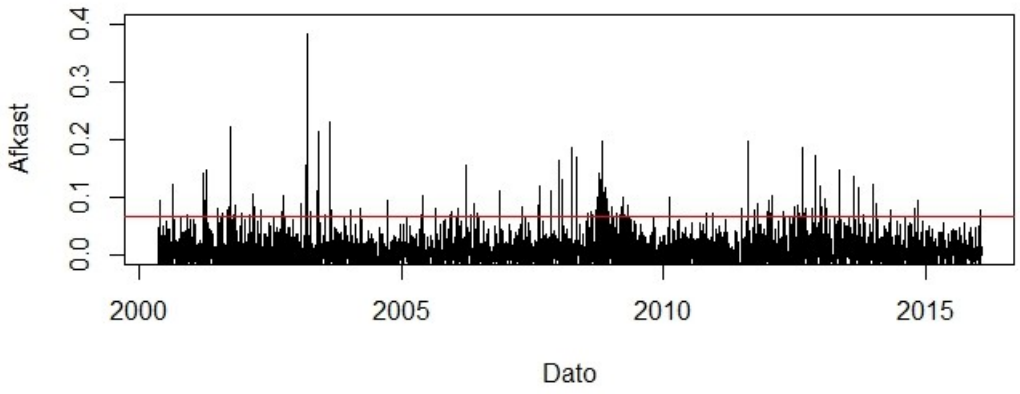
$$
\begin{aligned}
E[(\hat{\theta}-\theta^*)^2] &= E\left[(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta^*)^2\right] \\
&= E\left[(\hat{\theta}-E(\hat{\theta}))^2\right] + \Big(E(\hat{\theta})-\theta^*\Big)^2 + 2\left((E(\hat{\theta})-\theta^*))\cdot \underbrace{E(\hat{\theta}-E(\hat{\theta})}_{E(\hat{\theta})-E(\hat{\theta})=0}\right) \\
&= E[(\hat{\theta}-E(\hat{\theta}))^2] + \Big[E(\hat{\theta})-\theta^*\Big]^2 \\
&= Var(\hat{\theta}) + B(\hat{\theta})^2. \qquad (96)
\end{aligned}
$$

MSE'en kan altså udtrykkes som summen af variansen og Bias af den estimerede parameter. Hvis et parameterestimat derfor er unbiased, det vil sige $B(\hat{\theta})^2 = 0$, vil MSE metoden stadig tage højde for den varians, der naturligt fremkommer ved stokastiske variable. Det er ikke altid fordelagtigt at have en fuldstændig unbiased parameter, da det kan medføre en højere varians og dermed en potentiel højere MSE. Derfor er MSE et tradeoff mellem varians og Bias, hvor der ønskes en så lav værdi som muligt.

5 Punktprocesser

Vi har nu studeret teorien bag hvordan i.i.d. data kan fittes til to ekstremværdifordelinger samt forskellige estimationsmetoder, med henblik på at kunne benytte VaR og ES risikomålene.

I denne afhandling ønskes det at undersøge brugen af EVT, mere specifikt forekomsten af store finansielle tab på Vestas aktien. Det er velkendt, at finansielle afkast kan have tendens til klyngedannelse, hvormed der vil være en form for afhængighed i data. I figur (9) er negative afkasts absolutte værdi for Vestas aktien i perioden 10.05.2000 til 28.01.2016 illustreret. Den horisontale linje angiver thresholdværdien, som i dette tilfælde er $6,5\%$, hvor observationer der overskrider denne, anses som værende ekstreme værdier.



Figur 9: Negative afkasts absolutte værdi for Vestas aktien, med $u = 0,065$.

Der er i figur (9) en tendens til klyngedannelse, og især omkring finanskrisen i 2008, som også var forventeligt. Derudover bemærkes det, at et stort tab sjældent står alene, hvilket kan tyde på, at der i Vestas datasættet er klyngedannelse. Denne klyngedannelse kan forklares ved, at der er heteroskedasticitet i data, på den måde at der er en form for tidsvariation i den betingede varians. Det kan derfor være nyttigt at studere en metode til at modellere data som ikke er i.i.d..

I de følgende afsnit tages udgangspunkt i en Poisson punktproces, som antager i.i.d. data, og den tilpasses således at der opnås en ny model, som tager højde for eventuelle klynger i data. Ud

fra denne nye model kan der estimeres fraktiler, og derudfra kan risikomålene findes.

Først gennemgås den generelle teori omkring en punktproces, derefter Poisson punktprocessen, og til sidst to specifikke former for Hawkes punktprocesser: Hawkes POT med forudsigelige og uforudsigelige mærker. Hawkes POT modellerne medtages med henblik på at afhjælpe udfordringen med klyngedannelse bedst muligt. De to første underafsnit har, hvis ikke andet er angivet, reference til [20, McNeil et al.,2015, 165-168].

5.1 Generelt om punktprocesser

En punktproces er en stokastisk proces, hvor udfaldet er en samling af punkter, som falder inden for et givet område. Punktprocesser benyttes til at bestemme et såkaldt punkt, som ofte repræsenterer en lokation eller tidspunkt for en given hændelse. Da der i denne afhandling studeres ekstreme finansielle tab, anses punkterne som værende hændelser i tid.

En generel punktproces kan defineres ved, at studere en sekvens af stokastiske variable $Y_1, Y_2, ..., Y_n$ som antager værdier i et givet udfaldsrum $\mathcal{X}$, enten endimensionelt eller flerdimensionelt. At en punktproces siges at være endimensionel, betyder at der kun tages højde for tidspunktet, hvorpå en hændelse indtræffer. En todimensionel punktproces medtager udover tiden også størrelsen af hændelserne.

De stokastiske variable i en punktproces kan ifølge [20, McNeil et al., 2015, 165] for enhver delmængde $A \subset \mathcal{X}$ defineres som

$$N(A) = \sum_{i=1}^{n} I_{\{Y_i \in A\}}, \tag{97}$$

hvor $I_{\{Y_i \in A\}}$er en indikatorfunktion og $N(A)$ er dermed en stigende tælleproces, der tæller alle de stokastiske variable Y_i i mængden A. En generel punktproces defineres som $N(\cdot)$, og specificeres ud fra en given sandsynlighedsfordeling. Denne punktproces kan dermed beskrive en sandsynlighedsmæssig

struktur af forskellige sæt af punkter i et metrisk rum[4].

For at kunne definere hvilken punktproces der kan benyttes, antages det, fra tidligere teori, at tabsfordelingen er i MDA af en ekstremværdifordeling, og jævnfør afsnit 2.2.2 er følgende udtryk også opfyldt

$$\lim_{n\to\infty} P(M_n \leq b_n x + a_n) = \lim_{n\to\infty} F^n(b_n x + a_n) = G(x), \quad x \in \mathbb{R}. \tag{98}$$

Udtrykket i ligning (98) er som tidligere nævnt gældende for en ikke-degenereret grænsefordeling G_ξ, med normaliseringskonstanter b_n og a_n. Tages logaritmen på begge sider af lighedstegnet, opnås følgende

$$\lim_{n\to\infty} n\, ln\, F(b_n x + a_n) = \ln G_\xi(x). \tag{99}$$

Der antages nu en sekvens af thresholds, defineret ved $u_n(x) := b_n x + a_n$, hvor n er antallet af thresholds. Ud fra egenskaberne ved logaritmefunktionen vides det, at $-\ln(y) \sim 1 - y$ for $y \to 1$, og tabsfordelingen kan således udtrykkes som halen af fordelingsfunktionen $\overline{F} = 1 - F$. Ud fra ligning (99) vil der altså, for en sekvens af thresholds, gælde

$$n\overline{F}(u_n(x)) \sim -n\, \ln\, F(u_n(x)) \,\rightarrow\, -\ln\, G_\xi(x)\ ,\ n \to \infty. \tag{100}$$

Det betyder at tabsfordelingen for ethvert threshold u_n, kan udtrykkes som værende en ekstremværdifordeling.

Antallet af observationer som overskrider et givet threshold $u_n(x)$, er binomial fordelte stokastiske variable med fordelingen $N_{u_n(x)} \sim B(n, \overline{F}(u_n(x)))$, hvor n er antallet af observationer i det fulde

[4]Ifølge [9, Beltoft et al., 2009, 1] er et metrisk rum en mængde, hvor der er defineret en afstand mellem punkterne i mængden.

datasæt, og $\overline{F}(u_n(x))$ angiver sandsynligheden for at observationen overskrider et threshold u_n. Det medfører at den forventede værdi af det totale antal ekstreme observationer kan opskrives som $n\,\overline{F}(u_n(x))$. Ud fra generel sandsynlighedsfordelingsteori vides det, at binomialfordelingen, for $n \to \infty$ og $p \to 0$ kan approksimeres mod en Poissonfordeling. Sammenholdt med ligning (99) for $n \to \infty$ vil at antallet af overskridelser $N_{u_n(x)}$ konvergere mod en Poisson fordelt stokastisk variabel, med middelværdi $\lambda(x) = -\ln G_\xi(x)$, hvormed overskridelserne sker i henhold til en Poisson punktproces.

Middelværdien i en Poisson punktproces kan ved at isolere x i udtrykket for $u_n(x) = b_n x + a_n$ skrives som

$$\lambda(x) = -\ln G_\xi(x) = -\ln G_\xi((u - a_n)/b_n). \tag{101}$$

Hvis normaliseringskonstanterne b_n og a_n henholdsvis erstattes med $\sigma > 0$ og μ, opnås en Poisson proces med rate $-\ln G_{\xi,\mu,\sigma}(x)$.

5.2 Poisson punktprocessen

Det er nu vist, at ekstreme værdier også kan modelleres ved at benytte Poisson punktprocessen, hvorfor det vælges at gå videre med denne. Helt generelt har POT-modellen følgende antagelser

1. Overskridelser sker i henhold til en homogen Poisson proces[5] i tid
2. Størrelsen af overskridelserne over det fastlagte threshold er i.i.d., og uafhængig af tidspunktet af overskridelsen
3. Størrelsen af overskridelserne er GP fordelt

[5]Ifølge [15, Haals et al., 2009, 14] siges en punkt proces at være homogen, hvis intensiteten af punkterne i processen er jævnt fordelt over hele observationsområdet. En homogen punktproces kan også defineres ud fra, at intensiteten $N(\cdot)$ er konstant.

Modellen der opfylder disse antagelser, kaldes også for en mærket Poisson punktproces, hvor tidspunkterne for overskridelserne er punkter, og de GP fordelte overskridelser er mærker.

I denne afhandling ønskes det ikke kun at studere en endimensionel punktproces, hvor fordelingen beskrives ud fra hvornår de ekstreme hændelser sker i tid. Det ønskes derimod at studere en todimensionel punktproces, hvor fordelingen både beskrives ud fra hvornår de ekstreme hændelser sker og selve størrelsen af dem. Det betyder, at i den mærkede punktproces er hvert punkt beskrevet ud fra en tid og en værdi i form af et mærke, og er dermed ikke lige så simpel som den almindelige endimensionelle punktproces.

For at kunne studere en Poisson punktproces nærmere tages der udgangspunkt i den ikke-mærkede Poisson punktproces, som kan defineres ud fra to krav. Det første krav er, at der for alle underrum i et metrisk rum $A \subset \mathcal{X}$ skal gælde, at sandsynligheden for et bestemt antal af indtrufne punkter skal følge en Poissonfordeling, hvis fordelingsfunktion generelt er givet ved

$$P(N(A) = k) = \begin{cases} e^{-\Lambda(A)} \frac{\Lambda(A)^k}{k!} & , \ \Lambda(A) < \infty \\ 0 & , \ \Lambda(A) = \infty \end{cases}, \tag{102}$$

hvor Λ er Poissonfordelingens parameter, som angiver intensiteten af punkterne. Det andet krav som skal være opfyldt, er at hvis $A_1, ..., A_m$ indbyrdes er disjunkte delmængder af $\mathcal{X}$, da skal de stokastiske variable $N(A_1), ..., N(A_m)$ være uafhængige for alle $m \geq 1$.

Intensiteten $\Lambda(A)$ svarer til middelværdien $E(N(A)) = \Lambda(A)$, og denne intensitet kan betegnes som den afledede af intensitetsfunktionen $\lambda(x)$:

$$\Lambda(A) = \int_A \lambda(x)\, dx. \tag{103}$$

Intensitetsfunktionen $\lambda(x)$ kan altså implicit udtrykkes ud fra fordelingsfunktionen i ligning (102).

For nu at danne en mærket Poisson punktproces tages der udgangspunkt i en tre-parameter GPD. Da fordelingens overskridelser sker i henhold til en homogen Poisson proces i tid, kan man med antagelsen om regulært fordelte stokastiske variable opskrive intensitetsfunktionen for en todimensionel Poisson punktproces som

$$\lambda(t,x) = \frac{1}{\sigma}\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi-1}, \tag{104}$$

hvor t og x er stokastiske, og $(1+\xi(x-\mu)/\sigma) > 0$, for ellers vil $\lambda(t,x) = 0$. Udtrykket for intensitetsfunktionen i ligning (104) afhænger ikke af tiden t men af x, hvorfor intensitetsfunktionen for den todimensionelle Poisson proces kan skrives som $\lambda(x) := \lambda(t,x)$.

Ud fra relationen i ligning (103) kan intensiteten $\Lambda(A)$ for en generel todimensionel punktproces i en delmængde $A = (t_1,t_2)\times(x,\infty) \subset \mathcal{X}$ skrives som

$$\Lambda(A) = \int_{t_1}^{t_2}\int_{x}^{\infty}\lambda(y)\,dy\,dt = -(t_2-t_1)\ln G_{\xi,\mu,\sigma}(x). \tag{105}$$

Da der i den todimensionelle Poisson punktproces udelukkende fokuseres på størrelsen af de observationer, som er større end et threshold u, tages der dermed ikke hensyn til tidspunktet, hvorpå de indtræffer. Herudfra, samt ud fra teorien om punktprocesser generelt, kan den implicitte endimensionelle punktproces for overskridelser, siges at være en homogen Poisson punktproces med rate $\tau(x) := -\ln G_{\xi,\mu,\sigma}(x)$.

