CHALMERS | GÖTEBORGS UNIVERSITET

Om skattningar av sannolikheter för extrema händelser

Examensarbete för kandidatexamen i matematik vid Göteborgs universitet

Amanda Hårsmar
Jari Martikainen
Martin Rensfeldt

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2013

Om skattningar av sannolikheter för extrema händelser

Examensarbete för kandidatexamen i matematisk statistik vid Göteborgs universitet

Amanda Hårsmar Martin Rensfeldt

Examensarbete för kandidatexamen i matematisk statistik inom matematikprogrammet vid Göteborgs universitet

Jari Martikainen

Handledare: Olle Nerman
Examinator: Hjalmar Rosengren

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2013

Sammanfattning

Att skatta sannolikheter för extrema händelser är svårt eftersom de sällan inträffar och underlaget att bygga skattningar utifrån är begränsat. Ofta används skattningsmetoder grundade på asymptotiska resultat från extremvärdesteorin som inte självklart är uppfyllda när man utgår från en verklig datamängd. I den här uppsatsen har en sådan metod, som kallas peak over threshold-metoden, jämförts med en modifierad metod som inte bygger på sådan asymptotik; fastypsmetoden. I peak over threshold-metoden skattas den betingade sannolikheten att befinna sig långt ut i svansen av en fördelning med hjälp av en generaliserad paretofördelning. I fastypsmetoden skattas istället den betingade sannolikheten att befinna sig långt ut i svansen med en fastypsfördelning. Resultaten från simuleringar i denna första pilotstudie visar inte på några tydliga skillnader mellan metodernas precision. Peak over threshold-metoden visar sig dock ibland ge svansapproximationer med ändligt fördelningsstöd, vilket är problematiskt eftersom stickprovet sällan kan antas ha en övre deterministisk begränsning.

Abstract

Extreme events seldom occur and basic data for estimation is often limited. It is thus difficult to estimate probabilities of extreme events. Estimation methods based on asymptotic results from extreme value theory are widely used even though these results are not always well motivated when dealing with real data sets. In this report such a method, called the peak over threshold-method, has been compared with a modified method, called the phase type-method, that is not based on such asymptotic results. According to the peak over threshold-method the conditional probability of obtaining a value in the tail of a distribution is approximated by means of a generalised pareto distribution. According to the phase type-method this probability is instead approximated with a phase type distribution. The results from the simulations in this initial pilot study do not show any evident differences between the precision of the methods. However, the peak over threshold-method sometimes results in tail approximations with finite distribution support. This is problematic since the sample cannot generally be expected to have an upper deterministic limit.

Innehåll

1 Inledning **5**

2 Teori **6**
2.1 Felintensiteter . . . 6
2.2 Fastypsfördelningar och bakomliggande markovteori . . . 6
2.2.1 Markovitet och tidshomogenitet . . . 7
2.2.2 Absorption i en markovprocess . . . 7
2.2.3 Generatorn till markovprocessen och fördelningen för hopptider . . . 7
2.2.4 Fastypsfördelningens definition . . . 8
2.2.5 Övergångssannolikheter i en markovprocess . . . 9
2.2.6 Fastypsfördelningens fördelningsfunktion . . . 10
2.2.7 Exempel på fastypsfördelningar . . . 11
2.2.8 Fastypsfördelningens felintensitet . . . 12
2.2.9 Tät klass . . . 13
2.3 Extremvärdesteori . . . 13
2.3.1 Klassisk extremvärdesteori . . . 13
2.3.2 Peak over threshold . . . 14
2.3.3 Betingningsstabilitet . . . 15
2.3.4 Konvergens mot betingningsstabilitet . . . 16
2.3.5 Sammankoppling av metoderna . . . 18
2.4 Argument för att fördelningarna i vår undersökning konvergerar . . . 20
2.4.1 Paretofördelningen . . . 20
2.4.2 Exponentialfördelningen . . . 20
2.4.3 Fastypsfördelningen . . . 20
2.4.4 Normalfördelningen (0,1) . . . 21
2.4.5 Lognormalfördelningen (0,1) . . . 21
2.4.6 Skattning av parametrar till stickprov . . . 22

3 Undersökning med simulering **23**
3.1 Genomgång av undersökningens metoder . . . 23
3.1.1 POT-metoden . . . 23
3.1.2 Fastypsmetoden . . . 24
3.2 Detaljer kring undersökningen . . . 24
3.3 Resultat . . . 26
3.3.1 Normalfördelning (0,1) . . . 26
3.3.2 Lognormalfördelning (0,1) . . . 27
3.3.3 Paretofördelning . . . 27
3.3.4 Kanonisk coxiansk fördelning av ordning 10 . . . 30
3.3.5 T-fördelning med 6 frihetsgrader . . . 30
3.3.6 Exponentialfördelning . . . 31

4 Illustration av metoderna med nederbördsdata **31**
4.1 Beskrivning av datamängden . . . 31
4.2 Förbehandling av data . . . 31
4.3 Tillvägagångssätt . . . 31
4.4 Resultat . . . 32
4.4.1 Resultat i en av de 25 skattningarna . . . 32

5 Slutsatser för undersökningen med simuleringar **33**
5.1 Ändligt stöd i generaliserade paretofördelningen . . . 34

6 Diskussion **36**

A Appendix **38**
A.1 Entydigheten hos den kvasistationära fördelningen $\boldsymbol{\nu} = (0, 0, ..., 1)$ för en kanonisk coxiansk fördelning 38
A.2 Mer detaljerade beräkningar för hur ett uttryck fås fram för felintensiteten $h(t)$: 38
A.3 Härledning av skattningen $\hat{q}_{0.01}$ 39
A.4 Programkod 39

Förord

Planeringen av arbetet har gjorts gemensamt. En viktig del av materialet; Enger och Grandells lärobok i markovteori, Maritas Olssons avhandling om fastypsfördelningar och några artiklar om fastypsfördelningar lästes parallellt av alla i gruppen. Även informationssökandet skedde parallellt av alla i gruppen. Materialet vi hittade delade vi upp mellan oss och bytte sedan fram och tillbaka för att alla skulle få del av viktiga resultat.

Martin och Jari har haft större ansvar vad gäller programmerings- och simuleringsdelarna. Dock har tillvägagångssätt och problem som uppstått diskuterats i hela gruppen.

Att ange en huvudansvarig författare för respektive avsnitt blir missvisande eftersom hela texten har skapats i kontinuerlig dialog mellan gruppens medlemmar. Flera personer har tillfört, lagt till och ändrat i så gott som varje avsnitt. För mycket av texten har skapats i dialog för att se det som ren korrekturläsning av varandras texter.

En individuell tidslogg samt en gemensam dagbok över de medverkandes prestationer har förts under arbetet.

1 Inledning

Det som definierar extrema händelser är att de är osannolika. När man vill bedöma sannolikheten för osannolika händelser är informationen mycket begränsad. Det gör att vanliga statistiska metoder blir otillräckliga och andra tillvägagångssätt krävs. Då är det vanligt att vända sig till extremvärdesteorin. Där ges en grund för att fördelningen för extrema värden i stora stickprov efter lämplig skalning konvergerar mot någon fördelning i en särskild klass av extremvärdesfördelningar. En sådan extremvärdesfördelning kan användas för att göra uttalanden om osannolika händelser.

Teorin kräver dock att stickprovsstorleken är stor nog för att konvergensen ska ha ägt rum. Huruvida så är fallet går i verkliga statistiska situationer inte att avgöra. Det finns situationer där fördelningen inte konvergerar och det finns situationer där fördelningen visserligen konvergerar men där konvergensen är extremt långsam. Eftersom det handlar om så osannolika händelser finns helt enkelt inte ett tillräckligt underlag för att granska konvergensen. Det är svårt att uttala sig om händelser som så gott som aldrig händer.

Vi har utgått från en variant av extremvärdesteorin där man betingar med att värden är extrema. Utifrån den betingade sannolikhetsfördelningen får man konvergens mot en viss klass fördelningar som kallas generaliserade paretofördelningar.

I fall med långsam konvergens kan fördelningen förväntas ha konvergerat för händelser som är extremt osannolika. Däremot blir antagandet mer tveksamt för händelser som är ganska osannolika. (Exakt vad som är en tillräckligt extrem händelse beror på faktorer som hastigheten hos konvergensen och annat som delvis faller utanför denna uppsats. Vi nöjer oss med att konstatera att det finns händelser som är för osannolika för vanliga statistiska metoder men inte tillräckligt osannolika för att uppfylla kravet på konvergens.) I avsaknad av annat används ofta extremvärdesteorin trots brister eftersom det är bättre att basera bedömningar på en teori som ger en osäker grund än ingen grund alls.

Fastypsfördelningar, tid till absorption i ändliga, tidskontinuerliga markovprocesser med ett absorberande tillstånd, har flera tillämpningar och kan bland annat användas för att approximera andra fördelningar. Att approximera fördelningar är av intresse i sammanhang där man vill göra uttalanden som sträcker sig utanför den information som finns att tillgå, så kallad extrapolering. Speciellt skulle fastypsfördelningar kunna utgöra ett alternativ till extremvärdesfördelningar när det anses troligt att fördelningen ännu inte har konvergerat. Det vill säga i fall där händelserna är just ganska osannolika. Syftet med denna uppsats är att undersöka om fastypsfördelningar är ett bra alternativ till extremvärdesfördelningar i sådana situationer.

Efter en teoretisk genomgång har vi jämfört extremvärdesmetoder baserade på den så kallade peak over threshold-metoden, även kallad POT-metoden, med en modifierad POT-metod som använder fastypsfördelningar istället för generaliserade paretofördelningar. En sådan jämförelse har genomförts för sex teoretiska fördelningar. Vi har använt värden ovanför tre olika trösklar i fördelningarnas svans; 95-, 97,5- och 99-percentilen. För var och en av dessa trösklar har vi med respektive metod skattat sannolikheten att befinna sig ovanför 99,9-percentilen i den teoretiska fördelningen. Stickprovsstorlekarna har varit minst 1000. I den teoretiska fördelningen är naturligtvis sannolikheten en promille att ett värde finnns bland de en promille mest extrema. Vi har därför jämfört vilken av de två metoderna som ger resultat närmast en promille. Till sist har vi använt metoderna på nedebördsdata från SMHI, som en illustration av hur de används i verkligheten.