Da det tidligere blev vist, at tabsfordelingen $\overline{F}$ er i MDA af en ekstremværdifordeling, kan der med udgangspunkt i fordelingsfunktionen for GEV fordelingen i ligning (24), findes frem til halen af overskridelsesfordelingen $\overline{F}_u(x)$. På baggrund af resultatet i [20, McNeil et al., 2015, 150] kan $\overline{F}_u(x)$ udtrykkes som forholdet mellem raten af overskridelserne over henholdsvis $(u+x)$ og u:

$$\overline{F}_u(x) = \frac{\tau(u+x)}{\tau(u)} = \left(1+\frac{\xi x}{\sigma+\xi(u-\mu)}\right)^{-1/\xi} = \overline{G}_{\xi,\beta}(x), \tag{106}$$

hvor skalaparameteren $\beta = \sigma + \xi(u - \mu) > 0$. Udtrykket i ligning (106), er netop halen af GPD'en for overskridelser over et threshold u, som den kendes fra ligning (36). Der er dermed en teoretisk sammenhæng mellem GEV fordelingen og Poisson modellen, hvor observationer der overskrider et threshold u anses som værende hændelser i tid. Da det gælder, at GPD'en er i MDA af GEV fordelingen, er der en implicit sammenhæng mellem GPD'en og denne Poisson model.

I punktprocesser, hvor intensiteten er modelleret ud fra GPD'en, er der ligesom i de tidligere gennemgåede fordelinger nogle ukendte parametre, som ønskes estimeret. Den mest benyttede estimationsmetode inden for punktprocesser er MLE metoden. Tilgangen er dog her lidt anderledes, og den nedenstående teori omkring parameterestimation har derfor reference til [6, Christophersen, 2011, 15-16].

Likelihoodfunktionen $L(\theta|X)$ kan ligesom tidligere generelt opskrives ud fra tæthedsfunktionen $f_\theta(x)$, men da X er en punktproces på tidslinjen, udtrykkes tætheden for et punkt t_i ud fra den betingede tæthedsfunktion $f(t_i|\mathcal{H}_{t_i})$:

$$L(\theta|x) = L(\theta) = f_\theta(x) = f(t_1|\mathcal{H}_{t_i})...f(t_n|\mathcal{H}_{t_n})(1 - F(T|\mathcal{H}_{t_T})). \tag{107}$$

$\mathcal{H}_t$ er et filter indeholdende alt information om tiden op til og med tidspunkt t, og $(1 - F(T|\mathcal{H}_{t_T}))$ angiver sandsynligheden for ikke at have nogle punkter efter tidspunkt t_n, hvor t_n er det sidste punkt før tidspunkt t. Ud fra den betingede tæthedsfunktion $f(t|\mathcal{H}_t)$ og den tilhørende fordelingsfunktion $F(t|\mathcal{H}_t)$ kan den betingede intensitetsfunktion skrives som

$$\lambda^*(t) = \frac{f(t|\mathcal{H}_t)}{1 - F(t|\mathcal{H}_t)}, \tag{108}$$

hvor $f(t|\mathcal{H}_t)$ og $F(t|\mathcal{H}_t)$ er givet ved

$$f(t|\mathcal{H}_t) = \lambda^*(t)e^{\left(-\int_{t_n}^{t} \lambda^*(s)ds\right)} \qquad \text{og} \qquad F(t|\mathcal{H}_t) = 1 - e^{\left(-\int_{t_n}^{t} \lambda^*(s)ds\right)}. \tag{109}$$

Ud fra udtrykket i ligning (108) og (109) kan likelihoodfunktionen i ligning (107) omskrives til

$$\begin{aligned} L(\theta) &= \left(\prod_{i=1}^{n} f(t_i|\mathcal{H}_{t_i})\right) \frac{F(T|\mathcal{H}_T)}{\lambda^*(T)} \\ &= \left(\prod_{i=1}^{n} \lambda^*(t_i) e^{\left(-\int_{t_{i-1}}^{t_i} \lambda^*(s)ds\right)}\right) \frac{\lambda^*(T) e^{\left(-\int_{t_n}^{t} \lambda^*(s)ds\right)}}{\lambda^*(T)} \\ &= \left(\prod_{i=1}^{n} \lambda^*(t_i)\right) e^{\left(\int_0^T \lambda^*(s)ds\right)}, \end{aligned} \tag{110}$$

hvor $t_0 = 0$, $\lambda^*(t)$ er den betingede intensitetsfunktion, og $\{t_1, t_2, ..., t_n\}$ angiver punkterne i tidsintervallet $[0, T]$. Likelihoodfunktionen for en punktproces udtrykkes altså ved hjælp af processens intensitetsfunktion, i modsætning til den generelle likelihoodfunktion som tager udgangspunkt i en tæthedsfunktion.

I tilfældet med en endimensionel punktproces kan raten i udtrykket for likelihoodfunktionen i ligning (110) erstattes med notationen $-\tau(u)$, da det tidligere er vist at $\tau(x) = -\ln G_{\xi,\mu,\sigma}(x)$. Det betyder, at likelihoodfunktionen jævnfør [20, McNeil et al. 2015, 168] kan udtrykkes ud fra $-\tau(u)$ samt produktet af intensitetsfunktioner for alle j overskridelsesobservationer $X_1, ..., X_{N_u}$:

$$L(\theta; \tilde{X}_1, ..., \tilde{X}_{N_u}) = e^{-\tau(u)} \prod_{j=1}^{N_u} \lambda(\tilde{X}_j). \tag{111}$$

Maksimeres denne likelihoodfunktion med hensyn til de ukendte parametre ξ, σ og μ som vektoren θ består af, kan parameterestimaterne opnås på samme måde som ved en standard MLE metode.

6 Self-Exciting punktprocesser

I teorien er der indtil nu antaget i.i.d. data, men da der kan være tendens til klyngedannelse, vil Self-Exciting modellen i dette afsnit gennemgås. En punktproces siges at være Self-Exciting, hvis intensiteten afhænger af overskridelserne op til tidspunkt t, hvilket vil sige, at punktprocessen er tidsafhængig.

Der findes to typer af Self-Exciting modeller: Hawkes og ETAS, og vi har i denne afhandling valgt

kun at studere Hawkes punktprocessen. Først gennemgås den generelle Hawkes punktproces hvorefter egenskaberne for Poisson punktprocessen tilføjes, og Hawkes POT processen opnås. Modellerne i dette afsnit har reference til [20, McNeil et al., 2015, 578-581].

6.1 Hawkes

I den generelle Hawkes proces antages et datasæt bestående af $X_1, ...X_n$ observationer, et threshold u og N_u antal overskridelser. Overskridelserne betegnes som $(T_j, \widetilde{X}_j)$ for $j = 1, ..., N_u$, hvor T_j angiver tidspunkterne og $\tilde{X}_j$ er mærkerne.

En punktproces $N(\cdot)$ for overskridelser antages at være en Self-Exciting proces, hvor der i intensiteten betinges med de tidligere overskridelser, og den betingede intensitet kan skrives på formen

$$\begin{aligned} \lambda^*(t) &= \tau + \psi \textstyle\sum_{j:0<T_j<t} h(t - T_j,\, \tilde{X}_j - u) \\ &= \tau + \psi v^*(t). \end{aligned} \tag{112}$$

Det skal her gælde at parametrene $\tau > 0$ og $\psi \geqslant 0$, og at funktionen h kun kan antage positive værdier. I udtrykket for den betingede intensitet i ligning (112) er der udover de to parametre, også udtrykket $(t - T_j)$, som angiver tiden siden den forgående overskridelse, og $(\tilde{X}_j - u)$ som beskriver størrelsen af den j'te overskridelsesobservation. Det medfører, at de foregående overskridelser $(T_j, \widetilde{X}_j)$ har betydning for processens betingede intensitet, da de både påvirker tidspunktet for observationen og størrelsen. Den betingede intensitet beskriver risikoen for en ny overskridelse af thresholdværdien på tidspunkt t, ligesom raten i en standard Poisson proces. Den betingede intensitet er i sig selv en stokastisk proces, som afhænger af informationen op til, men ikke inklusiv tidspunkt t.

Valget af h-funktionen specificerer hvilken proces det ønskes at modellere, og de to mest benyttede versioner af h-funktionen er Hawkes og ETAS processen:

- $h(s, x) = e^{\delta x - \gamma s},\ \delta, \gamma > 0$, er den 'simple' Hawkes model.

- $h(s,x) = e^{\delta x}(s+\gamma)^{-(\rho+1)},\ \delta,\gamma,\rho > 0$, er 'Epidemic Type After-Shock' (ETAS) modellen.

ETAS modellen benyttes ofte, som navnet antyder, til modellering af forekomster af jordskælv, hvor modellen tager højde for såkaldte efterskælv. Det ses ud fra forskrifterne for h-funktionen, at Hawkes modellen er en del mere simpel end ETAS modellen. Vi har som tidligere nævnt kun valgt at fokusere på Hawkes punktprocessen, hvor parameterestimaterne kan finde ved at maksimere likelihoodfunktionen, som fremkommer på samme måde som likelihoodfunktionen i ligning (111):

$$L(\theta : T_1, ...T_{N_u}) = e^{\left(-\int_0^n \lambda^*(s)ds\right)} \prod_{i=1}^{N_u} \lambda^*(T_i), \tag{113}$$

hvor θ er en vektor af de ukendte parametre i intensitetsfunktionen: τ, ψ, γ og δ som ønskes estimeret.

6.2 Hawkes POT

Hawkes POT modellen fremkommer med udgangspunkt i Poisson punktprocessen, hvor egenskaberne for en Hawkes proces tilføjes.

I dette afsnit opstilles der en model, med to forskellige forudsætninger for mærkerne. Den første model er en mærket Self-Exciting punktproces, hvor mærkerne er GP fordelte og uforudsigelige. At mærkerne er uforudsigelige betyder, at de ikke afhænger af tidligere hændelser, og kan siges at være i.i.d.. Den anden model er en Self-Exciting punktproces, med samme antagelse om GP fordelte mærker, men hvor mærkerne er forudsigelige. Mærkerne er forudsigelige i den forstand, at skalaparameteren i GPD'en afhænger af historisk data, op til tidspunktet hvor mærket indtræffer.

6.3 Hawkes POT med uforudsigelige mærker

For at opstille en Hawkes proces med uforudsigelige mærker tages der udgangspunkt i 'Self-Exciting' elementet i ligning (112), som er givet ved

$$v^*(t) = \sum_{j:0<T_j<t} h(t - T_j, \tilde{X}_j - u). \tag{114}$$

Denne funktion kaldes for en 'Self-Exciting' funktion, hvor hver tidligere overskridelse $(T_j, \tilde{X}_j)$ påvirker både tidspunktet, hvorpå observationen indtræffer, og størrelsen af denne observation. Det kan siges, at denne funktion er afhængighedselementet i Self-Exciting modellerne, som tager højde for eventuel afhængighed i data. Det ønskes at finde en intensitet, hvor der både tages højde for GPD'en, og 'Self-Exciting' funktionen i ligning (114).

For at kunne finde frem til intensiteten for en Hawkes POT model med uforudsigelige mærker, er det nødvendigt at lave en reparametrisering af intensiteten for Poisson punktprocessen. I følge [20, McNeil et al., 2015, 168] kan Poisson punktprocessens rate τ omskrives til $\tau := \tau(u) = -\ln G_{\xi,\mu,\sigma}(u)$. Hvis der antages en positiv skalaparameter på formen $\beta = \sigma + \xi(u - \mu)$, kan intensiteten for en Hawkes POT med uforudsigelige mærker opskrives som

$$\lambda(x) = \lambda(t, x) = \frac{\tau}{\beta}\left(1 + \xi\frac{x - u}{\beta}\right)^{-1/\xi - 1}, \tag{115}$$

hvor $\xi \in \mathbb{R}$, og τ og $\sigma > 0$. Kort sagt ønskes en kombination af den endimensionelle intensitet fra Hawkes modellen i ligning (112), og intensiteten fra reparametriseringen af Poisson punktprocessen i ligning (115). Hermed opnås intensiteten

$$\lambda^*(t, x) = \frac{\tau + \psi v^*(t)}{\beta}\left(1 + \xi\frac{x - u}{\beta}\right)^{-1/\xi - 1}, \tag{116}$$

hvor $\lambda^*(t, x)$ ligger i udfaldsrummet $\mathcal{X} = (0, n] \times (u, \infty)$. For parametrene skal det gælde at $\tau > 0$ og $\psi \geq 0$. Det bemærkes, at hvis parameteren $\psi = 0$ vil $\psi v^*(t) = 0$, og udtrykket for intensiteten i ligning (116) ender ud med at være det samme udtryk som i standard Poisson punktprocessen, hvor der ikke er nogen afhængighed i data. Det betyder, at 'Self-Exciting' elementet udelukkende ligger i udtrykket for $\psi v^*(t)$.

Raten τ for at en given observation vil overskride et threshold u til tidspunkt t, givet den historiske information der er tilgængelig, kan findes ved at integrere intensitetsfunktionen. Det giver følgende udtryk

$$\tau^*(t,x) = \int_x^\infty \lambda^*(t,y)dy = (\tau + \psi v^*(t)) \left(1 + \xi \frac{x-u}{\beta}\right)^{-1/\xi}. \tag{117}$$

Ud fra raten $\tau^*(t,x)$ kan fordelingen af tab som overskrider et threshold u, på samme måde som i ligning (103), ske i henhold til en GPD :

$$\frac{\tau^*(t,u+x)}{\tau^*(t,u)} = \left(1 + \frac{\xi x}{\beta}\right)^{-1/\xi} = \overline{G}_{\xi,\beta}(x). \tag{118}$$

Forskellen på ligning (118) og på udtrykket i ligning (106) er at tidsaspektet her er med, hvilket også kendetegner Hawkes processen. Tidsaspektet har dog ingen indflydelse på fordelingen af overskridelserne, hvorfor der stadig modelleres med GP fordelte overskridelser. Hawkes POT med uforudsigelige mærker kan dermed være fordelagtig at benytte når der modelleres med ekstreme værdier.

6.4 Hawkes POT med forudsigelige mærker

Intensiteten i ligning (116) kan generaliseres ved at erstatte den ikke-tidsafhængige skalaparameter β med en tidsafhængig 'Self-Exciting' funktion $\beta + \alpha v^*(t)$, og følgende intensitet opnås

$$\lambda^*(t,x) = \frac{\tau + \psi v^*(t)}{\beta + \alpha v^*(t)} \left(1 + \xi \frac{x-u}{\beta + \alpha v^*(t)}\right)^{-1/\xi - 1}, \tag{119}$$

hvor $\beta > 0$ og $\alpha \geq 0$. Denne model er generaliseret i den forstand, at værdien af α angiver, om modellen har forudsigelige eller uforudsigelige mærker. Hvis $\alpha = 0$, er udtrykket $\alpha v^*(t) = 0$, og mærkerne har derfor ingen indflydelse på modellen. I tilfældet hvor $\alpha = 0$ opnås udtrykket for en model med uforudsigelige mærker som i ligning (116). Det skal nævnes, at forskellen mellem modellerne med forudsigelige og uforudsigelige mærker ligger i selve fordelingen af mærkerne. I

begge modeller antages GP fordelte mærker, hvor mærkerne i den uforudsigelige model er i.i.d., og i den forudsigelige model antages at følge en betinget GPD, som afhænger af tiden.

Overskridelsesraten for en Hawkes POT med forudsigelige mærker kan findes ved at integrere intensitetsfunktionen, og er givet ved

$$\tau^*(t,x) = \int_x^\infty \lambda^*(t,y)dy = (\tau + \psi v^*(t)) \left(1 + \xi \frac{x-u}{\sigma + \alpha v^*(t)}\right)^{-1/\xi}. \tag{120}$$

Overskridelserne sker her, på samme måde som i ligning (118), i henhold til en GPD idet

$$\frac{\tau^*(t,u+x)}{\tau^*(t,u)} = \left(1 + \frac{\xi x}{\beta + \alpha v^*(t)}\right)^{-1/\xi} = \overline{G}_{\xi,\beta+\alpha v^*(t)}(x), \tag{121}$$

hvor den eneste forskel ligger i skalaparameteren β, som nu indeholder det tidsafhængige 'Self-Exciting' element $\beta + \alpha v^*(t)$. I modellen med forudsigelige mærker vil observationerne, givet at de indtræffer på tidspunkt t og givet informationen op til tidspunkt t, følge en GPD med formparameter ξ og skalaparameter $\beta + \alpha v^*(t)$.

Til at estimere processens parametre benyttes igen MLE metoden, hvor likelihoodfunktionen jævnfør ligning(111) kan opskrives som

$$L(\theta : T_1, ...T_{N_u}) = e^{-n\tau - \psi \int_0^n v^*(s)ds} \prod_{j=1}^{N_u} \lambda^*(T_j, \tilde{X}_j). \tag{122}$$

Denne likelihood benyttes både for Hawkes POT med forudsigelige og uforudsigelige mærker, hvor intensiteten for den specifikke model indsættes, og likelihoodfunktionen maksimeres.

6.5 Risikomål

En Self-Exciting POT model kan også benyttes til at estimere risiko, såsom betinget VaR og ES, og modellen giver en anderledes tilgang til beregningerne sammenlignet med risikomålene i forbindelse

med GPD'en. Denne tilgang er beskrevet med reference til [20, McNeil et. al., 2015, 581].

VaR er stadig defineret som den p'te fraktil af en fordeling F, $\text{VaR}_p = F^{-1}(p)$, hvor VaR for et niveau p kan beregnes ved at løse ligningen

$$\tau^*(t+, x) = (1-p), \qquad x \geqslant u, \tag{123}$$

hvor $t+$ angiver den betingede overskridelsesintensitet til et tidspunkt lige efter t, hvor fraktilniveauet er $1-p$. For Hawkes POT med forudsigelige mærker er det kun muligt at løse udtrykket i ligning (123) hvis $\tau + \psi v^*(t+) > 1-p$. VaR estimatoren kan findes ved først at substituere udtrykket for $\tau^*(t+, x)$ ind i ligning (120), hvorefter udtrykket sættes lig med $(1-p)$ og x isoleres:

$$\text{VaR}_p^t = u + \frac{\beta + \alpha v^*(t+)}{\xi}\left(\left(\frac{1-p}{\tau + \psi v^*(t+)}\right)^{-\xi} - 1\right). \tag{124}$$

ES målet findes med udgangspunkt i at fordelingen af den betingede overskridelse over VaR_p^t, givet information op til tid t, er GP fordelt med formparameteren ξ og skalaparameteren $\beta + \alpha v^*(t+) + \xi(\text{VaR}_p^t - u)$. Det resulterer i, at ES kan skrives på formen

$$\text{ES}_p = \frac{\text{VaR}_p^t}{1-\xi} + \frac{\beta + \alpha v^*(t+) - \xi u}{1-\xi}. \tag{125}$$

De ovenstående risikomål er betingede VaR og ES mål, hvilke estimeres ved at fitte en Self-Exciting proces for overskridelsestiderne og værdierne i GPD modellen, hvorimod den ubetingede VaR er den der beregnes ud fra POT modellen.

6.6 Test af punktproces modeller

For at kunne sammenligne modellerne for punktprocesserne analyseres de ud fra Goodness-of-Fit tests i form af Akaikes og Bayesian informationskriterie: AIC og BIC. Grunden til at modellerne ikke vurdereres direkte ud fra værdien af den maksimerede likelihoodfunktion, er at den ikke tager højde

for modelkompleksitet, hvilket vil sige, at der kan vælges en unødig kompleks model. AIC og BIC benyttes til at finde et tradeoff mellem modelfit og modelkompleksitet, hvor det ønskes at maksimere fittet, og minimere kompleksiteten. Informationskriterierne kan med reference til [24, Ruppert, 2011, 103] beregnes ud fra formlerne

$$\begin{aligned} AIC &= -2\ln(L) + 2k \\ BIC &= -2\ln(L) + ln(n)k, \end{aligned} \tag{126}$$

hvor k angiver antallet af parametre i den pågældende model, n er antallet af observationer, og L angiver værdien af den maksimerede likelihoodfunktion.

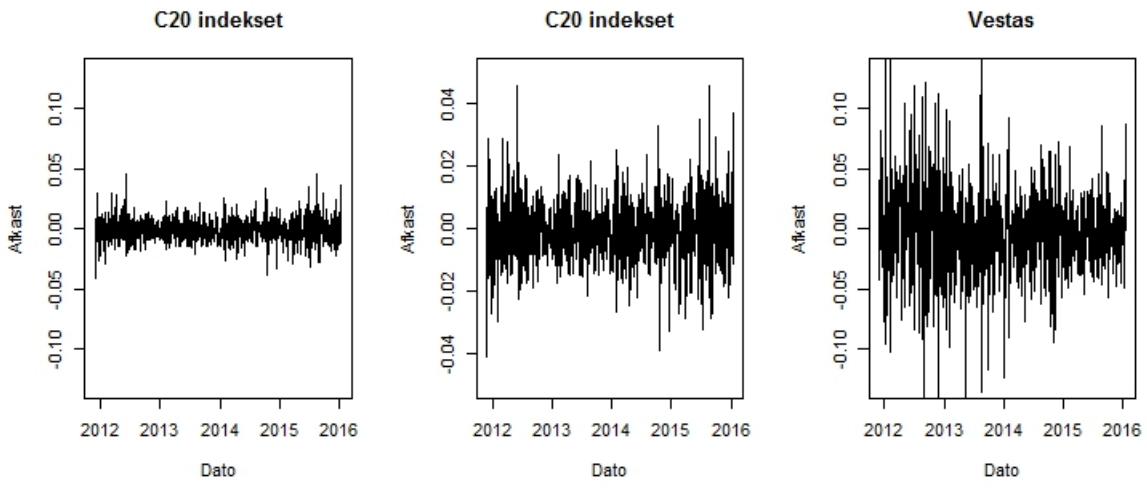
Del II

Analyse

7 Analyse af ekstremværdier

7.1 Data

Som tidligere beskrevet ønskes det i denne afhandling at modellere ekstreme tab på finansielt data. Den første tanke var at benytte negative afkast fra C20 indekset, for at finde ekstreme tab på det samlede danske aktiemarked. Da C20 indekset består af flere aktier, som er korrelerede, er afkastene ikke lige så volatile som på den enkelte aktie. I figur (10) er afkastene illustreret for henholdsvis C20 indekset og på en enkelt udvalgt aktie: Vestas aktien.

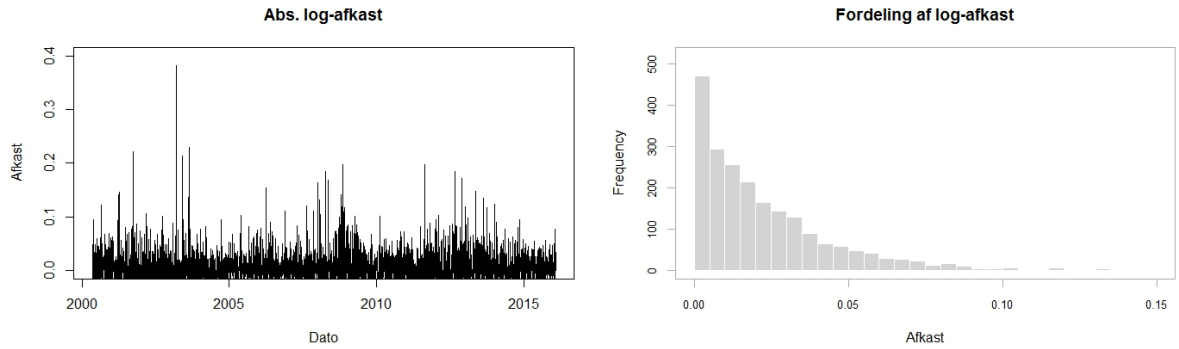


Figur 10: Log-afkast på C20 indekset og Vestas aktien i perioden 10.05.2000 til 28.01.2016.

Afkastene på C20 indekset er som forventet ikke lige så store som på Vestas aktien. Det største daglige tab på C20 indekset er -4%, hvor det på Vestas aktien er på -18%. Tabene skal ses relativt i forhold til resten af datasættet, hvor -4% på C20 indekset er ekstremt. Hvis der zoomes ind på C20 indekset, og sammenlignes med Vestas aktien, er den enkelte aktie mere volatil, hvilket vil give flere ekstreme værdier at modellere. I denne afhandling har vi derfor valgt at benytte log-afkast beregnet ud fra daglige 'åbne' kurser fra Vestas aktien i perioden 10.05.2000 til 28.01.2016. Log-afkastene er beregnet på følgende måde

$$r_t = \ln(S_t) - \ln(S_{t-1}) = \ln\left(\frac{S_t}{S_{t-1}}\right). \tag{127}$$

Da der kun ønskes at studere ekstreme tab, og ikke ekstreme gevinster, udtages kun de negative log-afkast, og der analyseres videre på dem som absolutte værdier. Dette giver i alt 2.129 observationer, hvor det vurderes, at det er muligt at opnå nok ekstreme værdier til at lave en valid analyse.

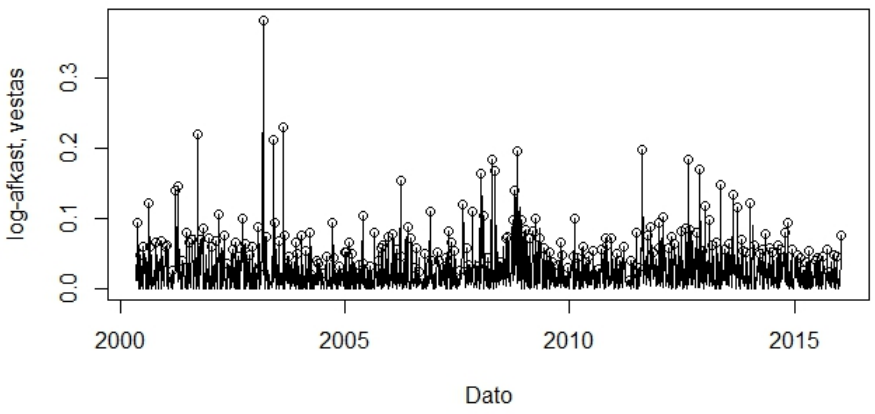


Figur 11: Absolutte værdier af negative afkast på Vestas aktien, samt histogram der viser fordelingen af log-afkastene.

Den første graf i figur (11) viser de absolutte værdier af de negative log-afkast, hvor det nu er tydeligere, at afkastene er meget volatile. Histogrammet viser som forventet, at størstedelen af

observationerne er lave, men antyder også, at fordelingen af data har en højre fed hale. Det kan på grund af denne højre fede haletendens tyde på, at EVT'en kan benyttes til at modellere de ekstreme værdier. Der skal som det første i analysen, bestemmes en metode til at specificere hvilke observationer i datasættet for Vestas aktien der er ekstreme.

Det bemærkes ud fra det første plot i figur (11) at de store udsving i datasættet har tendens til at forekomme lige efter hinanden, eller tæt på hinanden, og det er dermed relevant at undersøge, om der er tendens til klyngedannelse. I den første del af dette analyseafsnit antages der i.i.d. data, hvor der i den anden del af analyseafsnittet, ved hjælp af punktprocesser, undersøges om klyngetendensen er reel. Det kan dog stadig være afgørende for udvælgelsen af de ekstreme observationer hvordan data ser ud. I figur (12) er de månedlige Blok Maksima (BM) værdier hørende til de absolutte negative log-afkast for Vestas aktien illustreret.



Figur 12: Blok Maksima værdier for Vestas aktien på månedsbasis.

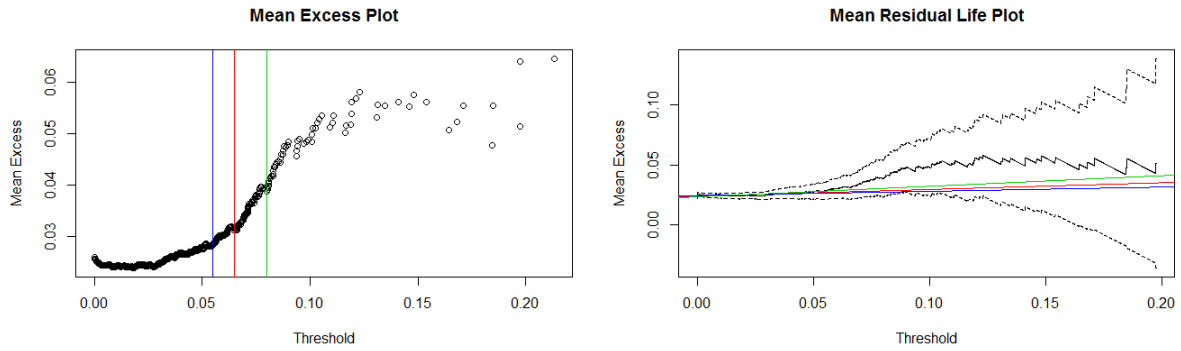
Vi har vurderet, at den naturlige opdeling af blokke er på månedsbasis, da årlig data kun vil give 15 ekstreme observationer, hvor den månedlige opdeling resulterer i 189 ekstreme observationer. Ved udvælgelse med BM metoden opfanges mange ekstreme observationer, men der er stadig

nogle høje observationer, som ikke er medtaget. Derudover udvælger BM metoden også en del lave observationer, som egentlig ikke er ekstreme.