I uppsatsen används både begreppet kvantil och percentil. Med en α-kvantil menas att sannolikheten att hamna ovan denna punkt är α. För en percentil anges istället sannolikheten att hamna nedanför denna punkt.

2 Teori

I detta kapitel ges en genomgång av teorin som ligger till grund för vår undersökning. Kapitlet börjar med en genomgång av begreppet felintensitet som sedan tas upp både i avsnittet om fastypsfördelningar och som utgångspunkt i teorin kring peak over threshold. Fastypsfördelningar introduceras med hjälp av markovteori. I avsnittet om extremvärdesteori beskrivs först två olika metoder och sedan sambanden mellan dem. Avslutningsvis ges argument för att de fördelningar som används i vår undersökning uppfyller villkor på konvergens som gör dem relevanta för undersökningen.

2.1 Felintensiteter

För förståelse av extremvärdesteorin spelar felintensiteter en viktig roll.

Definition 2.1. *Felintensiteten för en fördelningsfunktion med kontinuerlig täthet definieras som*

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t).$$

(Aalen et al. (2008))

Om t är tiden till ett fel inträffar så kan man tolka felintenstieten i punkten t som det förväntade antalet "händelser" (fel) per tidsenhet, givet att "händelsen" (felet) inte har inträffat fram till tidpunkten t.

För felintensiteten gäller $h(t) = \frac{f(t)}{1-F(t)}$, där $f(t)$ är täthetsfunktionen och $F(t)$ är fördelningsfunktionen. (Aalen et al. (2008))

Man brukar inte tänka på felintensiteteten för annat än positiva stokastiska variabler men den kan definieras även för negativa t.

Det finns en koppling mellan felintensiteten och fördelningsfunktionen.

Sats 2.2. *Låt X vara en ickenegativ stokastisk variabel med felintensitet h och fördelningsfunktion F. Då bestäms h och F väsentligen entydigt av varandra och det gäller att*

$$F(t) = 1 - e^{-\int_0^t h(y)dy}.$$

(Aalen et al. (2008))

I avsnittet om peak over threshold, som är en extremvärdesmetod, används felintensiteten som utgångspunkt.

Vi kommer att arbeta med en del fördelningar som är definierade på hela reella talaxeln. Då gäller att

$$F(t) = 1 - e^{-\int_{-\infty}^t h(y)dy}.$$

2.2 Fastypsfördelningar och bakomliggande markovteori

Fastypsfördelningens definition bygger på teori om markovprocesser. Vi inleder därför med en genomgång av grundläggande teori för markovprocesser i kontinuerlig tid. I genomgången av markovteori utgår vi från kompendiet *Markovprocesser och köteori* av Enger och Grandell.

2.2.1 Markovitet och tidshomogenitet

En markovprocess i kontinuerlig tid med diskret utfallsrum är en stokastisk process $\{X(t), t \geq 0\}$ på ett diskret utfallsrum $E = \{0, 1, 2, ...\}$ som uppfyller markovegenskapen, dvs. att

Definition 2.3. *En stokastisk process $\{X(t), t \geq 0\}$ är markovsk om och endast om*

$$P(X(t_{n+1}) = i_{n+1} | X(t_n) = i_n, X(t_{n-1}) = i_{n-1}, ..., X(t_0) = i_0) = \\ = P(X(t_{n+1}) = i_{n+1} | X(t_n) = i_n),$$

när

$$i_0, i_1, ..., i_{n-1}, i_n, i_{n+1} \in E, \\ 0 < t_0 < t_1 < ... < t_{n-1} < t_n < t_{n+1}.$$

Markovegenskapen säger alltså att sannolikheten för sista övergången till tillstånd i_{n+1} endast beror på det tillstånd som processen befinner sig i innan hoppet (och på tidpunkterna t_n, t_{n+1}) .

De markovprocesser som ligger till grund för fastypsfördelningar är tidshomogena. Därför begränsar vi framställningen av markovteorin till att gälla tidshomogena markovprocesser.

Definition 2.4. *En markovprocess $\{X(t), t \geq 0\}$ är tidshomogen om och endast om*

$$P(X(t + h) = i_n | X(t) = i_{n-1}) = P(X(h) = i_n | X(0) = i_{n-1}).$$

Det innebär att sannolikheten att ta sig till tillstånd i_n på tiden h givet att man befinner sig i tillstånd i_{n-1}, är densamma oavsett när i processen man befinner sig i tillstånd i_{n-1}.

2.2.2 Absorption i en markovprocess

Ett tillstånd i sägs leda till tillstånd j om det i ett ändligt antal steg är möjligt att komma från i till j. Ett tillstånd är absorberande om kedjan stannar i tillståndet med sannolikhet 1, givet att den kommit dit.

Definition 2.5. *Ett tillstånd som direkt eller indirekt leder till ett absorberande tillstånd kallas genomgångstillstånd.*

Vi kommer härifrån bara att betrakta markovprocesser med ändliga tillståndsrum $E = \{0, 1, ..., n\}$.

2.2.3 Generatorn till markovprocessen och fördelningen för hopptider

Markovprocessen kan beskrivas med hjälp av en generator. Tiden till hopp från det tillstånd processen befinner sig i till de andra tillstånden är exponentialfördelad. Exponentialfördelningens intensitet kan vara olika för de olika tillstånden. I ett givet tillstånd i kan man tänka på processen som att det finns händelser $A_j, j \in E \backslash \{i\}$, som kan inträffa, där A_j = hopp till tillstånd j. Låt V_j vara tiden till respektive A_j inträffar. Då gäller alltså att $V_j \sim Exp(\lambda_j)$, där λ_j är någon intensitet som ges av tillståndet i som processen befinner sig i innan hoppet. Man kan tänka på dessa V_j som oberoende stokastiska variabler. Det är bara den händelse som sker först som verkligen inträffat.

Definition 2.6. *Generatorn T till en markovprocess är matrisen av övergångsintensiteter*

$$\Lambda = \begin{pmatrix} \lambda_{0,0} & \lambda_{0,1} & \cdots & \lambda_{0,n} \\ \lambda_{1,0} & \lambda_{1,1} & \cdots & \lambda_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{n,0} & \lambda_{n,1} & \cdots & \lambda_{n,n} \end{pmatrix},$$

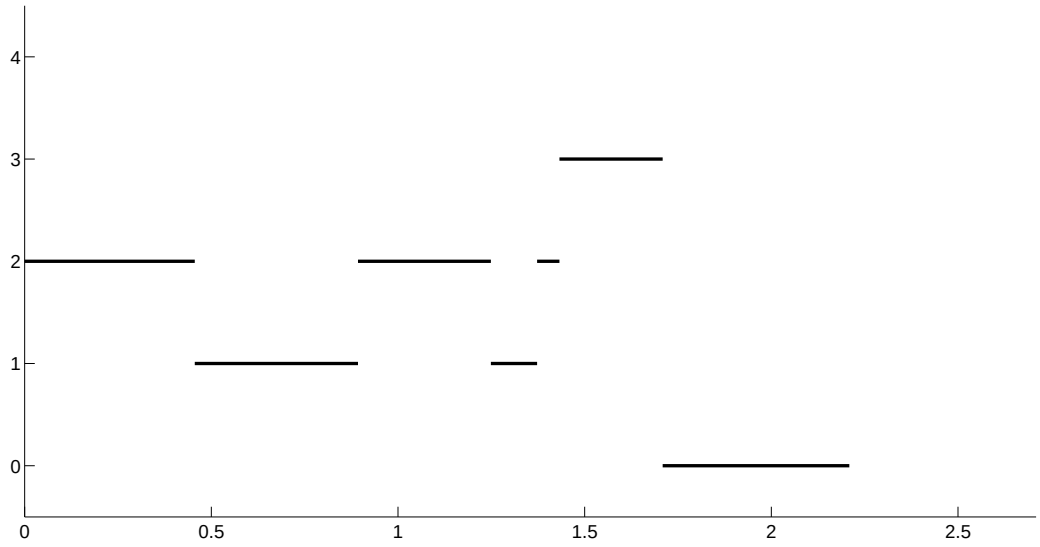
där

$$\lambda_{i,i} = -\sum_{j \neq i}^{n} \lambda_{i,j}.$$

Elementet $\lambda_{i,j}$ i matrisen Λ är alltså, givet att markovprocessen befinner sig i tillstånd i, intensiteten till variabeln V_j som anger tiden tills händelsen A_j inträffar.

Den sista summationen innebär att varje rad i generatorn summeras till noll. Det har sin förklaring i att varje rad i matrisen för övergångssannolikheter, som definieras i ett senare avsnitt, summeras till ett.

Figur 1 visar ett exempel på hur en realisering av en markovprocess kan se ut. Längden på de horisontella linjerna anger hur länge processen stannade i respektive tillstånd.



Figur 1: Exempel på en realisering av en markovprocess.

2.2.4 Fastypsfördelningens definition

Nu när vi har definierat en tidshomogen och tidskontinuerlig markovprocess, absorption, genomgångstillstånd och generatorn kan vi definiera fastypsfördelningen.

Låt τ vara tiden till absorption i en homogen tidskontinuerlig markovprocess $\{X(t), t \geq 0\}$ med ett ändligt tillståndsrum $E = \{0, 1, 2, ..., p\}$, där 0 är det absorberande tillståndet och p antalet genomgångstillstånd. Då är τ fastypsfördelad med parametrar som utgörs av en

startvektor $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_p)$ där $\pi_i = P(X(0) = i)$, och generatorn $\mathbf{T}$ för genomgångstillstånden 1, 2, ... , p. Detta betecknas $\tau \sim PH(\boldsymbol{\pi}, \mathbf{T})$. Fastypsfördelningen sägs vara av ordning p.

Startvektorn anger startsannolikheterna endast för genomgångstillstånden, detta för att processen inte tillåts börja i det absorberande tillståndet. Parametriseringen, som alltså innehåller $p^2 + p - 1$ parametrar, är inte unik. Flera parametriseringar kan ge upphov till samma fastypsfördelning. Det finns en unik parametrisering med $2p - 1$ parametrar som ges av de $2p - 1$ första momenten (Asmussen et al. (1996)). Den tas inte upp i den här uppsatsen eftersom EMpht-programmet (Olsson (1996)), som används i vår undersökning för att skatta parametrar i fastypsfördelningen, utgår från parametriseringen $(\boldsymbol{\pi}, \mathbf{T})$.

Med ovanstående parametrar $(\boldsymbol{\pi}, \mathbf{T})$ ges generatorn $\boldsymbol{\Lambda}$ till hela markovprocessen $\{X(t), t \geq 0\}$ på följande vis:

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & \mathbf{0} \\ \boldsymbol{\theta} & \mathbf{T} \end{pmatrix},$$

Där $\boldsymbol{\theta}$ är kolonnvektorn med tillståndsberoende absorptionsintensiteter, $\mathbf{0}$ en p-dimensionell radvektor av nollor, och $\mathbf{T}$ generatorn för genomgångstillstånden. Viktigt för att förstå varför hela $\boldsymbol{\Lambda}$ ges av $(\boldsymbol{\pi}, \mathbf{T})$ är insikten att $\boldsymbol{\theta} = -\mathbf{T}\mathbf{e}$, där $\mathbf{e}$ är den p-dimensionella kolonnvektorn av ettor (Olsson (1995)).

2.2.5 Övergångssannolikheter i en markovprocess

Generatorn är ett sätt att beskriva markovprocessen. En markovprocess med ändligt tillståndsrum kan också beskrivas med hjälp av matrisen av övergångssannolikheter.

Definition 2.7. *Matrisen för övergångssannolikheter $\mathbf{P}_t$ till en markovprocess är matrisen*

$$\mathbf{P}_t = \begin{pmatrix} p_{0,0}(t) & p_{0,1}(t) & \cdots & p_{0,n}(t) \\ p_{1,0}(t) & p_{1,1}(t) & \cdots & p_{1,n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ p_{n,0}(t) & p_{n,1}(t) & \cdots & p_{n,n}(t) \end{pmatrix},$$

där

$$p_{i,j}(t) = P(X(t+s) = j | X(s) = i)$$

och

$$\sum_{j=1}^{n} p_{i,j}(t) = 1.$$

Ett viktigt resultat för övergångssannolikheterna ges av Kolmogorov-Chapmans sats. Satsen är egentligen mer omfattande, och endast det för denna text viktigaste resultatet redogörs för.

Sats 2.8. *Kolmogorov-Chapmans sats*

Låt X vara en tidshomogen, tidskontinuerlig markovprocess med $\boldsymbol{p}(h) = (p_0(h), p_1(h), ..., p_p(h))$, där $p_i(h) = P(X(h) = i)$, och $\boldsymbol{P}_h$ matrisen av övergångssannolikheter $h \geq 0$. Då gäller att

$$\boldsymbol{p}(s+t) = \boldsymbol{p}(s)\boldsymbol{P}_t, \text{ med } s, t \geq 0. \tag{1}$$

Bevis. Låt $p_j(s+t)$ vara ett godtyckligt element i $\mathbf{p}(s+t)$. Då skall det visas att

$$p_j(s+t) = (\mathbf{p}(s)\mathbf{P}_t)_j \ .$$

Men det gäller att

$$(p(s)P_t)_j = (p_0(h), p_1(h), ..., p_p(h)) \begin{pmatrix} p_{0j}(t) \\ p_{1j}(t) \\ \vdots \\ p_{pj}(t) \end{pmatrix} = \sum_{i \in E} p_i(s) p_{ij}(t).$$

Dessutom gäller att

$$\begin{aligned} p_j(s+t) &= [\textit{lagen om total sannolikhet}] = \sum_{i \in E} P(X(s+t) = j | X(s) = i) P(X(s) = i) \\ &= \sum_{i \in E} p_i(s) p_{ij}(t). \end{aligned}$$

□

Genom att derivera $\mathbf{P}_t$ kan man få fram system av differentialekvationer som kan visas ha den entydiga lösningen $\mathbf{P}_t = exp(\mathbf{\Lambda} t) = \sum_{n=0}^{\infty} \frac{\mathbf{\Lambda}^n t^n}{n!}$.

2.2.6 Fastypsfördelningens fördelningsfunktion

För att härleda fastypsfördelningens fördelningsfunktion är matrisen för övergångssannolikheter en bättre utgångspunkt än generatorn $\mathbf{\Lambda} = \begin{pmatrix} 1 & \mathbf{0} \\ \boldsymbol{\theta} & \mathbf{T} \end{pmatrix}$.

Det visar sig att matrisen av övergångssannolikgeter $\mathbf{P}_t = exp(\mathbf{\Lambda} t)$ kan partitioneras:

$$\begin{aligned} \mathbf{P}_t &= exp(\mathbf{\Lambda} t) \\ &= \sum_{n=0}^{\infty} \frac{\mathbf{\Lambda}^n t^n}{n!} \\ &= \mathbf{I} + \sum_{n=1}^{\infty} \frac{\mathbf{\Lambda}^n t^n}{n!} \\ &= \mathbf{I} + \sum_{n=1}^{\infty} \frac{t^n}{n!} \begin{pmatrix} 0 & \mathbf{0} \\ \boldsymbol{\theta} & \mathbf{T} \end{pmatrix}^n \\ &= \mathbf{I} + \sum_{n=1}^{\infty} \frac{t^n}{n!} \begin{pmatrix} 0 & \mathbf{0} \\ \boldsymbol{T^{n-1}\theta} & \mathbf{T^n} \end{pmatrix} \\ &= \mathbf{I} + \sum_{n=1}^{\infty} \frac{t^n}{n!} \begin{pmatrix} 0 & \mathbf{0} \\ \boldsymbol{-T^n \mathbf{e}} & \mathbf{T^n} \end{pmatrix} \\ &= \mathbf{I} + \begin{pmatrix} 0 & \mathbf{0} \\ -\sum_{n=1}^{\infty} \frac{\mathbf{T}^n \mathbf{e} t^n}{n!} & \sum_{n=1}^{\infty} \frac{\mathbf{T}^n t^n}{n!} \end{pmatrix} \\ &= \begin{pmatrix} 1 & \mathbf{0} \\ -\sum_{n=1}^{\infty} \frac{\mathbf{T}^n \mathbf{e} t^n}{n!} & \mathbf{I} + \sum_{n=1}^{\infty} \frac{\mathbf{T}^n t^n}{n!} \end{pmatrix} \\ &= \begin{pmatrix} 1 & \mathbf{0} \\ -(exp(\mathbf{T}t)\mathbf{e} - \mathbf{I}\mathbf{e}) & exp(\mathbf{T}t) \end{pmatrix} \\ &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{e} - exp(\mathbf{T}t)\mathbf{e} & exp(\mathbf{T}t) \end{pmatrix} \end{aligned}$$

där $\mathbf{e}$ är p-dimensionell kolonnvektorn av ettor.

Det går nu rättframt att härleda fördelningsfunktionen till τ. En vanlig ansättning ger:

$$F_\tau(t) = P(\tau \leq t) = P(X(t) = 0) = [\mathbf{p}(t)]_1 = [\mathbf{p}(0+t)]_1 = [kolmogorov - chapman] = \\ = [\mathbf{p}(0)\mathbf{P}_t]_1 = [(0, \boldsymbol{\pi})\mathbf{P}_t]_1$$

Där $[\ .\ ]_1$ betecknar första elementet i vektor-matrisprodukten. Vi skriver ut sista vektor-matrisprodukten genom att använda vår alternativa formel för P_v. Vi får att

$$[(0, \boldsymbol{\pi})\mathbf{P}_t]_1 = \left[(0, \boldsymbol{\pi}) \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{e} - exp(\mathbf{T}t)\mathbf{e} & exp(\mathbf{T}t) \end{pmatrix}\right]_1 = \boldsymbol{\pi}(\mathbf{e} - exp(\mathbf{T}t)\mathbf{e}) = 1 - \boldsymbol{\pi} exp(\mathbf{T}t)\mathbf{e}.$$

Vilket alltså blir fördelningsfunktionen. Näst sista likheten fås av att vi endast är intresserade av första elementet i den av vektor-matrismultiplikationen resulterande vektorn. Den sista likheten fås av att elementen i $\boldsymbol{\pi}$ summerar sig till ett. För att få tätheten till τ deriverar vi helt enkelt fördelningsfunktionen m.a.p. s och erhåller på så vis (Bladt (2005))

$$f_\tau(t) = \boldsymbol{\pi} exp(\mathbf{T}t)\boldsymbol{\theta}.$$

2.2.7 Exempel på fastypsfördelningar

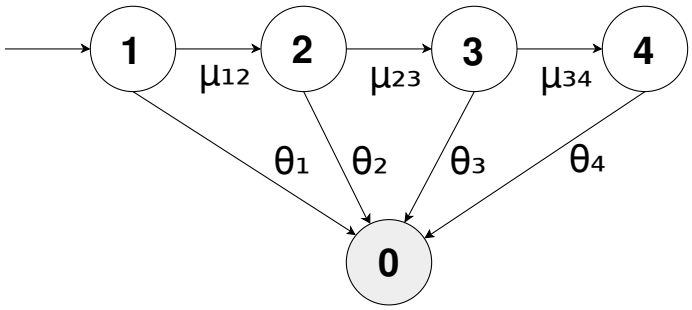
Här följer några exempel på fastypsfördelningar:

Coxiansk fastypsfördelning av ordning p. Har representationen:

$$\boldsymbol{\pi} = (1, 0, ..., 0),$$

$$\boldsymbol{T} = \begin{pmatrix} -\mu_{1,1} & \mu_{1,2} & 0 & \cdots & 0 \\ 0 & -\mu_{2,2} & \mu_{2,3} & \cdots & 0 \\ 0 & 0 & -\mu_{3,3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & -\mu_{p,p} \end{pmatrix}.$$

Det går att ordna tillstånden så att intensiteterna satisfierar $\mu_{1,1} \geq \mu_{2,2} \geq ... \geq \mu_{p,p} > 0$ utan att fördelningen ändras (Cumani (1982), O'Cinneide (1989)). Detta kallas kanonisk coxiansk form.