For at kunne opfange alle de relevante ekstreme observationer i datasættet, benyttes POT metoden.

7.2 Fastsættelse af threshold

Ud fra de absolutte negative log-afkast i figur (11) observeres det, at Vestas datasættet har en del ekstreme observationer, og der ønskes nu at finde et threshold så det kan specificeres hvilke observationer der er ekstreme. ME plottet i figur (13), er grafisk illustreret ud fra `meplot` funktionen i `R`, og viser de gennemsnitlige overskridelser op imod forskellige thresholdværdier.



Figur 13: Mean Excess plot og Mean Residual Life plot.

Det huskes fra tidligere, at ME plottet skal være tilnærmelsesvis lineært for at overskridelsesobservationerne fitter en GPD. Valget af threshold er defineret som det $u > 0$, hvor ME plottet er tilnærmelsesvis lineært for værdier af $x \geq u$. Det er sjældent, at plottet er perfekt lineært, og dette eksempel er ingen undtagelse. I thresholdintervallet $[0,05; 0,09]$ kan der anes en lineær tendens, men også et lille knæk i plottet ved en thresholdværdi på omkring $0,055$, $0,065$ og $0,08$. Det vil resultere i et datasæt med henholdsvis 199, 130 og 70 observationer. Idet ME plottet har en positiv hældning,

forventer vi at få positive estimater for formparameteren ξ.

I figur (13) har vi, ud fra `mrlplot` funktionen i `R`, også illustreret MRL plottet, som 'teoretisk' angiver det samme som ME plottet, men grafisk er lidt anderledes. MRL plottet viser, udover de gennemsnitlige overskridelser for forskellige threshold værdier, også deres tilhørende 95% konfidensintervaller. I MRL plottet, er der for alle tre threshold værdier: $u = 0,055$, $u = 0,065$ og $u = 0,08$ fittet lineære regressionslinjer til den lineære del af plottet, som ligger før det valgte threshold. De lineære regressionslinjer ligger inden for 95% konfidensintervallet, hvorfor det også ud fra dette plot kan tyde på, at thresholdværdierne umiddelbart kan være fine valg.

Valget af threshold testes nu, ved hjælp af en Bootstrap Goodness-of-Fit test. I `R` benyttes `gpd.test` funktionen, og resultaterne er illustreret i tabel (2).

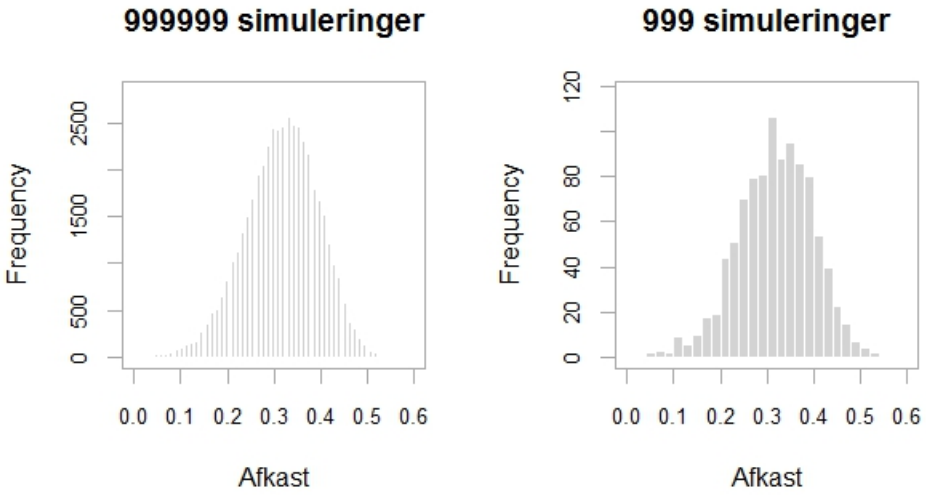
Threshold	$0,055$	$0,065$	$0,08$
P-værdi	$0,1131$	$0,2673$	$0,3654$
R-værdi	$0,9738$	$0,9760$	$0,9698$
R^2-værdi	$0,9483$	$0,9526$	$0,9405$
Nulhypotese	H_0^+	H_0^+	H_0^+

Tabel 2: Boostrap test af thresholdværdierne $u = 0,055$, $u = 0,065$ og $u = 0,08$.

Modellens nulhypotese H_0 er, at overskridelsesobservationerne fitter en GPD, og den er som tidligere forklaret inddelt i to delhypoteser H_0^+ og H_0^-. Det kan i tabel (2) aflæses, at alle tre thresholdværdier får p-værdier, som er større end signifikansniveauet på 5%. Det betyder, at hypotesen H_0^+ om at observationerne er GP fordelte med positiv formparameter, ikke kan afvises. I outputtet udskrives også R værdien, hvor R^2 værdien beregnes til at ligge på henholdsvis $R^2_{u=199} = 0,948$, $R^2_{u=130} = 0,953$ og $R^2_{u=70} = 0,941$, hvilket er fine værdier tæt på en. For at understøtte disse valg af thresholds, er Bootstrap-testen udført for forskellige værdier, som ligger tæt på $0,055$, $0,065$ og $0,08$. Dette resulterede dog ikke i pænere R^2 værdier, hvorfor analysen fortsættes med disse tre thresholdværdier.

7.3 Estimation af parametre

Der er nu fundet tre mulige grænser, som thresholdværdien potentielt kan fastsættes til, hvormed det nu er muligt at fitte de overskredne observationer til GPD'en. Vi har benyttet fem forskellige metoder til at estimere de ukendte parametre i GPD'en, og det ønskes at studere disse parameterestimater, de tilhørende Standard Errors (SE) og konfidensintervaller. Vælger man at beregne konfidensintervaller ud fra den generelle formel i (93), antages asymptotiske normalfordelte estimater. For at undersøge om denne antagelse er opfyldt, har vi i figur (14) illustreret fordelingen af de asymptotiske ξ estimater fundet ved MLE metoden for henholdsvis 999999 og 999 simuleringer.



Figur 14: Asymptotiske fordeling af Bootstrap estimater for 999999 og 999 simuleringer.

I figur (14) bekræftes det at for $n \to \infty$, går fordelingen af de asymptotiske estimater mod en normalfordeling. I denne afhandling benyttes mange estimationsmetoder, hvoraf en af dem er numerisk, og dermed en beregningstung metode. På baggrund af dette vælger vi at benytte færre simuleringer, 999, og estimerer konfidensintervallerne ved hjælp af Bootstrap metoden, så eventuelle fede haler og anormalitet opfanges.

Til at estimere de ukendte parametre i GPD'en, har vi som tidligere nævnt benyttet MLE, PWM, EPM, MOM og LMOM metoderne. MLE og PWM estimaterne er beregnet ud fra `gpd` funktionen i `R`, hvor der i funktionen specificeres, hvilken metode der ønskes benyttet. EPM estimaterne er fundet ud fra træet i figur (8), som angiver strukturen af algoritmen. Vi har implementeret denne algoritme i `R`. MOM estimaterne er beregnet ud fra lukkede formler, og LMOM momenterne er fundet ud fra `lmom` funktionen, hvorudfra estimaterne beregnes ved hjælp af `pelgpa` funktionen. Parameterestimaterne samt tilhørende SE og 95%-Bootstrap konfidensintervaller er illustreret i tabel (3). SE og konfidensintervallerne er beregnet ud fra 999 Bootstrap simuleringer, som er beregnet ud fra tre datasæt bestående af henholdsvis 199 observationer for $u = 0,055$, 130 observationer for $u = 0,065$, og 70 observationer for $u = 0,08$.

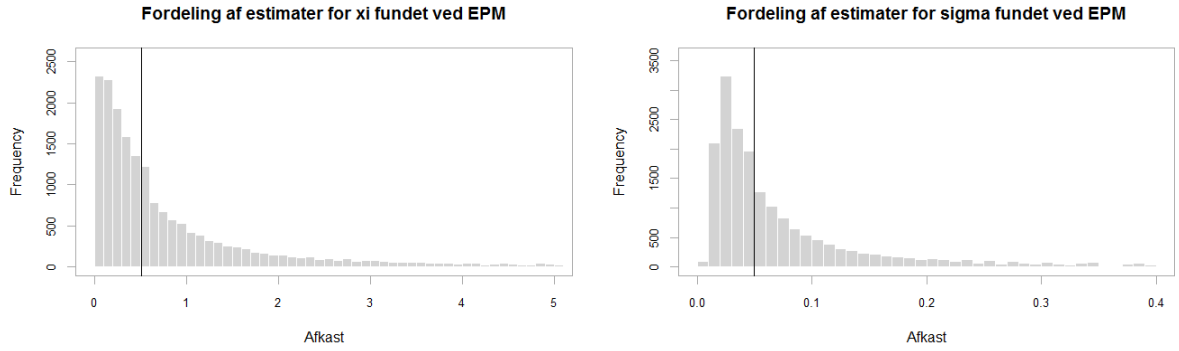
u		0,055		0,065		0,08	
		ξ	σ	ξ	σ	ξ	σ
MLE	Est.	$0,293$	$0,020$	$0,359$	$0,020$	$0,373$	$0,025$
	SE	$(0,092)$	$(0,002)$	$(0,108)$	$(0,003)$	$(0,215)$	$(0,008)$
	KI	$[0,141; 0,450]$	$[0,017; 0,025]$	$[0,121; 0,622]$	$[0,015; 0,028]$	$[-0,048; 0,760]$	$[0,011; 0,042]$
PWM	Est.	$0,285$	$0,020$	$0,335$	$0,021$	$0,335$	$0,026$
	SE	$(0,094)$	$(0,002)$	$(0,079)$	$(0,003)$	$(0,115)$	$(0,006)$
	KI	$[0,118; 0,406]$	$[0,017; 0,025]$	$[0,156; 0,467]$	$[0,016; 0,027]$	$[0,080; 0,532]$	$[0,014; 0,045]$
EPM	Est.	$0,509$	$0,050$	$0,500$	$0,057$	$0,453$	$0,069$
	SE	$(0,261)$	$(0,018)$	$(0,275)$	$(0,022)$	$(0,322)$	$(0,030)$
	KI	$[0,713; 0,832]$	$[0,060; 0,076]$	$[0,707; 0,845]$	$[0,068; 0,091]$	$[0,686; 0,880]$	$[0,080; 0,121]$
MOM	Est.	$1,749$	$0,230$	$1,949$	$0,285$	$2,307$	$0,391$
	SE	$(0,730)$	$(0,053)$	$(0,898)$	$(0,077)$	$(1,284)$	$(0,136)$
	KI	$[0,978; 3,576]$	$[0,174; 0,363]$	$[1,042; 4,053]$	$[0,208; 0,456]$	$[1,212; 5,342]$	$[0,275; 0,735]$
LMOM	Est.	$0,301$	$0,0198$	$0,314$	$0,022$	$0,218$	$0,032$
	SE	$(0,019)$	$(0,000)$	$(0,019)$	$(0,000)$	$(0,013)$	$(0,000)$
	KI	$[0,140; 0,4209]$	$[0,016; 0,026]$	$[0,112; 0,0,447]$	$[0,0162; 0,031]$	$[0,088; 0,403]$	$[0,0213; 0,053]$

Tabel 3: Parameterestimater (Est.) for ξ og σ samt tilhørende Standard Errors (SE) og 95%-konfidensintervaller (KI) for de fem estimationsmetoder.

Det ses i tabel (3) at alle ξ estimater er positive, hvilket også var forventeligt ud fra den positive hældning i ME plottet. Estimaterne for ξ og σ for $u = 0,065$ og $u = 0,08$ fundet ud fra MLE og PWM metoderne ligger pænt på samme niveau, med kun små afvigelser fra hinanden, hvorimod estimaterne for $u = 0,055$ er en smule lavere for begge metoder. Estimaterne fundet ved hjælp af LMOM metoden ligger på nogenlunde samme niveau som MLE og PWM estimaterne for $u = 0,055$ og $u = 0,065$,

men afviger for $u = 0,08$, hvor $\xi = 0,218$ og $\sigma = 0,032$. Estimaterne fra EPM metoden er generelt højere for alle thresholdværdier, og estimaterne fra MOM metoden er væsentligt højere. Hvis de tilhørende SE værdier studeres, har LMOM metoden overordnet de mindste værdier, efterfulgt af MLE og PWM, som ligger meget tæt. Dette er også forventeligt, da estimaterne i metoderne ligger så tæt. MOM har meget høje SE værdier, hvilket kan skyldes, at de tilhørende estimater er en del højere end estimaterne fundet ud fra de andre estimationsmetoder, hvilket kan skyldes en højere varians. EPM metoden har også en forholdsvis høj SE værdi.

EPM metoden er som tidligere nævnt en numerisk metode, som beregner et estimat for alle sæt af observationer (i, j), og finder det endelige estimat ved medianerne af disse. I figur (15) er der illustreret to histogrammer, som viser fordelingen af estimaterne fundet for alle kombinationer af i og j. Den vertikale linje i histogrammerne illustrerer parameterestimaterne for en tresholdværdi på $0,055$, hvor $\xi = 0,509$ og $\sigma = 0,050$.



Figur 15: Fordeling af estimater ξ og σ for kombinationer af alle par af i og j i EPM metoden.