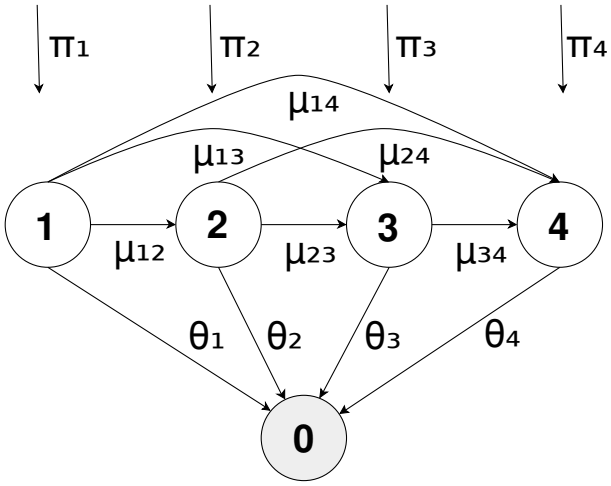
Figur 2: Coxiansk fastypsfördelning av ordning 4 där $\mu_{i,i+1}$ är övergångsintensiterna och θ_i absorptionsintensiteterna $i = 1, .., 4$.

Acyklisk fastypsfördelning, har efter lämplig omsortering av tillstånden en övertriangulär representation:

$$\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_p)$$

$$\boldsymbol{T} = \begin{pmatrix} -\mu_{1,1} & \mu_{1,2} & \mu_{1,3} & \cdots & \mu_{1,p} \\ 0 & -\mu_{2,2} & \mu_{2,3} & \cdots & \mu_{2,p} \\ 0 & 0 & -\mu_{3,3} & \cdots & \mu_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & -\mu_{p,p} \end{pmatrix}$$

Varje övertriangulär fastypsfördelning kan representeras i coxiansk form och därmed även i kanonisk coxiansk form (Cumani (1982), O'Cinneide (1989)). Det är den här egenskapen som gör att vi väljer att använda en coxiansk fastypsfördelning för skattning av svanssannolikheterna i vår undersökning. Att den coxianska fastypsfördelningen har färre parametrar än en övertriangulär gör nämligen att körningen av programvaran går snabbare. I figur 2 och 3 ses övergångsdiagram för en coxiansk respektive en kanonisk coxiansk fördelning.



Figur 3: Acyklisk fastypsfördelning av ordning 4.

2.2.8 Fastypsfördelningens felintensitet

Fastypsfördelningen har felintensitet

$$h(t) = \frac{\boldsymbol{\pi} exp(\mathbf{T}t)\boldsymbol{\theta}}{\boldsymbol{\pi} exp(\mathbf{T}t)\mathbf{e}} = \frac{\boldsymbol{\pi} exp(\mathbf{T}t)}{\boldsymbol{\pi} exp(\mathbf{T}t)\mathbf{e}}\boldsymbol{\theta} \tag{2}$$

Där $\frac{\boldsymbol{\pi} exp(\mathbf{T}t)}{\boldsymbol{\pi} exp(\mathbf{T}t)\mathbf{e}}$ är vektorn med sannolikheter att befinna sig i genomgångstillstånden vid tiden t, givet att absorption ännu inte inträffat. Om denna vektor når en gränsfördelning $\boldsymbol{\nu} = (\nu_1, \nu_2, ..., \nu_p)$ som bevaras då tiden $t \to \infty$, sägs $\boldsymbol{\nu}$ vara en kvasistationär fördelning.

Enligt föregående avsnitt kan varje acyklisk samt coxiansk fastypsfördelning representeras i kanonisk coxiansk form. Om man vid tiden t har nått det sista tillståndet är sannolikheten för absorption exponentialfördelad med intensitet θ_p. Vid betingning med att ännu inte ha

nått absorption bevaras denna fördelning då $t \to \infty$. En fördelning med all sannolikhetsmassa i det sista tillståndet, $\boldsymbol{\nu} = (0, 0, ..., 1)$, är alltså en kvasistationär fördelning. För en kanonisk coxiansk fördelning är detta den enda kvasistationära fördelningen (se appendix) och fördelningen konvergerar mot den vid betingning med att inte ha nått absorption.

Om en kanonisk coxiansk fördelning inte har nått absorption går den alltså mot en exponentialfördelning då $t \to \infty$. Av detta följer att även felintensiteten går mot felintensiteten hos en expontentilafördelning, som är konstant. Detta ses även i uttrycket (2). Det är en av egenskaperna hos fastypsfördelningen som gör den intressant för svansskattningar, vilket beskrivs närmare i avsnittet om extremvärdesteori (Asmussen et al. (1996)).

2.2.9 Tät klass

Fastypsfördelningar är en tät klass. Det innebär att för varje fördelning på positiva talaxeln finns en följd av fastypsfördelningar som konvergerar mot fördelningen när ordningen $p \to \infty$. Därmed kan fastypsfördelningen approximera alla positiva, kontinuerliga fördelningar godtyckligt nära (Bladt (2005)). Även coxianska fördelningar är en tät klass (Johnsson, Taaffe (1988)).

2.3 Extremvärdesteori

Extremvärdesteori handlar om att man vill få fram sannolikhetsfördelningen för extrema värden. Här ges två sätt att angripa problemet.

För den första metoden ges en kort sammanfattning av teorin i syfte att ge en övergripande bild av viktiga problemställningar och resultat inom extremvärdesteori. Vår undersökning bygger på metod nummer två. Förklaringen av denna är därför mer ingående. Teorin kring de båda metoderna länkas sedan samman av lämpliga satser.

2.3.1 Klassisk extremvärdesteori

Ett tillvägagångssätt inom extremvärdesteori är att studera fördelningen för maximum i ett växande stickprov. Nedan ges en övergripande sammanfattning utan bevis eller djupare förklaringar. Våra resonemang nedan bygger i allt väsentligt på framställningen i *An Introduction to Statistical Modeling of Extreme Values* av Coles.

För n stycken oberoende, likafördelade stokastiska variabler $X_1, ..., X_n$ med fördelningsfunktion $F(x)$ fås fördelningsfunktionen för maximum, $M_n = max(X_1, ..., X_n)$, av:

$$P(M_n \leq x) = P(X_1 \leq x, ..., X_n \leq x) = P(X_1 \leq x)...P(X_n \leq x) = (F(x))^n.$$

Om $F(x)$ är känd har man alltså fördelningen för maximum M_n. I praktiken intresserar man sig dock för fall där $F(x)$ inte är känd. Om man istället utgår från en approximation av $F(x)$ växer felen då $F(x)$ upphöjs till n och resultaten blir alltför osäkra.

Målet är att approximera en fördelning för M_n då n går mot oändligheten. Men $F(x) < 1$ ger $(F(x))^n \to 0$ då $n \to \infty$, så fördelningen av M_n urartar. För att fördelningen inte ska urarta måste man skala om M_n.

Klassisk extremvärdesteori handlar om att hitta fördelningar för vilka det går att finna följder av omskalningskonstanter a_n och b_n så att $(M_n - b_n)/a_n$ går mot en fördelning $G(x)$ då n går mot oändligheten. Det handlar även om att hitta fördelningar $G(x)$ som kan uppkomma som sådana gränsvärdesfördelningar.

Det visar sig att de enda fördelningar som kan uppkomma som sådana gränsvärdesfördelningar är generaliserade extremvärdesfördelningar.

Definition 2.9. *Gruppen av generaliserade extremvärdesfördelningar är fördelningar med fördelningsfunktion*

$$G(z) = exp\left(-\left(1+\frac{\beta(z-\mu)}{\sigma}\right)^{-\frac{1}{\beta}}\right),$$

definierade för $z > \mu - \frac{\sigma}{\beta}$ *och* $-\infty < \mu < \infty, \sigma > 0, -\infty < \beta < \infty$ *(Falk et al. (1994)).*

Nedan ges även en sats för maxstabilitet. Det är en egenskap vi kommer att använda oss av för att koppla samman de två extremvärdesmetoderna.

Definition 2.10. *En fördelning sägs vara maxstabil om det för varje* $n = 2, 3, ...$ *finns konstanter* $a_n > 0$ *och* b_n *så att* $G^n(a_n z + b_n) = G(z)$.

En fördelning är allstå maxstabil om varje heltalspotens ≥ 2 av fördelningen kan skalas om och translateras till fördelningen själv.

Sats 2.11. *En fördelning är maxstabil om och endast om den är en generaliserad extremvärdesfördelning.*

2.3.2 Peak over threshold

Metod två är det tillvägagångsätt som ligger till grund för vår undersökning. Den går ut på att studera sannolikhetsfördelningen givet att man befinner sig över en hög tröskel u. Det vill säga att undersöka $P(X - u \leq t | X > u) = \frac{F(t+u)-F(u)}{1-F(u)}$. Svårigheten ligger i att man i praktiska tillämpningar inte vet fördelningsfunktionen $F(x)$. Med hjälp av till exempel centrala gränsvärdessatsen kan man approximera $F(x)$. En sådan approximation blir dock dålig i fördelningens svansar, där de extrema värdena finns. En bättre metod är den så kallade peak over threshold-metoden, även kallad POT-metoden. I förklaringen av metoden utgår vi från felintensiteten istället för fördelningsfunktionen. Felintensiteten blir nämligen densamma oavsett om man betingar med att befinna sig över tröskeln u eller inte. Med $s = t + u$ fås nämligen

$$\begin{aligned} h(s|X > u) &= \frac{\frac{d}{ds}P(X \leq s|X > u)}{1 - P(X \leq s|X > u)} \\ &= \frac{\left(\frac{d}{ds}\frac{F(s)-F(u)}{1-F(u)}\right)}{\left(1 - \frac{F(s)-F(u)}{1-F(u)}\right)} \\ &= \frac{\left(\frac{f(s)}{1-F(u)}\right)}{\left(1 - \frac{F(s)-F(u)}{1-F(u)}\right)} \\ &= \frac{f(s)}{1-F(s)} = h(s) \end{aligned}$$

Intuitivt kan detta resultat förklaras av att felintensiteten vid tiden s kan tolkas som intensiteten för fel givet att fel inte har inträffat fram till tiden s.

Att utgå från felintensiteten är möjligt eftersom fördelningsfunktionen $F(x)$, enligt avsnittet om felintensiteter, kan karakteriseras av sin felintensitet $h(x)$ på intervallet $[0, c[$, där c är den minsta punkt sådan att $F(x) = 1$.Vi hoppar över argumenten för att felintensiteten alltid existerar.

2.3.3 Betingningsstabilitet

Antag att $F(x)$ är en positiv fördelning med kontinuerlig, deriverbar täthet. Om felintensiteten $h(x)$ uppfyller

$$h(s) = \frac{h(u + \frac{s}{h(u)})}{h(u)} \text{ för alla } s, u > 0 \tag{3}$$

gäller att

$$F(s) = \frac{F(u + \frac{s}{h(u)}) - F(u)}{1 - F(u)} \tag{4}$$

ty

$$\begin{aligned}
F(s) &= 1 - e^{-\int_0^s h(y)dy} \\
&= 1 - e^{-\int_0^s \frac{h(u+\frac{y}{h(u)})}{h(u)}dy} \\
&= \begin{bmatrix} x = u + \frac{y}{h(u)} \\ dx = \frac{dy}{h(u)} \\ y = 0 \Rightarrow x = u \\ y = s \Rightarrow x = u + \frac{s}{h(u)} \end{bmatrix} \\
&= 1 - e^{-\int_u^{u+\frac{s}{h(u)}} h(x)dx} \\
&= 1 - e^{-(\int_0^{u+\frac{s}{h(u)}} h(x)dx - \int_0^u h(x)dx)} \\
&= 1 - \frac{e^{-\int_0^{u+\frac{s}{h(u)}} h(x)dx}}{e^{-\int_0^u h(x)dx}} \\
&= 1 - \frac{1 - F(u + \frac{s}{h(u)})}{1 - F(u)} \\
&= \frac{1 - F(u)}{1 - F(u)} - \frac{1 - F(u + \frac{s}{h(u)})}{1 - F(u)} \\
&= \frac{F(u + \frac{s}{h(u)}) - F(u)}{1 - F(u)}
\end{aligned} \tag{5}$$

Observera att $(F(u + \frac{s}{h(u)}) - F(u))/(1 - F(u)) = P(h(u)(X - u) \leq u | X > u)$. $F(s)$ ska alltså vara en omskalad version av sin betingade fördelning över en hög tröskel. Om kravet är uppfyllt kan vi använda $F(x)$ för uttalanden om de extrema värden över tröskeln som vi är intresserade av. Omskalningen med felintensiteten är praktisk då en fördelning som uppfyller villkoret automatiskt även uppfyller villkoret att $f(0) = 1$. Det ger en skalparameter mindre i felintensiteten $h(t)$ som uppfyller dessa villkor.

Det går att få fram uttryck för felintensiteten $h(t)$. Genom att anta att vi får derivera $h(t)$ med avseende på t och sätta $t = 0$ får vi $h'(0) = \frac{h'(t)}{h(t)^2}$. Eftersom $h(0) = 1$ har differentialekvationen lösningarna $h(t) = 1$ om $h'(0) = 0$ och $h(t) = \frac{1}{1+At}$ för någon konstant A om $h(0) \neq 0$. (Se appendix för mer detaljerade beräkningar.)

Med hjälp av detta uttryck för felintensiteten kan man även få fram c, ändpunkten i intervallet för felintensiteten:

$$1 = F(c) = 1 - e^{-\int_0^c \frac{1}{At+1}dt},$$

så

$$0 = e^{-\int_0^c \frac{1}{At+1} dt}.$$

Vilket ger $c = \infty$ om $A > 0$ och $c = -\frac{1}{A}$ om $A < 0$.

De felintensiteter som fås fram leder till följande fördelningsfunktioner,

$$F(t) = \begin{cases} 1 - e^{-\int_0^t \frac{1}{As+1} ds} = 1 - e^{\left[-\frac{ln(1+As)}{A}\right]_0^t} = 1 - (1+At)^{-\frac{1}{A}}, & \text{för } h(t) = \frac{1}{1+At} \\ 1 - e^{-\int_0^t 1 ds} = 1 - e^{-[s]_0^t} = 1 - e^{-t}, & \text{för } h(t) = 1 \end{cases} \tag{6}$$

Resultatet ovan är den normerade generaliserade paretofördelningen. Kravet (4) som ställs på fördelningsfunktionen, dvs att den är en omskalad version av sin betingade fördelning över en hög tröskel, är en form av betingningsstabilitet. Begreppet betingningsstabilitet tillåter även omskalning med andra funktioner än felintensiteten och att parametrarna i fördelningen ändras.

Om vi släpper på kravet att $f(0) = 1$ så fås extra parametrar i (6). Den grupp fördelningar som då uppfyller kraven på betingningsstabilitet kallas generaliserade paretofördelningar, förkortas gp-fördelningar och brukar parametriseras enligt följande definition (Falk et al. (1994)).

Definition 2.12. *Gruppen av generaliserade paretofördelningar har fördelningsfunktion*

$$F(t) = \begin{cases} 1 - \left(1 + \frac{\beta(t-\mu)}{\sigma}\right)^{-\frac{1}{\beta}}, & \beta \neq 0 \\ 1 - e^{\frac{t-\mu}{\sigma}}, & \beta = 0 \end{cases}$$

definierad för $t \geq \mu$ *då* $\beta \geq 0$ *samt* $\mu < t < \mu - \frac{\sigma}{\beta}$ *då* $\beta < 0$.

En teori som bara gäller för gp-fördelningar är i praktiken för snäv för att vara användbar. Man behöver kunna uttala sig om extrema värden även när sådana krav på fördelningen inte är uppfyllda.

2.3.4 Konvergens mot betingningsstabilitet

Antag istället att X är en stokastisk variabel på reella talaxeln med fördelningsfunktion G och felintensitet g som uppfyller:

$$h(t) = \lim_{u \to \infty} \frac{g(u + \frac{t}{g(u)})}{g(u)} \text{ likformigt på varje intervall } t \in [0, c].$$

Då löser h ekvationssystemet (3):

Sätt

$$u' = u + \frac{v}{g(u)} \ .$$

Då fås

$$\begin{aligned} h(v+\frac{t}{h(v)}) &= h\left(v+t\lim_{u\to\infty}\frac{g(u)}{g(u')}\right) \\ &= \lim_{u\to\infty}\frac{g\left(u+\frac{v+t\frac{g(u)}{g(u')}}{g(u)}\right)}{g(u)} \\ &= \lim_{u\to\infty}\frac{g\left((u+\frac{v}{g(u)})+\frac{t}{g(u')}\right)}{g(u')}\frac{g(u')}{g(u)} \\ &= h(t)h(v). \end{aligned}$$

Så

$$h(t) = \frac{h\left(v+\frac{t}{h(v)}\right)}{h(v)}.$$

Det ger att den betingade, skalade fördelningsfunktionen för X konvergerar mot en positiv stokastisk variabel med fördelningsfunktion $F(t)$. Det vill säga

$$F(t) = \lim_{u\to\infty}\frac{G(u+t/g(u))-G(u)}{1-G(u)},$$

ty

$$\begin{aligned} F(t) &= 1-e^{-\int_0^t h(y)dy} \\ &= 1-e^{-\int_0^t\left(\lim_{t\to\infty}\frac{g(u+\frac{y}{g(u)})}{g(u)}\right)dy} \\ &= \begin{bmatrix} x = u+\frac{y}{g(u)} \\ dx = \frac{dy}{g(u)} \\ y=0 \Rightarrow x = u \\ y = t \Rightarrow x = u+\frac{t}{g(u)} \end{bmatrix} \\ &= \lim_{u\to\infty}(1-e^{-\int_u^{u+\frac{t}{g(u)}} g(x)dx}) \\ &= \lim_{u\to\infty}(1-e^{-(\int_{-\infty}^{u+\frac{t}{g(u)}} g(x)dx-\int_{-\infty}^{u} g(x)dx)}) \\ &= \lim_{u\to\infty}\left(1-\frac{e^{-\int_{-\infty}^{u+\frac{t}{g(u)}} g(x)dx}}{e^{-\int_{-\infty}^{u} g(x)dx}}\right) \\ &= \lim_{u\to\infty}\left(1-\frac{1-G(u+\frac{t}{g(u)})}{1-G(u)}\right) \\ &= \lim_{u\to\infty}\left(\frac{G(u+\frac{t}{g(u)})-G(u)}{1-G(u)}\right). \end{aligned}$$

Kraven som ställs på fördelningen kan göras ännu något allmännare men det tas inte upp i denna uppsats.

Om $F(t)$ är känd och tröskeln u är tillräckligt hög kan därför $F(t)$ användas för att skatta sannolikheten för värden över tröskeln. $F(t)$ är betingningsstabil enligt steg 1 och alltså en gp-fördelning. Gp-fördelningar kan med andra ord användas för att skatta sannolikheter även i fall där ursprungsfördelningen inte är en gp-fördelning. Detta är peak over threshold-metodens centrala resultat.

2.3.5 Sammankoppling av metoderna

Antag att $X_1, ..., X_n$ är n stycken oberoende, stokastiska variabler med gp-fördelning $F(t)$. För att kunna koppla samman POT-teorin med den klassiska extresmvärdesteorin vill vi undersöka fördelningen för maximum av dessa variabler.

Vi har att

$$P(max(X_1, ..X_n) \leq t) = P(X_1 \leq t, ..., X_n \leq t).$$

Men att alla observationer är mindre än t är detsamma som att ingen observation är större än t. Sannolikheten att en observation är större än t är $1 - F(t)$, så antalet observationer som är större än t är binomialfördelat$(n, 1 - F(t))$. Eftersom vi intresserar oss för stora t kommer sannolikheten att en observation är större än t vara liten, vilket betyder att $1-F(t)$ kommer att ligga nära noll. När n är stort och binomialfördelningens sannolikhetsparameter är liten kan man använda sig av poissonapproximation. Så antalet observationer större än t är approximativt poissonfördelat$(n(1 - F(t))$.