Ud fra figur (15) ses det, at spredningen for begge estimater er stor, hvilket kan have indflydelse på værdien af parameterestimaterne. I EPM metoden har vi til at beregne estimaterne benyttet en forløkke i `R`, som kører over alle kombinationer af (i, j), $i > j$, og beregner estimatet af formparameteren ξ ved hjælp af formlen

$$\hat{\xi}(i,j) = \frac{\ln(1 - x_{i:n}/\hat{\delta}(i,j))}{C_i}. \tag{128}$$

Udfordringen i dette tilfælde er, at metoden for nogle kombinationer af i og j får et udtryk $1 - x_{i;n}/\delta < 0$, hvor det ikke er muligt at beregne logaritmen. Det har resulteret i, at en del kombinationer ikke har kunne medtages i porteføljen af estimater, da forløkken har måtte springe dem over. Ved $u = 0,055$ drejer det sig om i alt 19.419 kombinationer ud af 39.205 kombinationer, for $u = 0,065$ om 8.296 kombinationer ud af 16.642 kombinationer, og for et threshold på $0,08$ drejer det sig om i alt 2.355 kombinationer ud af 4.762 kombinationer. Antallet af fejlberegninger er altså lige under halvdelen af beregningerne for alle tre thresholds, hvilket kan give et stort udsving i de endelige parameterestimater, og være grunden til at SE værdierne i denne metode er høje.

For at få et samlet overblik over præcisionen og om hvor gode parameterestimaterne er i forhold til den 'sande' værdi, er Bias og MSE for de tre thresholdværdier illustreret i tabel (4).

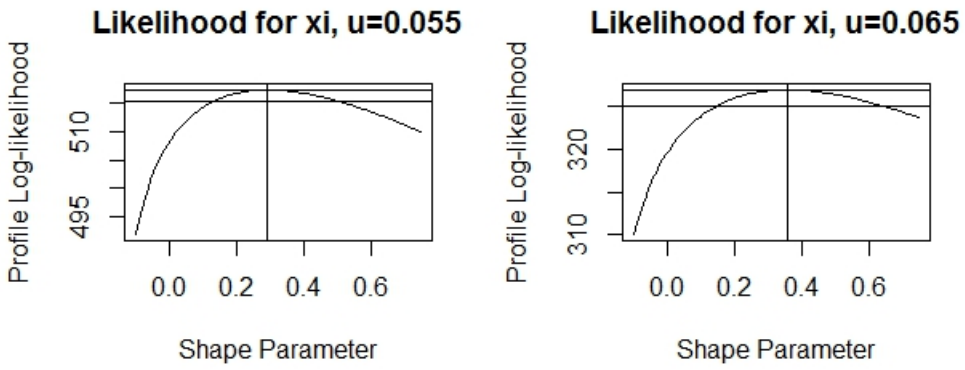
u		0,055		0,065		0,08	
	Estimat	ξ	σ	ξ	σ	ξ	σ
MLE	Bias	$0,0073$	$-0,0002$	$0,01537$	$-0,0004$	$0,0219$	$-0,0014$
	MSE	$0,0085$	0	$0,0120$	0	$0,0467$	0
PWM	Bias	$0,0157$	$-0,0003$	$0,0235$	$-0,0006$	$0,0289$	$-0,0013$
	MSE	$0,0091$	0	$0,0068$	0	$0,0141$	0
EPM	Bias	$-0,2585$	$-0,0176$	$-0,2627$	$-0,0214$	$-0,3182$	$-0,0285$
	MSE	$0,1347$	$0,0006$	$0,1499$	$0,0010$	$0,2048$	$0,0017$
MOM	Bias	$-0,2043$	$-0,0138$	$-0,3183$	$-0,0286$	$-0,4669$	$-0,0546$
	MSE	$0,5748$	$0,0030$	$0,9073$	$0,0068$	$1,8670$	$0,0214$
LMOM	Bias	$0,015$	$-0,0005$	$0,0241$	$-0,0006$	$0,0395$	$-0,0021$
	MSE	$0,00057$	0	$0,0009$	0	$0,0017$	0

Tabel 4: Bias og Mean Squared Error af de fundne parameterestimater for de fem estimationsmetoder: MLE, PWM, EPM, MOM og LMOM.

I MLE, PWM og LMOM metoden er både Bias og MSE for alle tre thresholds betydeligt lavere end for de to andre metoder. Især MOM har en høj MSE for ξ estimaterne, hvilket afspejler estimaterne fundet i tabel (3). Der er ingen af parameterestimaterne som er unbiased, men mange af

estimaterne har en Bias forholdsvis tæt på nul. Hvis der udelukkende analyseres på MSE værdien som både tager højde for bias og den naturlige varians, er det for alle estimationsmetoderne data, som ligger over et threshold på $0,055$, som fittes bedst til GPD'en. EPM parameterestimaterne i tabel (3) afveg ikke meget fra MLE, PWM og LMOM, sammenlignet med MOM estimaterne, men den har en høj MSE på henholdsvis $0,1347$, $0,1499$ og $0,2048$ for ξ estimaterne. Denne høje MSE, og dermed usikkerhed i estimaterne, kan skyldes fejlberegningerne i algoritmen, og vi vurderer derfor, at denne metode ikke er valid til at beskrive datasættet. MOM estimationsmetoden afskrives også på baggrund af estimaternes høje SE og MSE.

MLE metoden har den fordel, at det er muligt at validere resultatet ved at studere likelihoodfunktionen nærmere. I figur (16) er likelihoodfunktionen for ξ hørende til en thresholdværdi på henholdsvis $0,055$ og $0,065$ illustreret.



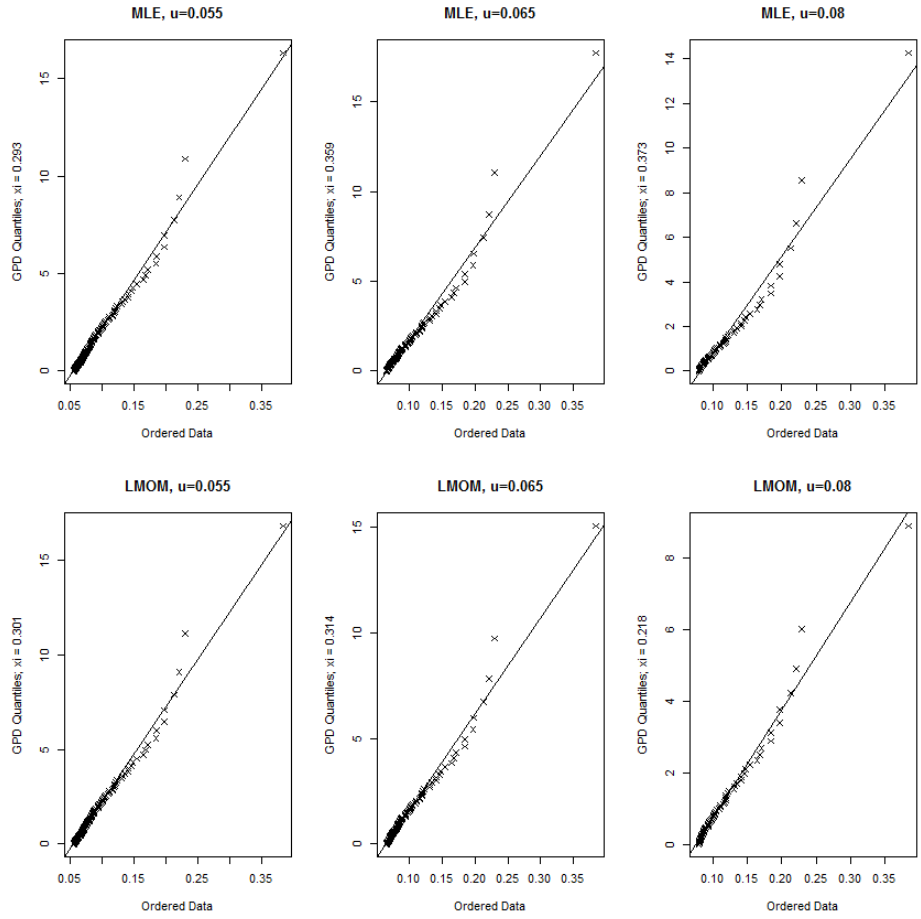
Figur 16: Likelihoodfunktion for ξ, med threshold $u = 0,055$ og $u = 0,065$.

Begge likelihoodfunktioner har kun ét toppunkt, hvilket bekræfter, at det er det korrekte toppunkt, som likelihooden maksimeres ud fra. Den vertikale linje som går igennem likelihoodfunktionens toppunkt, angiver parameterestimatet ξ for begge thresholdværdier. Det bemærkes at likelihoodfunktionen for $u = 0,065$ har et fladere toppunkt end for $u = 0,055$, hvilket betyder, at der er en højere varians på estimatet. Den nederste horisontale linje angiver i begge plots 95%-

konfidensintervallet, hvor der også her er en mindre varians på estimatet for $u = 0,055$ sammenlignet med $u = 0,065$. Havde datasættet bestået af flere ekstreme observationer kunne likelihoodfunktionen have været mere spids, og estimatet mere præcist. Da der i denne afhandling arbejdes med EVT, hvor man sjældent modellerer mange observationer, kan estimatet accepteres på baggrund af at funktionen er maksimeret korrekt.

I figur (17) er QQ-plots for MLE og LMOM metoderne for henholdsvis $u = 0,055$, $u = 0,065$ og $u = 0,08$ illustreret. Vi har valgt ikke at medtage PWM metoden, idet estimaterne og de tilhørende Bias og MSE værdier, samt QQ-plots er tilnærmelsesvis identiske med MLE metoden. Vi fortsætter dermed analysen med fokus på MLE og LMOM metoden.

QQ-plottene i figur (17) for begge metoder ligner hinanden meget for alle tre thresholdværdier, men der kan for begge anes et bedre fit for $u = 0,055$, idet overskridelsesobservationerne ligger en smule tættere på den rette linje. Der ses dog en afvigelse fra GPD'en i alle seks QQ-plots, hvor det er de højeste ekstreme observationerne som afviger. Denne afvigelse kan tyde på fede haler, og at der som tidligere nævnt, ikke nødvendigvis kan antages, at overskridelsesobservationerne er i.i.d. Har observationerne i datasættet tendens til en form for klyngedannelse, kan det medføre udsving i QQ-plottene, da klynger i data vil kunne skabe flere ekstreme observationer som en form for efterskælv.

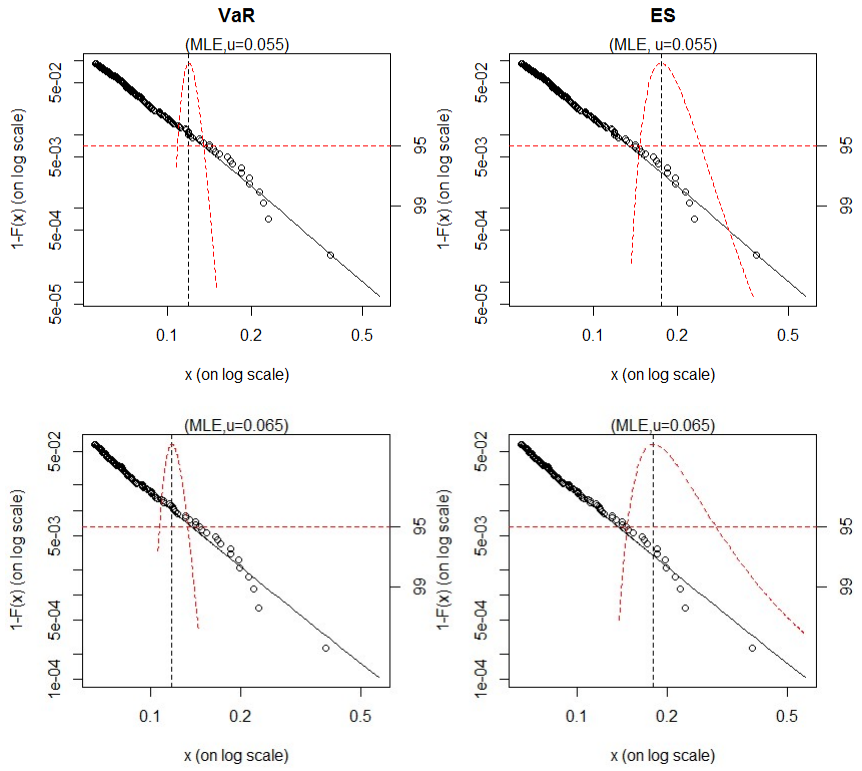


Figur 17: De tre øverste plots er QQ-plots for MLE metoden, for henholdsvis $u = 0,055$, $u = 0,065$ og $u = 0,08$, og de tre nederste QQ-plots er for LMOM metoden.

Ud fra resultaterne i tabel (4) og figur (17) er der to estimationsmetoder, som viser et acceptabelt modelfit: MLE og LMOM metoden. Da MSE værdierne er bedre for en thresholdværdi på $0,055$ og $0,065$ sammenlignet med $0,08$, vælger vi at fortsætte analysen for begge metoder med disse thresholdværdier.

7.4 Estimation af risikomål

Overskridelsesobservationerne er nu fittet til GPD'en ved hjælp af MLE og LMOM metoden for to thresholdværdier $u = 0,055$ og $u = 0,065$, og det ønskes at estimere de tilhørende VaR og ES risikomål. I `R` er funktionerne `gpd.q` og `gpd.sfall` benyttet til at beregne, og grafisk illustrere, risikomålene for MLE metodens estimater. Funktionerne udregner estimater for VaR og ES samt konfidensintervaller for høje fraktiler, som ligger over en fastsat thresholdværdi i en GPD. Det sikres derfor at VaR målet bliver beregnet på baggrund af GP fordelt data. Graferne for $\text{VaR}_{0,99}$, $\text{ES}_{0,99}$ og 95% konfidensintervallerne for MLE metoden for $u = 0,055$ og $u = 0,065$ er illustreret i figur (18).



Figur 18: $\text{VaR}_{0,99}$, $\text{ES}_{0,99}$ og 95% konfidensintervaller for MLE metoden for $u = 0,055$ og $u = 0,065$.

De to øverste grafer viser $\text{VaR}_{0,99}$ og $\text{ES}_{0,99}$ estimatet samt 95% konfidensintervallet for GPD

fittet for $u = 0,055$, og de to nederste grafer illustrerer de tilsvarende VaR og ES estimater for $u = 0,065$. Linjen gennem punkterne viser den estimerede hale-estimator, som angivet i ligning (87), og punkterne på grafen angiver overskridelsobservationerne. Den vertikale prikkede linje viser henholdsvis VaR og ES estimaterne, og de to skæringspunkter mellem den horisontale prikkede linje og den prikkede kurve angiver på samme måde 95%-konfidensintervallet for VaR og ES estimatet.