Vi får

$$\begin{aligned}
&P(max(X_1, ..X_n) \leq t) \\
&= P(\text{noll observationer } >\text{t}) \\
&\approx \frac{(n(1-F(t)))^0}{0!} e^{-(n(1-F(t))} \\
&= e^{-(n(1-F(t))} \\
&= \left(e^{-(1-F(t))}\right)^n \\
&= \left(exp\left(-\left(1-\left(1-\left(1+\frac{\beta(t-\mu)}{\sigma}\right)^{-\frac{1}{\beta}}\right)\right)\right)\right)^n \\
&= \left(exp\left(-\left(1+\frac{\beta(t-\mu)}{\sigma}\right)^{-\frac{1}{\beta}}\right)\right)^n \\
&= exp\left(-\left(1+\frac{\beta(t-\mu_n)}{\sigma_n}\right)^{-\frac{1}{\beta}}\right)
\end{aligned}$$

I näst sista likheten finner vi en generaliserad extremvärdesfördelning upphöjd till n. Men en generaliserad extremvärdesfördelning är maxstabil och fördelningen upphöjd till n är därför också en generaliserad extremvärdesfördelning. Maxstabiliteten ger därför en ny generaliserad extremsvärdesfördelning med parametrar μ_n och σ_n. Men då den generaliserade extremvärdesfördelningen och gp-fördelningen inte har samma stöd så måste parametrarna uppfylla

$$\lim_{n\to\infty} \frac{\mu_n}{\sigma_n} = \infty,$$

för att t ska kunna bli $< \mu$. Detta är inte något vi visar.

Resultatet innebär att maximum för många oberoende variabler med en betingningsstabil fördelning är approximativt maxstabil. Det visar på sambandet mellan de två metoderna och formparametern β.

Sambandet mellan metoderna gäller även åt andra hållet. Vi ger en sats och ett förenklat bevis för att en fördelning som konvergerar mot en generaliserad extremvärdesfördelning har en betingad överskottsfördelning som konvergerar mot en generaliserad paretofördelning.

Observera att denna sats ger ett starkare resultat eftersom den inte gäller endast för en generaliserad extremvärdesfördelning, utan också för de fördelningar som har maximum som konvergerar mot en sådan.

Sats 2.13. *Pickands-Balkema-de Haans sats*
Låt $X_1, ..., X_n$ vara en följd av oberoende slumpvariabler med gemensam fördelningsfunktion F och låt $M_n = max(X_1, ..., X_n)$. Antag att för tillräckligt stora n gäller $P(M_n \leq z) \approx G(z)$ där G är en generaliserad extremvärdesfördelning med parametrar μ, $\sigma_1 > 0$ och β. Då gäller för tillräckligt stora u att

$$(X - u | X > u) \approx H(z),$$

där H(z) är en gp-fördelning med parametrar β och σ_2, där $\sigma_2 = \sigma_1 + \beta(u - \mu)$.

Bevisskiss

Låt X vara en slumpvariabel med fördelningsfunktion F. Enligt satsens antagande gäller för tillräckligt stora n att

$$F^n(z) \approx exp\left(-\left(1 + \frac{\beta(z-\mu)}{\sigma_1}\right)^{-\frac{1}{\beta}}\right),$$

för parametrar μ, $\sigma_1 > 0$ och β.

Så

$$n \ln F(z) \approx -\left(1 + \frac{\beta(z-\mu)}{\sigma_1}\right)^{-\frac{1}{\beta}}.$$

För stora z ger Taylorutveckling att

$$n \ln F(z) \approx -(1 - F(z)).$$

Insättning av detta i föregående uttryck ger för tillräckligt stora u

$$1 - F(u) \approx \frac{1}{n}\left(1 + \frac{\beta(u-\mu)}{\sigma_1}\right)^{-\frac{1}{\beta}}.$$

På samma sätt gäller för y>0,

$$1 - F(u+y) \approx \frac{1}{n}\left(1 + \frac{\beta(u+y-\mu)}{\sigma_1}\right)^{-\frac{1}{\beta}}.$$

Vilket ger att

$$P(X > u + y | X > u) \approx \frac{\frac{1}{n}\left(1 + \frac{\beta(u+y-\mu)}{\sigma_1}\right)^{-\frac{1}{\beta}}}{\frac{1}{n}\left(1 + \frac{\beta(u-\mu)}{\sigma_1}\right)^{-\frac{1}{\beta}}}$$
$$= \left(1 + \frac{\frac{\beta(u+y-\mu)}{\sigma_1}}{1 + \frac{\xi(u-\mu)}{\sigma_1}}\right)^{-\frac{1}{\beta}}$$
$$= \left(1 + \frac{\beta y}{\sigma_2}\right)^{-\frac{1}{\beta}},$$

där $\sigma_2 = \sigma_1 + \beta(u - \mu)$. Så den betingade överskottsfördelningen är en generaliserad paretofördelning (Falk et al. (1994), Coles(2001)).

2.4 Argument för att fördelningarna i vår undersökning konvergerar

I vår undersökning har vi för ett antal teoretiska fördelningar jämfört POT-metoden med metoden att approximera svansen i fördelningen med en fastypsfördelning. POT-metoden, som förklaras mer ingående i nästa kapitel, går ut på att man använder observationerna över en tröskel i ett stickprov till att skatta parametrarna i en gp-fördelning. För att POT-metoden ska vara relevant har vi valt att undersöka teoretiska fördelningar där vi vet att den betingade överskottsfördelningen konvergerar mot en gp-fördelning.

Maximum för en t-fördelning (Zholud (2011)), normalfördelning, exponentialfördelning, lognormalfördelning och paretofördelning konvergerar alla mot en generaliserad extremvärdesfördelning (Castillo (1988)). Enligt satsen i föregående avsnitt konvergerar den betingade överskottsfördelningen därför mot en gp-fördelning. Vi ger inga utförliga bevis för konvergensen hos varje fördelning utan nöjer oss med att stärka uttalandet med en rad argument för följande fördelningar:

2.4.1 Paretofördelningen

Paretofördelningen tillhör klassen av gp-fördelningar. Som vi skrev i avsnittet om peak over threshold har generaliserade paretofördelningar egenskapen att även den betingade överskottsfördelningen är en gp-fördelning. POT-metoden är i det här fallet alltså uppenbart relevant.

2.4.2 Exponentialfördelningen

Även exponentialfördelningen, med fördelningsfunktion $F(x) = 1-e^{-x}$, tillhör gruppen generaliserade paretofördelningar. Här kan vi enkelt visa att fördelningen bevaras i den betingade överskottsfördelningen.

$$P(X - t \leq u | X > t) \quad = \frac{F(t+u) - F(u)}{1 - F(u)} = \frac{1 - e^{-(t+u)} - (1 - e^{-u})}{1 - (1 - e^{-u})} \quad = \frac{e^{-u}(1 - e^{-t})}{e^{-u}} = 1 - e^{-t}$$

2.4.3 Fastypsfördelningen

Fastypsfördelningens felintensitet går som tidigare nämnts mot en konstant. Det innebär att även den omskalade felintensiteten går mot en konstant. Den omskalade felintensiteten går alltså mot felintensiteten hos en exponentialfördelning, som är en generaliserad paretofördelning. Kopplingen mellan felintensiteten och fördelningsfunktionen ger därför enligt avsnittet om extremvärdesteori att fastypsfördelningen går mot en gp-fördelning i gräns.

2.4.4 Normalfördelningen (0,1)

För felintensiteten hos en normalfördelad variabel med väntevärde 0 och standardavvikelse 1 gäller

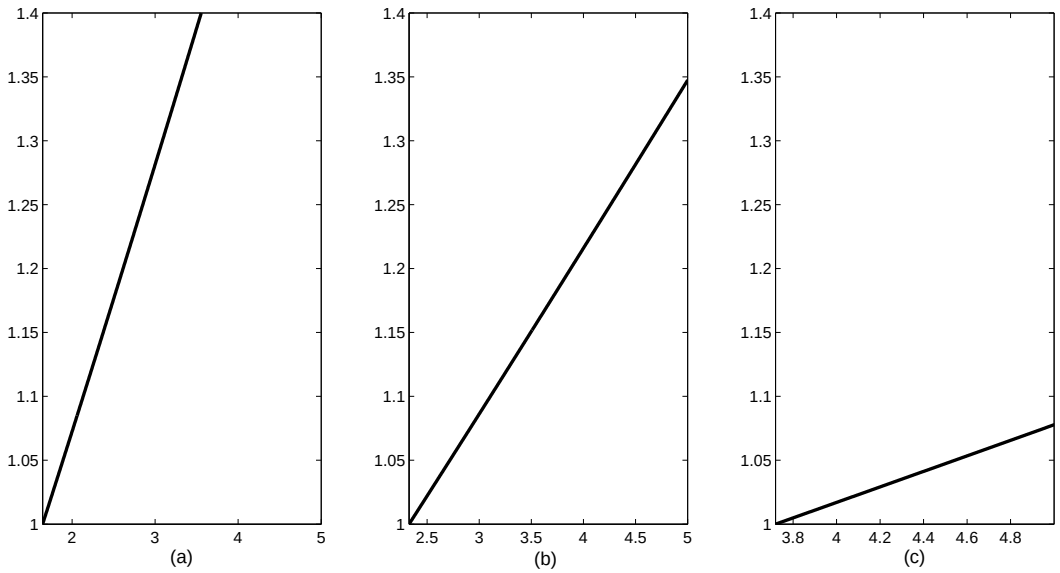
$$h(u) = \frac{\varphi(u)}{1 - \Phi(u)} \approx u,$$

där $\varphi(u)$ är tätheten och $\Phi(u)$ är fördelningsfunktionen (Råde, Westergren (2008)).

Så

$$\frac{h(t + \frac{u}{h(t)})}{h(t)} \approx \frac{t + \frac{u}{t}}{t} = 1 + \frac{u}{t^2} \to 1 \text{ då } t \to \infty.$$

Så den omskalade felintensiteten går mot en konstant som är felintensitetn hos en exponentialfördelning. Det ger, enligt samma argument som för fastypsfördelningen, att den omskalade överskottsfördelningen konvergerar mot en gp-fördelning. I figur 4 ses hur den omskalade felintensiteten för normalfördelningen blir flackare då tröskeln blir högre.



Figur 4: Felintensiteten för den omskalade överskottsfördelningen för normalfördelningen över tre trösklar: (a) 95-percentilen (≈ 1.96), (b) 99-percentilen (≈ 2.33), (c) 99.99-percentilen (≈ 3.72).