VaR og ES risikomålene for LMOM metoden er direkte beregnet ud fra formlerne i ligning (88), idet de indbyggede funktioner i `R` ikke kan benyttes. De indbyggede risikofunktioner kræver, at estimationsmetoden har fittet overskridelsesobservationerne til GPD'en ved hjælp af `gpd` funktionen, som ikke er muligt for LMOM metoden men for MLE.

De fundne estimater for VaR og ES for både et 95% og 99% fraktil niveau samt konfidensintervallerne (KI) er for begge thresholdværdier vist i tabel (5).

			95% fraktil	95% KI	99% fraktil	95%KI
0,055	MLE	VaR	$0,0690$	$[0,0666; 0,0730]$	$0,1194$	$[0,1088; 0,1347]$
		ES	$0,1036$	$[0,0952; 0,1196]$	$0,1748$	$[0,1474; 0,2414]$
	LMOM	VaR	$0,0687$	$[0,0661; 0,0721]$	$0,1183$	$[0,1064; 0,1320]$
		ES	$0,1029$	$[0,0935; 0,1143]$	$0,1739$	$[0,1417; 0,2076]$
0,065	MLE	VaR	$0,0693$	[Inf;-Inf]	$0,1180$	$[0,1074; 0,1331]$
		ES	$0,1041$	$[0,0959; 0,1269]$	$0,1800$	$[0,1494; 0,2898]$
	LMOM	VaR	$0,0695$	$[0,0683; 0,0714]$	$0,1185$	$[0,1059; 0,1336]$
		ES	$0,1036$	$[0,0945; 0,1138]$	$0,1749$	$[0,1443; 0,2077]$

Tabel 5: VaR og ES estimater samt tilhørende konfidensintervaller (KI) for MLE og PWM estimationsmetoderne.

Estimaterne for $\text{VaR}_{0,99}$ og $\text{ES}_{0,99}$ for begge thresholdværdier og estimationsmetoder er forholdsvis ens, da de først afviger på den anden eller tredje decimal, og det samme er gældende for 99% fraktilen. I plottet i figur (18) indikerer den horisontale prikkede linje det valgte konfidensniveau, og det kan her aflæses, at forskellen mellem 95% og 99% konfidensintervallet for VaR er meget lille. Modsat ligger forskellen mellem konfidensintervallerne for ES i værdien af den øvre grænse. Denne forskel kan indikere, at fordelingen af ES estimaterne har fede haler sammenlignet med fordelingen

af VaR estimaterne, hvilket også kan aflæses ud fra den prikkede kurves højre side i figur (18).

VaR målet er som tidligere beskrevet, et mål for hvor meget værdien af et aktiv kan falde over en given tidsperiode med en given sandsynlighed. Ved brug af MLE metoden for $u = 0,065$ fås for eksempel et 99% VaR estimat på $0,1180$, hvilket betyder at det med 99% sikkerhed forventes, at Vestas aktien maksimalt vil have et forventet negativt afkast på $0,1180$ på dagsbasis. Ved LMOM metoden for $u = 0,065$ fås et næsten identisk estimat på $0,1185$. Da datasættet består af daglige aktiekurser, er det en-dags VaR der estimeres. Disse risikomål indikerer, at man med kun 1% sandsynlighed kan forvente at tabe mere end VaR estimaterne på $0,118$ og $0,117$ på daglig basis. Det samme er selvfølgelig gældende for $u = 0,055$.

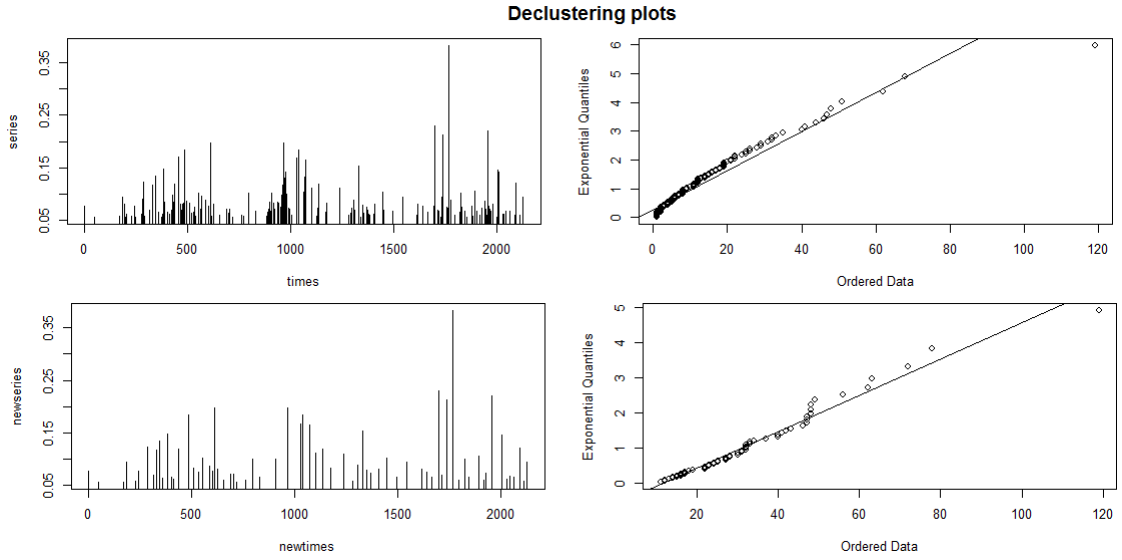
VaR målet siger ikke noget om hvor stort et tab man kan forvente, hvorfor det er relevant at studere de tilhørende ES mål. De fundne estimater for $ES_{0,99}$ beskriver størrelsen af det tab, som man med 1% sandsynlighed kan forvente. Størrelsen af tabet vil for MLE metoden for $u = 0,065$ være på $0,18$, og $u = 0,055$ være på $0,1748$. VaR og ES estimaterne for begge metoder er nærmest ens, og afviger først på anden eller tredje decimal, hvilket ikke er overraskende, da parameterestimaterne også ligger tæt på hinanden. Det giver derfor ikke en betydelig forskel for VaR og ES målene, om MLE eller LMOM metoden benyttes.

Vi har indtil nu studeret de ekstreme observationer med antagelsen om i.i.d. data, og fået en indikation af at overskridelsesobservationerne i Vestas datasættet kan antages at være GP fordelte, og parameterestimaterne ud fra Bias og MSE værdierne kan antages at være valide. På den anden side viser QQ-plottene i figur (17), at der findes haleobservationer som med antagelsen om i.i.d. data ikke kan forklares, og denne problemstilling ønsker vi at studere nærmere ved at tage hånd om eventuel klyngedannelse i datasættet. Dette gøres ved at fitte overskridelsesobservationerne til GPD'en ved implementering af punktprocesser, og undersøge om det kan give et bedre fit.

7.5 Punktprocesser

Det er nu studeret, hvordan ekstreme værdier med antagelsen om i.i.d. data kan modelleres i praksis, og det ønskes at se om GPD fittet kan forbedres ved at tage højde for eventuel klyngedannelse. Der fokuseres i dette afsnit kun på thresholdværdien $u = 0,055$, og der undersøges om en bedre model og dermed bedre fit, kan opnås.

Klyngedannelse i data kan afhjælpes på flere måder, hvor en af de simpleste metoder er 'runs declustering'. Denne metode går ud på, at der tilføres et filter til data, så efterskælvene i datasættet filtreres ud, og der opnås et datasæt med tilnærmelsesvis i.i.d. overskridelser. I metoden vælges et minimum interval for hvor stor adskillelsen mellem observationerne må være, også kaldet en 'run' længde. I figur (19) ses øverst de oprindelige overskridelsesobservationer over thresholdværdien $0,055$, hvor der kan anes en tendens til klynger. Studeres det øverste QQ-plot af de eksponentielle kvartiler, altså GPD'en med $\xi = 0$, er der en tendens til, at data ligger over den rette linje.

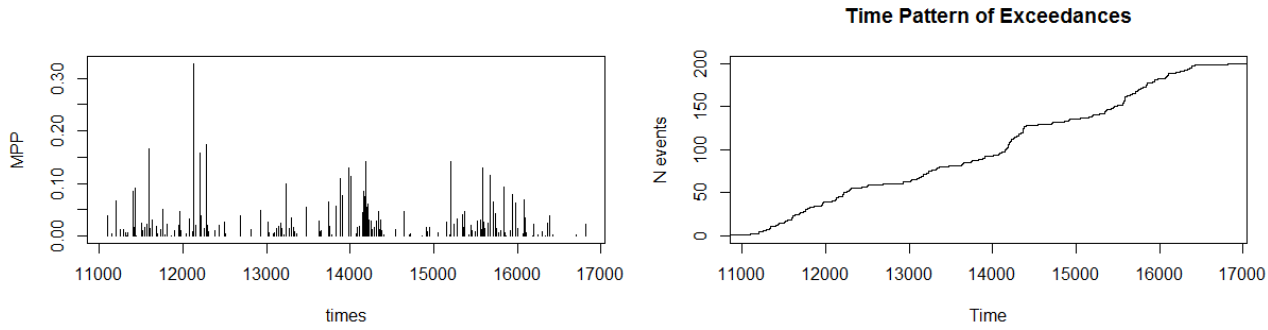


Figur 19: Effekt af 'declustering'. Plot af overskridelser hvor 'times' udgør antal dage i tidsperioden fra 10.05.2000 til 28.01.2016 og QQ-plot for Vestas aktien med $u = 0,055$ og run= 10.

Der er i 'declustering' metoden benyttet en 'run' længde på 10, og resultaterne af dette er

illustreret i de to nederste plots i figur (19). Effekten af at benytte 'declustering' metoden er at observationerne ligger pænere spredt omkring den rette linje. Metoden benyttes dog ikke, da vi vurderer, at det ikke vil give en valid analyse af de ekstreme observationer, idet for mange værdier vil blive trukket ud af analysen. I dette eksempel, med et 'run' på 10, reduceres datasættet fra 199 observationer til 69 observationer, hvilket ikke vil give et reelt billede af hvordan det oprindelige data egentlig er fordelt. Vi prøver derfor at afhjælpe udfordringen med klynger i data ved hjælp af punktprocesser, og undersøger hvilken effekt dette kan have på de ekstreme observationer.

Til at estimere den mærkede Poisson punktproces benyttes funktionen `extremalPP` i `R`, som også danner grundlaget for Hawkes POT modellerne med forudsigelige og uforudsigelige mærker. Tidsperioden vil ud fra denne funktion nu udgøres af det numeriske dataformat i `R`, hvor perioden fra den 10.05.2000 til den 28.01.2016 kan angives som intervallet $[11099; 16818]$.



Figur 20: 199 overskridelser i en mærket Poisson punktproces, samt den kumulerede hyppighed.

Det første plot i figur (20) illustrerer mærkerne som overskrider thresholdet på $u = 0,055$ i en Poisson punktproces, hvor der både i højere og lavere grad er en tendens til klynger. Specielt i tidsintervallet fra 14.000 til 14.500, som angiver år 2008, bemærkes en særlig stor klyngetendens. I det andet plot vises den kumulerede hyppighed af overskridelserne, som ved i.i.d. data tilnærmelsesvis vil være en ret linje. Det observeres i dette plot også, at der i år 2008 forekommer klynger, idet at linjen er stejlere i denne periode. Det vil sige, jo mindre lineær linjen er, jo flere observationer

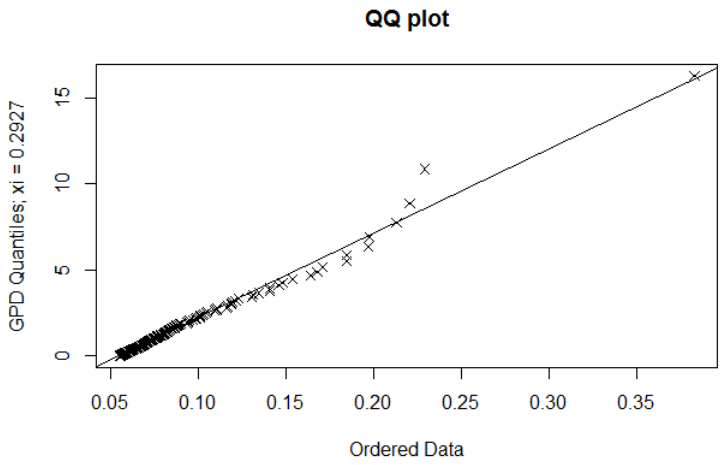
klynger sig sammen i den pågældende tidsperiode.

Den første model som i denne analyse er studeret, er POT modellen med en homogen Poisson proces, hvor de tilhørende parameterestimater er givet i tabel (6).

Model	Overskridelse	Estimat (SE)
POT model	Homogen Poisson proces	$\xi = 0,2928$ $(0,0674)$ $\sigma = 0,0102$ $(0,0021)$ $\tau = 0,0201$ $(0,0047)$ $\beta = 0,0204$
Self-Exciting (Hawkes POT) uforudsigelige mærker	Intensiteten afhænger af tiden siden forgående observation $(t - T_j)$ og størrelsen af observationen $(\tilde{X}_j - u)$.	$\tau = 0,0156$ $(0,0042)$ $\psi = 0,0072$ $(0,0023)$ $\gamma = 0,0129$ $(0,0043)$ $\xi = 0,2931$ $(0,0920)$ $\beta = 0,0204$ $(0,0023)$
Self-Exciting (Hawkes POT) forudsigelige mærker	Intensiteten afhænger af tiden siden forgående observation $(t - T_j)$ og størrelsen af observationen $(\tilde{X}_j - u)$.	$\tau = 0,0156$ $(0,0042)$ $\psi = 0,0072$ $(0,0023)$ $\gamma = 0,0129$ $(0,0043)$ $\xi = 0,2943$ $(0,0923)$ $\beta = 0,0197$ $(0,0038)$ $\alpha = 0,0002$ $(0,0008)$

Tabel 6: Parameterestimater for POT og Hawkes POT med uforudsigelige og forudsigelige mærker.

Til at estimere parametrene i POT modellen har vi i `R` benyttet `pot`-funktionen som fitter overskridelsesobservationerne til en Poisson punktproces, og følgende parameterestimater opnås $\xi = 0,2928$, $\sigma = 0,0102$, $\tau = 0,0201$ og $\beta = 0,0204$. Den betingede GPD's skalaparameter β findes som tidligere beskrevet ud fra formlen $\beta = \sigma + \xi(u - \mu)$, og er i dette tilfælde lig med $0,0204$. Det betyder, at der tilnærmelsesvis er opnået de samme estimater for form-og skalaparameteren, som blev fundet i POT modellen med et GPD fit. I QQ-plottet i figur (21) observeres den samme tendens som i den tidligere analyse, hvor klyngedannelse stadig kan være forklaringen på afvigelsen fra linjen i højre side af plottet.



Figur 21: QQ-plot af mærket Poisson punktproces.

Der er derfor i denne del af analysen modelleret yderligere to punktprocesser: Hawkes POT med forudsigelige og uforudsigelige mærker, hvortil `fit.seMPP` funktionen i `R` er benyttet, som fitter en mærket Self-Exciting proces til en mærket Poisson punktproces. I funktionen udspecificeres valget af metode, som i dette tilfælde er en Hawkes model, og om hvorvidt fokus er på en model med forudsigelige eller uforudsigelige mærker. I tabel (6) bemærkes det, at ξ og β estimaterne for begge Hawkes POT modeller afviger en smule fra hinanden, idet parametrene i den mærkede Hawkes POT model med uforudsigelige mærker fås til $\xi = 0,2931$ og $\beta = 0,0204$ og parametrene i modellen med forudsigelige mærker er givet ved $\xi = 0,2943$ og $\beta = 0,0197$. Parametrene ψ og γ er identiske, med en værdi på henholdsvis $0,0072$ og $0,0129$. Det var forventeligt at få en positiv værdi af ψ, for hvis $\psi = 0$ havde det resulteret i en standard POT model uden en 'Self-Exciting' struktur. I Hawkes POT modellen med forudsigelige mærker fremkommer parameterestimatet for α, $0,0002$, som indikerer, at mærkerne har indflydelse på modellen. Er parameterestimatet $\alpha = 0$, vil en Hawkes POT model med uforudsigelige mærker opnås. Værdien af α er dog meget lille, hvorfor vi ikke forventer store afvigelser mellem de to modeller videre i analysen.

Model	VaR	ES
POT model	$\text{VaR}_{0,95} = 0,0690$ $\text{VaR}_{0,99} = 0,1193$	$\text{ES}_{0,95} = 0,1036$ $\text{ES}_{0,99} = 0,1748$
Self-Exciting (Hawkes POT) uforudsigelige mærker	$\text{VaR}_{0,95} = 0,0505$ $\text{VaR}_{0,99} = 0,0897$	$\text{ES}_{0,95} = 0,0774$ $\text{ES}_{0,99} = 0,1329$
Self-Exciting (Hawkes POT) forudsigelige mærker	$\text{VaR}_{0,95} = 0,0505$ $\text{VaR}_{0,99} = 0,0897$	$\text{ES}_{0,95} = 0,0774$ $\text{ES}_{0,99} = 0,1330$

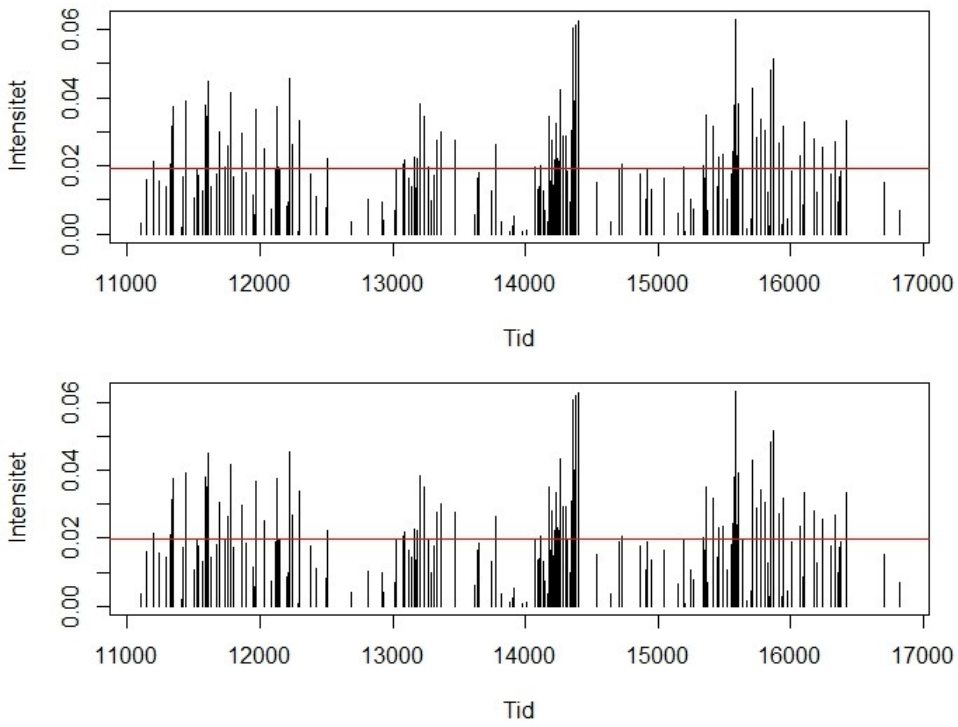
Tabel 7: VaR og ES estimater POT og Hawkes POT med uforudsigelige og forudsigelige mærker.

Vi har ud fra de tre modeller for punktprocesser beregnet VaR og ES på både 95% og 99% niveau. Som forventet opnår POT modellen samme estimater for både VaR og ES som i POT modellen med GPD fit. Niveauet for VaR og ES i Hawkes POT modellerne er lavere end i POT modellen, og næsten ens. Den eneste afvigelse er ES værdien på et 99% niveau, hvor værdien afviger på tredje decimal. Der opnås dermed den samme risikovurdering, lige meget om modellen har forudsigelige mærker eller ej.

Det huskes fra teorien at intensiteten hørende til en Hawkes POT model med forudsigelige mærker, er givet ved

$$\lambda^*(t,x) = \frac{\tau + \psi v^*(t)}{\beta + \alpha v^*(t)} \left(1 + \xi \frac{x-u}{\beta + \alpha v^*(t)}\right)^{-1/\xi - 1}, \tag{129}$$

hvor intensiteten hørende til Hawkes POT modellen med uforudsigelige mærker kan opskrives på samme form med $\alpha = 0$. Denne forskel observeres i parameterestimaterne, hvor modellen med forudsigelige mærker som beskrevet har $\alpha = 0,0002$. I figur (22) er intensiteten hørende til de to modeller, samt middelværdien illustreret.



Figur 22: Plot af intensitet hørende til Hawkes POT model med forudsigelige mærker (øverst) og uforudsigelige mærker (nederst).

De to Hawkes POT modeller er næsten ens, hvor middelværdien af intensiteten for den uforudsigelige model er $0,01968$, og for den forudsigelige model er $0,01939$. De afviger altså først på fjerde decimal, hvilket afspejles i den meget lave α-værdi som i dette tilfælde har en meget lille indflydelse på modellen.

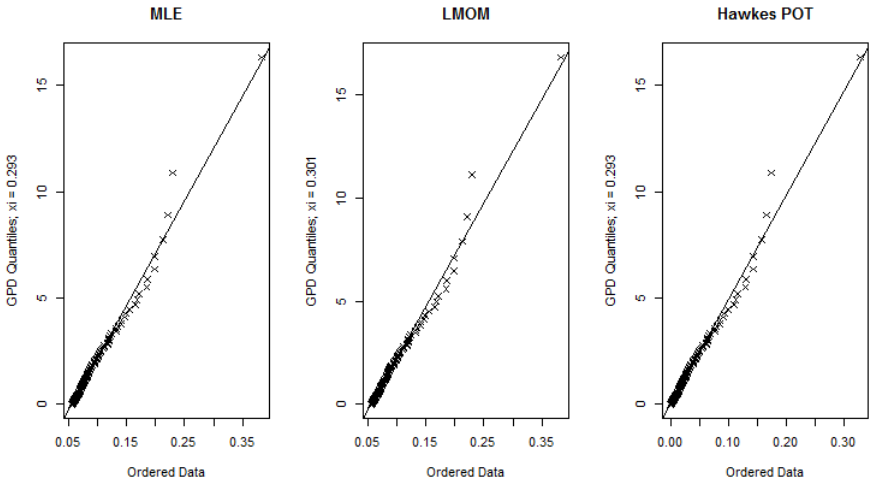
Det ønskes nu at sammenligne POT modellen, og de to Hawkes POT modeller med forudsigelige og uforudsigelige mærker ud fra AIC og BIC målene. Som beskrevet angiver k antallet af parametre i den pågældende model, n er antallet af observationer, som ved et threshold på $0,055$ er 199, mens L angiver værdien af den maksimerede likelihoodfunktion. Disse værdier, samt resultaterne af AIC og BIC, er vist i tabel (8).

	POT model	Mærket Hawkes Uforudsigelige mærker	Mærket Hawkes forudsigelige mærker
L	$-350,335$	$-337,008$	$-336,985$
k	3	5	6
AIC	$706,6698$	$684,016$	$685,9694$
BIC	$716,5497$	$700,4825$	$705,7292$

Tabel 8: AIC og BIC mål for POT, og Hawkes POT model med forudsigelige mærker og uforudsigelige mærker.

AIC og BIC målene vurderes ud fra 'smaller-is-better' princippet, hvor det bemærkes at resultaterne for Hawkes POT modellen med uforudsigelige mærker ved begge kriterier har den laveste værdi. POT modellen har den laveste likelihoodværdi, og samtidig også de højeste AIC og BIC mål. Idet de to Hawkes POT modeller har næsten samme værdi af likelihoodfunktionen, og det kun er én parameter der adskiller dem, bemærkes effekten af at BIC målet straffer modelkompleksitet mere end AIC målet gør.

På baggrund af de ovenstående resultater i tabel (8) sammenlignes Hawkes POT med POT modellen med et GPD fit fra den tidligere analyse. I figur (23) illustreres et QQ-plot for Hawkes POT modellen med uforudsigelige mærker op imod de tidligere fundne QQ-plots for LMOM og MLE metoden. Vi har valgt ikke at illustrere begge Hawkes POT modeller, da de er tilnærmelsesvis ens.



Figur 23: QQ-plot for MLE, LMOM og Hawkes POT med uforudsigelige mærker.

Det var forventet at der i QQ-plottet for Hawkes POT modellen kunne anes en forbedring sammenlignet med MLE og LMOM estimationsmetoden for GPD'en, men ud fra de tre QQ-plots i figur (23) ses ikke en forbedring. Dette resultat kunne antyde, at der ikke er nok klyngedannelse i data til at modellering med punktprocesser er nødvendig.

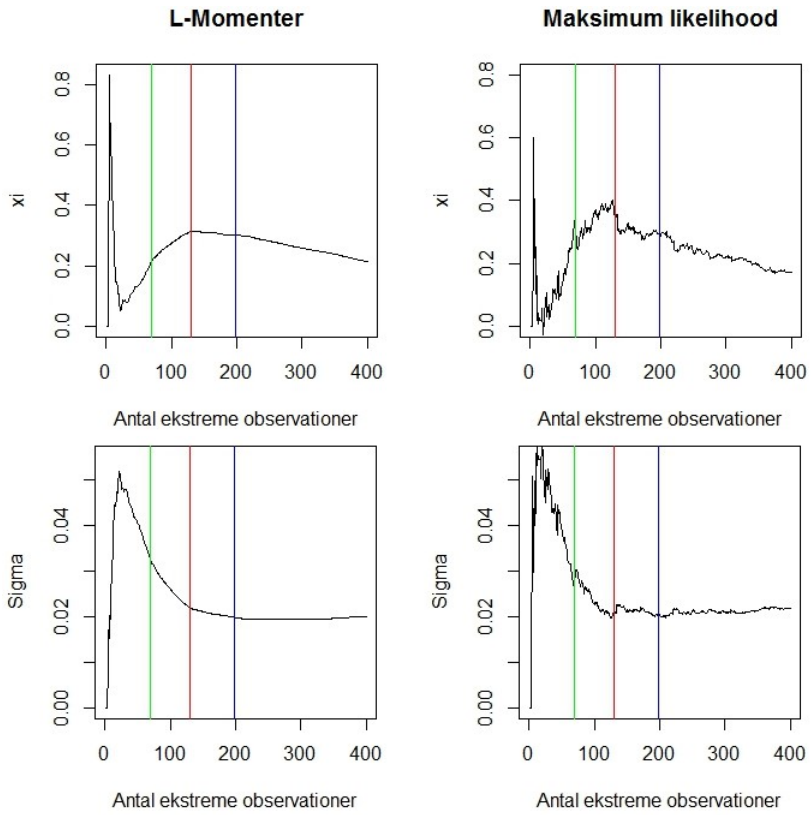
8 Diskussion og perspektivering

Vi har i denne afhandling løbende måtte træffe nogle valg og fravalg, omkring hvilke resultater og metoder som er valide og fordelagtige at benytte på det valgte datasæt: Vestas aktien. Disse valg giver incitament til at diskutere resultaterne og de benyttede modeller.

Til at starte med er valget af datagrundlag vigtigt, idet det vil have indflydelse på antallet af ekstreme observationer og dermed resultaterne for de statistiske metoder. Det valgte datasæt for Vestas aktien består af 2129 observationer, hvilket vi har antaget, er et tilstrækkeligt antal observationer for analysen - men det er selvfølgelig en vurderingssag.

I starten af analysen argumenterede vi for, at POT metoden var at foretrække til udvælgelse af ekstreme observationer, idet metoden sikrer, at alle relevante ekstreme observationer opfanges. Et videre interessant studie kunne være at sammenligne resultaterne i denne afhandling med et GEV fordelings fit, hvor de ekstreme observationer udvælges på baggrund af BM metoden. Fordelen ved at benytte GEV fordelingen frem for GPD'en er, at analysen ville være simplere, i den forstand at studiet omkring punktprocesser ikke nødvendigvis havde været relevant at medtage. Idet de ekstreme observationer i BM metoden udvælges på baggrund af en fast tidsperiode, er klyngedannelse ikke mulig. Udfordringen er, at nogle at de lave observationer kan anses som værende ekstreme, og høje observationer som ligger tæt op af hinanden, kan blive fravalgt afhængigt af blok inddelingen.

I tabel (3) kan det aflæses, at parameterestimaterne for formparameteren ξ, for især LMOM estimationsmetoden, varierer meget i forhold til thresholdværdien. Denne variation er interessant, og vi har derfor udført en sensitivitetsanalyse af estimaterne for ξ og σ for forskellige thresholdværdier, hvor antallet af ekstreme observationer ligger i intervallet $[0:400]$. Udviklingen af parametrene for LMOM og MLE metoden er illustreret i figur (24).