2.4.5 Lognormalfördelningen (0,1)

Om X är en lognormalfördelad variabel med väntevärde 0 och standardavvikelse 1 gäller att $\ln X$ är normalfördelad med samma parametrar. Så

$$F(u) = P(X \leq u) = P(\ln X \leq \ln u) = \Phi(\ln u)$$

och

$$f(u) = \varphi(\ln u)\frac{1}{u}$$

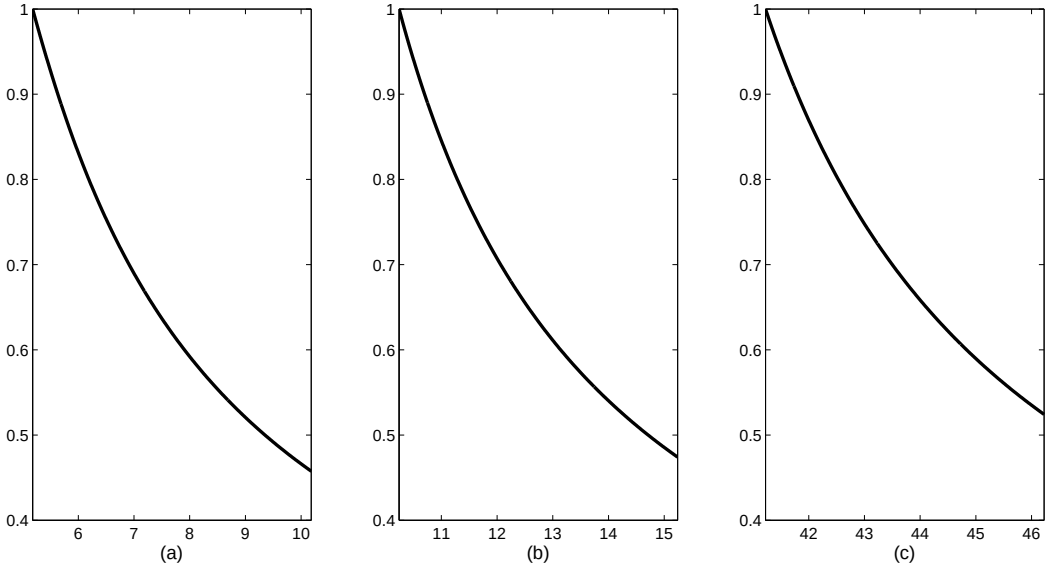
med felintensitet

$$h(u) = \frac{f(u)}{1 - F(u)} = \frac{\varphi(\ln u)}{u(1 - \Phi(\ln u))} \approx \frac{\ln u}{u} \ .$$

På samma sätt som för normalfördelningen fås därför att

$$\begin{aligned}
\frac{h(t + \frac{u}{h(t)})}{h(t)} &\approx \frac{\left(\frac{\ln\left(t + \frac{u}{\ln t}\right)}{t + \frac{u}{\ln t}} \right)}{\left(\frac{\ln t}{t}\right)} \\
&= \frac{\ln\left(t + \frac{tu}{\ln t}\right) t}{\left(t + \frac{tu}{\ln t}\right) \ln t} \\
&= \frac{\ln t}{\ln t + u} + \frac{\ln\left(1 + \frac{u}{\ln t}\right)}{\ln t + u} \\
&= \frac{1}{1 + \frac{u}{\ln t}} + \frac{\ln\left(1 + \frac{u}{\ln t}\right)}{\ln t + u} \to 1 \text{ då } t \to \infty \ .
\end{aligned}$$

Så även en lognormalfördelad variabel har en omskalad överskottsfördelning som går mot en gp-fördelning. I figur 5 ses den approximerade felintensiteten för den omskalade överskottsfördelningen för en lognormalfördelning över tre trösklar. Skillnaden i utseendet för felintensiteten för de olika trösklarna är knappt märkbar, konvergensen mot en konstant felintensitet är uppenbarligen mycket långsam.



Figur 5: Den approximerade felintensiteten för den omskalade överskottsfördelningen för log normalfördelningen över tre trösklar: (a) 95-percentilen (≈ 5.18), (b) 99-percentilen (≈ 10.24), (c) 99.99-percentilen (≈ 41.22).

2.4.6 Skattning av parametrar till stickprov

Ovan ger vi argument för att den omskalade överskottsfördelningen konvergerar mot en gp-fördelning. I verkligheten skalas data inte om utan man använder ett stickprov av överskotts-

värden som det är. Låt X vara en stokastisk variabel som har en överskottsfördelning som omskalad med $k(u)$ konvergerar mot en gp-fördelad stokastisk variabel Z med parametrar β, μ, σ. Då inses att det inte är ett problem ty

$$\begin{aligned} P(k(u)(X-u) < t | X > t) &\approx P(Z < t) \\ &\Downarrow \\ P(k(u)(X-u) < tk(u) | X > t) &\approx P(Z < tk(u)) \\ &\Updownarrow \\ P((X-u) < t | X > t) &\approx P(Z < tk(u)) = \\ &= 1 - \left(1 + \beta \frac{t - \mu/k(u)}{\sigma/k(u)}\right)^{-\frac{1}{\beta}} . \end{aligned}$$

Det innebär att man kan skatta parametrar i gp-fördelningen direkt från ett stickprov av överskottsvärden.

3 Undersökning med simulering

Vi utvärderar två olika metoder för att skatta sannolikheter för värden långt ut i svansen. Detta görs för data från sex olika teoretiska fördelningar. I utvärderingen undersöker vi hur bra metoderna är på att skatta den verkliga sannolikheten att i en fördelning hamna ovanför $q_{0.001} : P(X > q_{0.001}) = 0.001$. Vi har valt att para resultaten och jämföra avståndet till 0.001. Detta eftersom undersökningens omfattning är för liten för att analysera systematiskt fel och standardfel på ett meningsfullt sätt. Anledningen till att en större undersökning inte kunnat genomföras är på grund av begränsningar i EMpht-programmet, som används för att skatta parametrarna i fastypsfördelningar.

3.1 Genomgång av undersökningens metoder

3.1.1 POT-metoden

POT-metoden (peak over threshold) används för att från ett stickprov skatta sannolikheter för extrema värden. För ett givet stickprov behåller man endast de värden som överstiger en viss tröskel u, vilket resulterar i ett nytt stickprov som innehåller överskottsvärden. I enlighet med resultaten från föregående kapitel approximerar man sedan svansen över tröskeln u med en gp-fördelning. Detta eftersom en fördelnings överskott konvergerar i fördelning mot en gp-fördelning, givet att den alls konvergerar.

Det är alltså något vanskligt att approximera överskottsfördelningen med en gp-fördelning, eftersom det inte är säkert att ursprungsfördelningen har en överskottsfördelning som konvergerar. En anledning till att ändå använda sig av denna approximation är att det inte finns några andra modeller att förlita sig på, och den används bland annat inom ekonomi och meteorologi (Coles, (2001)).

En viktig fråga är hur tröskeln u ska väljas. Eftersom approximationen bygger på ett asymptotiskt resultat för u, borde den rimligen fungera bättre då u flyttas i riktning mot ∞. Men då detta i praktiken betyder att antalet observationer som finns kvar i stickprovet minskar är det inte bra att välja tröskeln allt för högt. Det handlar om att göra en avvägning mellan systematiskt fel och spridning (Coles, 2001). Är tröskeln för låg ökar det systematiska felet och är den för hög gör det lilla antalet observationer att skattningen får stor varians.

Parametrarna i den generaliserade paretofördelningen skattas med maximum likelihood-metoden, genom en inbyggd funktion i MATLAB. Ett problem som kan uppstå är att maximum likelihood-skattningarna kan ge parameteruppsättningar som gör att fördelningen har

ändligt stöd, dvs. har sannolikheten noll för utfall över en ändlig tröskel.

Eftersom sannolikheterna vi vill skatta ligger långt ut i svansen kommer ändligt stöd att bli ett stort problem om stödets övre gräns befinner sig nedanför $q_{0.001}$.

3.1.2 Fastypsmetoden

Fastypsfördelningarnas approximationsegenskaper gör att de kan användas för att approximera en fördelnings svans. Fastypsmetoden går ut på att skatta parametrarna till en fastypsfördelning från ett stickprov av överskottsvärden. Vilket innebär att man modifierar POT-metoden genom att byta ut antagandet om överskottsfördelningar från en gp-fördelning till en fastypsfördelning av fix ordning p. Vi tror att denna ansats kanske är ny.

Enligt avsnittet om fastypsfördelningar kan en generell fastypsfördelning approximera vilken annan positiv, kontinuerlig fördelning som helst godtyckligt nära. Enligt Asmussen et al. (1995) fungerar ofta en approximering med en coxiansk fördelning lika bra som en med generell fastypsfördelning. Dessutom har, som redan nämnts, alla acykliska fastypsfördelningar en representation i form av en kanonisk coxiansk fördelning.

Ovanstående resultat för den coxianska klassen av fastypsfördelningar motiverar varför vi endast använder oss av dessa då vi utvärderar fastypsmetoden. Det är en naturlig inskränkning, dels med anledning av redan nämnda resultat, dels för att minska körtiden för programvaran, då det i en coxiansk fördelning blir avsevärt färre parametrar att skatta. Viktigt att förstå är att det vid en sådan skattning av parametrar i en fastypsfördelning inte nödvändigtvis existerar någon tolkning av de olika tillstånden i den bakomliggande markovkedjan.

För att utvärdera metoden behåller vi precis som tidigare de observationer i ett stickprov som överstiger tröskeln u. För stickprovet med överskottsvärden skattas sedan parametrarna i en coxiansk fördelning. Parametrarna skattas med hjälp av programvaran EMpht, skriven av Marita Olsson. EMpht använder EM-algoritmen för att beräkna maximum likelihood-skattningar av parametrar. EM-algoritmen garanterar att likelihooden för parameteruppsättningen ökar i varje iteration, samt att den konvergerar (Olsson (1995)). En nackdel med EM-algoritmen är att man inte vet om den parameteruppsättning som programmet konvergerar mot är ett globalt maximum för likelihoodfunktionen. Vid användning av EMpht bör man därför köra programmet flera gånger, och i varje körning använda ett nytt frö för att upptäcka om man får samma skattningar eller ej. Det har vi inte gjort, eftersom det både är tidsödande att köra programmet flera gånger per stickprov och hade försvårat jämförbarheten mellan de två metoderna.