Figur 24: Sensitivitetsanalyse af ξ og σ i forhold til forskellige thresholdværdier. Indtegnet er de tre thresholds $u = 0,055$ med 199 observationer, $u = 0,065$ med 130 observationer og $u = 0,08$ med 70 observationer.

I figur (24) er tendensen i variationen for de to estimationsmetoder tilnærmelsesvis ens, men forskellen består i at kurven for LMOM estimaterne er mere glat, hvor den for estimaterne fundet ved MLE metoden er mere ujævn. Det kan skyldes, at MLE metoden, som beregner estimaterne ved at maksimere likelihoodfunktionen, er mere sensitiv over for små ændringer i antallet af observationer. Det observeres, at der for begge parameterestimater og metoder, er en stor varians for få observationer. Det kan medføre, at man ved en for højt valgt thresholdværdi kan opnå et misvisende parameterestimat. I figur (24) er de tre thresholdværdier, som analysen er baseret på: $u = 0,055$, $u = 0,065$ og $0,08$ indtegnet. Disse thresholdværdier resulterer som tidligere nævnt i henholdsvis

199, 130 og 70 antal ekstreme observationer. Det kan ud fra figur (24) tyde på at 70 observationer, for Vestas datasættet, ikke er nok til at opnå stabile estimater. Derudover er estimatet af ξ for begge modeller støt faldende for 130 ekstreme observationer og derover. Det gælder altså om at finde et threshold, som er højt nok, til at det kun er de ekstreme observationer som medtages, men samtidig ikke er så højt, at variansen på parameterestimaterne er for stor. Ud fra figur (24) bekræftes det dermed at en thresholdværdi på $u = 0,055$ kan være det mest optimale valg.

EPM metodens parameterestimater for ξ og σ ligger på et højt niveau sammenlignet med estimaterne fundet ved LMOM, MLE og PWM metoden. De høje estimater kan skyldes, at de endelige estimater ender med at blive beregnet ved medianen af kun cirka halvdelen af kombinationerne. Dette resultat strider imod artiklen [4, Castillo et al., 1997, 1611], hvor EPM metoden antages at virke for alle mulige værdier af parametre, hvilket skal forstås som, at der ikke er nogen restriktioner for parametrene. Derudover beskrives EPM metoden som den eneste mulighed, når estimater fra andre metoder ikke eksisterer, eller er inkonsistente. Det er korrekt, at det i denne afhandling er muligt at finde et endeligt parameterestimat, men det er på baggrund af de beregningstekniske udfordringer nødvendigt at være kritisk overfor resultatet.

LMOM metoden er beskrevet i [22, Pandey et al., 2001, 3], hvor den fremstilles som værende en estimationsmetode, som er efficient for mange fordelinger, når der skal estimeres parametre i tilfældet med et lille antal observationer. Det ses i figur (24) at LMOM metoden for Vestas datasættet har en lavere varians på parameterestimaterne for få observationer end MLE metoden. Ud fra QQ-plottene i figur (17) kan det også bekræftes at LMOM metoden opnår et bedre modelfit for $u = 0,08$ end MLE metoden, hvor der er en betydeligt større afvigelse fra den rette linje for færre observationer.

Idet der i datasættes struktur kunne anes en tendens til klyngedannelse, startede vi analysen af punktprocesser med at studere en simpel 'Declustrering' metode: 'Runs Declustrering'. Metoden blev fravalgt på baggrund af for få observationer, mere præcist blev datasættet på 199 observationer reduceret til 69 observationer. I punktproces analysen bar resultatet præg af, at der ikke var så

meget klyngedannelse at tage højde for som forventet. I figur (23) hvor QQ-plottet for Hawkes POT punktprocessen sammenlignes med MLE og LMOM metoden ses ikke en stor forbedring. Det kan indikere at en simpel 'Run-Declustering' med en lavere 'run' værdi end tidligere, ville have været en tilstrækkelig metode at benytte.

Denne analyse af metoderne i EVT'en blev udført med henblik på, at beregne de tilhørende VaR og ES risikomål. Det blev tidligere beskrevet, at VaR målet for ikke elliptiske fordelinger ikke nødvendigvis opfylder subadditivitet, hvorfor ES risikomålet er medtaget. At VaR målet i denne afhandling ikke opfylder subadditivitet, giver ikke store konsekvenser da der kun analyseres på en enkelt aktie og ikke en portefølje af aktier. Havde analysen taget udgangspunkt i en portefølje af aktier, ville det have været relevant at studere VaR målet yderligere.

Vi har i denne afhandling valgt at benyttes os af punktprocesser til at løse udfordringen med klyngedannelse i data. En anden tilgang til dette kunne være at benytte en GARCH model, som tager højde for tidsvariationen i den betingede varians, altså 'volatilitets clustering'. Ved brug af denne metode, i stedet for punktprocesser, forventes dog de samme resultater, hvorfor vi ikke har valgt ikke at fokusere på denne tilgang.

9 Konklusion

Formålet med denne afhandling var overordnet, at studere statistiske modeller til beskrivelse af ekstreme værdier. Ekstremværditeorien tager udgangspunkt i egenskaberne bag maksima, hvor Fisher-Tippets sætning siger, at normaliserede maksima konvergerer mod den generaliserede ekstremværdi (GEV) familie, som består af Gumbel, Fréchet og Weibull fordelingerne. Ekstreme observationer som antages at være GEV fordelte, udvælges ved brug af Blok Maksima metoden. Vi har studeret ekstreme tab ud fra daglige aktiekurser på Vestas aktien, hvor der visuelt var en tendens til, at observationerne forekom tæt på hinanden, hvilket gjorde at vi benyttede en anden udvælgelsesmetode: Peaks-over-Threshold (POT) metoden. I POT metoden udvælges de ekstreme observationer, som de observationer der ligger over en given thresholdværdi. Ud fra Pickands-Balkema-de Haans sætning er det den generaliserede paretofordeling (GPD), som er den mest korrekte sandsynlighedsfordeling til modellering af overskridelser over et fastsat threshold. Til bestemmelse af threshold blev Mean Excess og Mean Residual Life plottet benyttet, som begge indikerede tre mulige thresholdværdier: $u = 0,055$, $u = 0,065$ og $u = 0,08$. Da det kan være svært at vælge thresholds grafisk, udførte vi også en Bootstrap Goodness-of-Fit test, som tester hvorvidt overskridelsesobservationer fitter en GPD, hvilken for alle thresholdværdier var signifikant.

Med antagelsen om i.i.d. data blev der for ekstreme observationer, valgt ud fra POT metoden, benyttet flere estimationsmetoder til at fitte de ekstreme observationer til en GPD: Maksimum Likelihood Estimation (MLE), Probability-Weighted-Moments (PWM), Elemental-Percentile-Method (EPM), Method-of-Moments (MOM) og L-Moments-Method (LMOM). Estimation ved hjælp af EPM metoden gav beregningsmæssige udfordringer, hvor vi vurderede, at estimaterne ikke var valide. Estimaterne fra MOM metoden var meget høje, hvormed denne metode heller ikke blev vurderet til at kunne estimere parametrene i GPD'en. MLE og PWM metoden gav tilnærmelsesvis ens estimater, hvorfor vi udover at gå videre med LMOM metoden kun valgte at fortsætte med MLE. Vi erfarede i sensitivitetsanalysen at estimaterne for LMOM metoden var robuste, og metoden var mere

efficient til modellering af små datasæt, end MLE metoden, hvorfor denne metode kunne være at foretrække. De to estimationsmetoder gav dog næsten identiske Value-at-Risk (VaR) og Expected Shortfall (ES) risikomål, hvorfor det ikke giver en stor forskel på de endelige resultater, hvilken af de to estimationsmetoder der benyttes.

I POT metoden kan man også vælge at fitte de ekstreme observationer til en Poisson punktproces, som stadig antager i.i.d. data. For at tage højde for udfordringen omkring klyngedannelse, blev det studeret hvordan man ved at kombinere Poisson punktprocessen og en Self-Exciting proces kunne opnå en Hawkes POT model med forudsigelige og uforudsigelige mærker. Vi sammenlignede de tre modeller og erfarede, at Poisson punktprocessen som forventet fik tilnærmelsesvis samme estimater som i POT modellen med GPD fit. De to Hawkes POT modeller fik også næsten identiske estimater, da forskellen ligger i parameteren α, som i modellen med forudsigelige mærker var meget lille: $0,0002$. Mærkerne har derfor ikke en stor indflydelse på Hawkes POT modellen. Idet estimaterne i Hawkes POT modellerne var tilnærmelsesvis identiske, testede vi modelkompleksiteten ved hjælp af Akaike- og Bayesian-informationskriterierne, hvorudfra Hawkes POT modellen med uforudsigelige mærker, som havde de laveste værdier, var at foretrække.

I Poisson punktprocessen var VaR og ES målene henholdsvis $0,0690$ og $0,1193$ og igen tilnærmelsesvis de samme som resultaterne fra POT modellen med GPD fit. I Hawkes POT modellerne var risikomålene ens, på henholdsvis $0,0505$ og $0,0897$, som er et betydeligt lavere niveau.

Til sidst i afhandlingen sammenlignede vi, ved hjælp af QQ-plots, Hawkes POT modellen med uforudsigelige mærker med GPD fit ud fra MLE og LMOM estimationsmetoderne. Det kunne herudfra konkluderes, at vi, med antagelsen om klyngedannelse, ikke kunne opnå et bemærkelsesværdigt bedre fit af de ekstreme observationer. I denne afhandling hvor Vestas datasættet benyttes gør det altså ikke nogen forskel på modelfittet, om de ekstreme observationer statistisk modelleres ud fra en GPD eller en punktproces tilgang i POT modellen.

Litteratur

[1] Pennstate eberly college of science. https://onlinecourses.science.psu.edu/stat414/node/193.

[2] Pennstate eberly college of science. https://onlinecourses.science.psu.edu/stat504/node/27.

[3] Alva, J. A. V. and E. González-Estrada (2009). A bootstrap goodness of fit test for the generalized pareto distribution. *Elsevier*.

[4] Castillo, E. and A. S. Hadi (1997). Fitting the generalized pareto distribution to data. *Journal of the American Statistical Association*.

[5] Chavez-Demoulin, V., A. C. Davison, and A. J. McNeil (2007). Estimating value-at-risk: a point process approach. *Routledge, Taylor and Francis Group*.

[6] Christophersen, R. S. (2011, Maj). Hawkes punktprocessen: En model for jordskælv i danmark.

[7] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistic.

[8] Cruz, M. G., G. W. Peters, and P. V.Shevchenko (2015). *Fundemental Aspects of Operational Risk and Insurance Analytics:A Handbook og Operational Risk*. John Wiley & Sons.

[9] Dan Beltoft, K. T. (2009). Kontinuitet og integraler matematisk analyse 1.

[10] de Silva, N. An introduction to r: Examples for actuaries. http://toolkit.pbworks.com/f/R

[11] de Zea Bermudez, P. and S. Kotz (2009). Parameter estimation of the generalized pareto distribution - part i. *Journal of Statistical planning and Inference*.

[12] Embrechts, P., C. Klüppelberg, and T. Mikosch (2012). *Modelling Extremal Events for Insurance and Finance*. Springer.

[13] Ferreira, A. and L. D. Haan (2014). On the block maksima method in extreme value theory:pwm estimators. *The Annals of statistics*.

[14] Gilli, M. and E. Kellezi (2006). An application of extreme value theory for measuring financial risk. *Computational Economics(2006) 27: 207-228*.

[15] Haals, N. P. and P. Jensen (2009). Rumlige punkt-og linjeprocesser, en introduktion med henblik på beskrivelse af punktmønstre med linjetendenser og en modelopstimodel for datasæt af gravhøje.

[16] Hosking, J. and J. Wallis (1997). *Regional Frequency Analasis, An Approach Based on L-Moments*. Cambridge University Press.

[17] Hosking, J. R. M. (1990). L-moments: Analysis and estimation og ddistribution using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 52, No. 1*.

[18] Jockovic, J. (2012). Quantile estimation for the generalized pareto distribution with application to finance. *Yugoslav Journal of Operations Reserch 22(2012), Number 2, 297-311*.

[19] Markose, S. and A. Alentorn (2005). Option pricing and the implied tail index with the generalized extreme value (gev) distribution. *Centre of Computational Finance and Economics Agents (CCFEA)*.

[20] McNeil, A. J., R. Frey, and P. Embrechts (2015). *Quantitative Risk Management, concepts, techniques and tools*. Princeton university Press.

[21] Nygaard, C. (2011). *Samfundsvidenskabelige analysemetoder*. Forlaget Samfundslitteratur.

[22] Pandey, M. and P. van Gelder & J.K. Vrijling (2001). The estimation of extreme quantiles of wind velocity using l-moments in the ppeak-over-threshold approach. *Elsevier*.

[23] Rau-Bredow, H. (2004). *Risk Measures for the 21st Century*. wiley.

[24] Ruppert, D. (2011). *Statistics and Data Analysis for Financial Engineering*. Springer Texts in Statistics.

[25] Whalen, T. M., G. T. Savage, and G. D. Jeong (2003). An evaluation of the self-demined probability-weighted-moment method for estimating extreme wind speeds. *Journal of wind engineering*.

Liste over forkortelser

AIC	–	Akaike Information Criterion
BIC	–	Bayesian Information Criterion
BIS	–	Bank for International Settlements
BM	–	Blok Maksima
CLT	–	Den centrale grænseværdisætning
EPM	–	Elemental Percentile Method
ES	–	Expected Shortfall
ETAS	–	Epidemic Type After Shock
EVT	–	Ekstremværdi teori
GARCH	–	Generalized Autoreggressive Conditional Heteroskedasticity
GEV	–	Den generaliserede ekstremværdifordeling
GPD	–	Den generaliserede paretofordeling
i.i.d.	–	Independent and Identically Distributed
KI	–	Konfidensinterval
LMOM	–	L – Moments Method
MDA	–	Maximum Domain of Attraction
ME	–	Mean Excess
MLE	–	Maksimum Likelihood Estimation
MOM	–	Method of Moments
MRL	–	Mean Residual Life
MSE	–	Mean Squared Error
POT	–	Peaks over Threshold
PWM	–	Probability Weighted Moments
SE	–	Standard Error
VaR	–	Value at Risk