3.2 Detaljer kring undersökningen

Vi har gjort en första undersökning av hur rimligt resultat man får med fastypsmetoden i förhållande till POT-metoden för ett antal teoretiska fördelningar. I undersökningen har ett stickprov på 1000 värden simulerats från var och en av sex teoretiska fördelningar. Därefter har skattningar med POT-metoden och med fastypsmetoden av ordning 4 jämförts m.a.p. fördelningsfunktionen i 99.9-percentilen för olika trösklar. De fördelningar som använts är:

- Normalfördelning med parametrar $\mu = 0, \sigma = 1$.
- Lognormalfördelning med parametrar $\mu = 0, \sigma = 1$.
- Paretofördelningen, $f(x) = \alpha \frac{\mu^\alpha}{x^\alpha}$ då $x \geq \mu$, med parametrarna $\alpha = 3$ och $\mu = 3$.
- t-fördelningen med 6 frihetsgrader.

- Kanonisk coxiansk fördelning av ordning $p = 10$.
- Exponentialfördelning med intensitet $\lambda = 3$.

För ett givet stickprov har sedan observationer ovanför tröskeln u behållits. Undersökningen har genomförts för tre olika värden på u. Det är från stickprovet av överskottsvärden som parametrarna i en gp-fördelning samt en coxiansk fördelning av ordning 4 skattas. Parametrarna i den coxianska fördelningen skattas iterativt med programvaran EMpht, som för alla skattningar körts med 4000 iterationer. I fallet då data kom från den kanonisk coxianska fördelningen har även parametrarna till en coxiansk fördelning av ordning 2 skattats. Det skedde endast för det högsta värdet på u.

Detta resulterar i två approximationer av fördelningens svans ovanför u. Vi kallar dem F_1^*, som är gp-approximationen, respektive F_2^*, som är fastypsapproximationen. För att kunna evaluera metoderna behövs skattningar av $P(X > q_{0.001})$, där X är en stokastisk variabel med samma fördelning som stickprovet och q är motsvarande kvantil. Som dessa skattningar används

$$0.001 = \gamma = P(X > q_{0.001}) \approx \frac{n'}{n}(1 - F_i^*(q_{0.001})) = \hat{\gamma}_i, \quad i = 1, 2, \tag{7}$$

där n' är antalet observationer i stickprovet av överskottsvärden och n antalet observationer i det ursprungliga stickprovet. Att detta är rimliga skattningar av $P(X > q_{0.001})$ inses enklast genom att anta att F_i^* ger den bästa tänkbara approximationen, nämligen

$$F_i^*(s) = \frac{F(s) - F(u)}{1 - F(u)}, \quad s \in [u, \infty[.$$

Då blir skattningarna

$$\hat{\gamma}_i = \frac{n'}{n}\left(1 - \frac{F(q_{0.001}) - F(u)}{1 - F(u)}\right) = \frac{n'}{n}\left(\frac{1 - F(q_{0.001})}{1 - F(u)}\right).$$

Men eftersom man kan se n' som utfallet av en stokastisk variabel $N' \sim Bin(n, 1-F(u))$, så är $\frac{n'}{n}$ är en väntevärdesriktig skattning av $P(X > u) = 1 - F(u)$. Detta leder till att

$$\begin{aligned} E[\hat{\gamma}_i] &= E\left[\frac{n'}{n}\left(\frac{1 - F(q_{0.001})}{1 - F(u)}\right)\right] = E\left[\frac{n'}{n}\right]\left(\frac{1 - F(q_{0.001})}{1 - F(u)}\right) \\ &= 1 - F(q_{0.001}) = P(X > q_{0.001}). \end{aligned}$$

Så $\hat{\gamma}_i, i = 1, 2$, är en väntevärdesriktig skattning då svansapproximationen är exakt. Det är såklart något som inte gäller i realiteten, men är ett belägg för att skattningarna är användbara om approximationerna med gp-fördelning respektive coxiansk fördelning är bra.

Tröskeln u sätts först till 95-percentilen sedan till 97,5-percentilen samt 99-percentilen. För varje värde på u görs undersökningen på samma ursprungliga stickprov. För vart och ett av tröskelvärdena fås alltså ett nytt (mindre eller lika stort) stickprov av överskottsvärden.

För var och en av de sex fördelningarna genomförs ovanstående undersökning 10 gånger. Allt som allt resulterar alltså undersökningen i 30 skattningar för varje fördelning. Skattningar av $P(X > q_{0.001})$ från samma stickprov är beroende av varandra, övriga skattningar är oberoende.

Vi har valt att jämföra metoderna parvis. Vi undersöker med andra ord hur bra skattningar de ger för samma data. Detta eftersom vi har haft en begränsning i hur många simuleringar vi har kunnat göra, och samtidigt vill minimera onödig varians i hur de presterade.

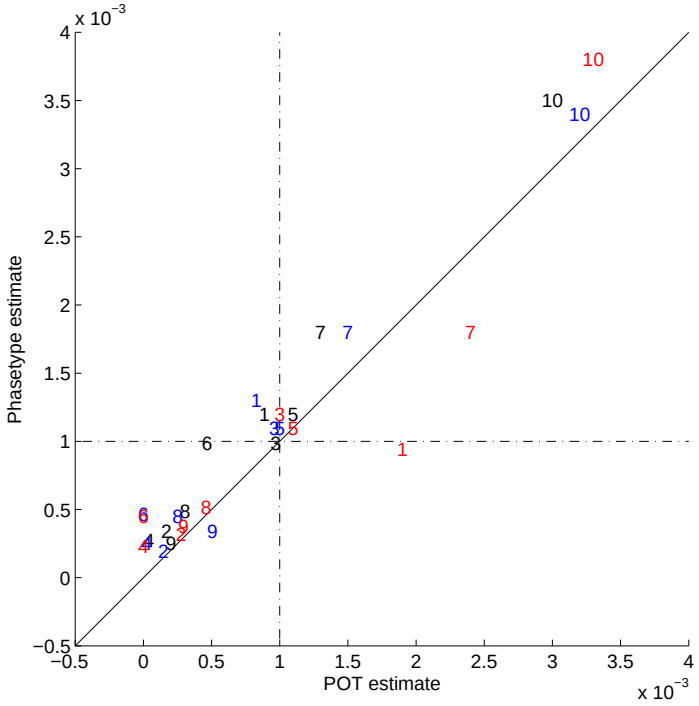
3.3 Resultat

Här redogörs för resultatet av undersökningen som den är beskriven ovan. Fokus ligger på att undersöka i hur många fall respektive metod presterade bäst. Med bäst avses att ha en skattning $\hat{\gamma}_i, i = 1, 2$, som ligger närmast $\gamma = 0.001$. Eftersom vi gör så få simuleringar blir slutsatserna inte helt entydiga.

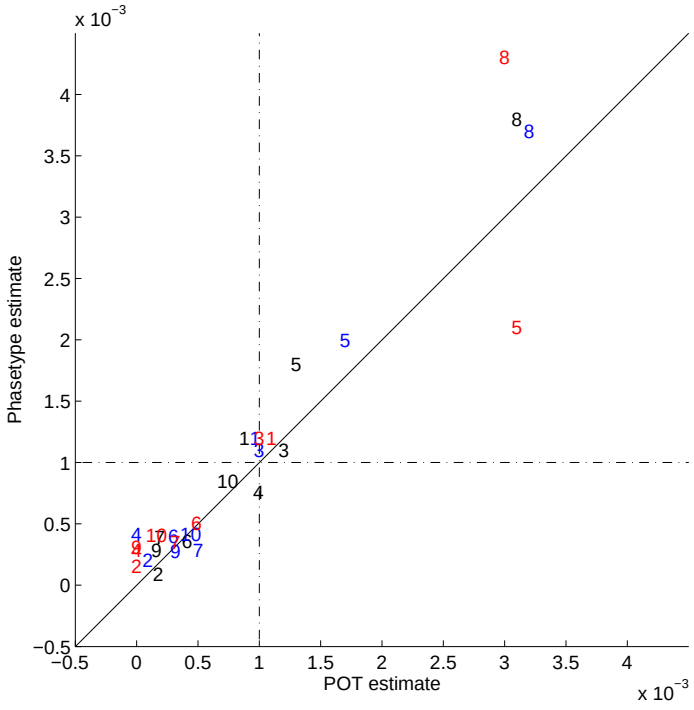
Till varje fördelning finns ett spridningsdiagram som visar skattningen från båda metoderna. I spridningsdiagrammen syns värdet på båda metoderna för alla trettio skattningar. X- och y-axeln anger värdet på skattningen från POT- respektive fastypsmetoden. Siffrorna i spridningsdiagrammen anger vilken av de tio simuleringarna som skattningen hör till. Färgen på siffrorna anger vilken tröskel som skattningen hör till; svart betyder 95-percentilen, blå betyder 97.5-percentilen och röd betyder 99-percentilen. Om en siffra hamnar på diagonalen betyder det att båda metoder har presterat lika bra. I den nedre rutan till vänster har fastypsmetoden presterat bäst för punkter ovanför diagonalen och POT-metoden har presterat bäst för punkter under diagonalen. I den övre rutan till höger har istället POT-metoden presterat bäst för punkter ovanför diagonalen och fastypsmetoden för punkter under.

3.3.1 Normalfördelning (0,1)

För den standardiserade normalfördelningen presterade fastypsmetoden något bättre. Mest anmärkningsvärt i resultaten är att en märkbar majoritet av simuleringarna resulterade i F_1^* med ändligt stöd. Resultatet för normalfördelningen ses i figur 6.



Figur 6: Resultatet från undersökningen då data simulerades från en normalfördelning. Svart siffra = 95-percentilen, blå siffra = 97.5-percentilen och röd siffra = 99-percentilen.



Figur 7: Resultatet från undersökningen då data simulerades från en lognormalfördelning. Svart siffra = 95-percentilen, blå siffra = 97.5-percentilen och röd siffra = 99-percentilen.

3.3.2 Lognormalfördelning (0,1)

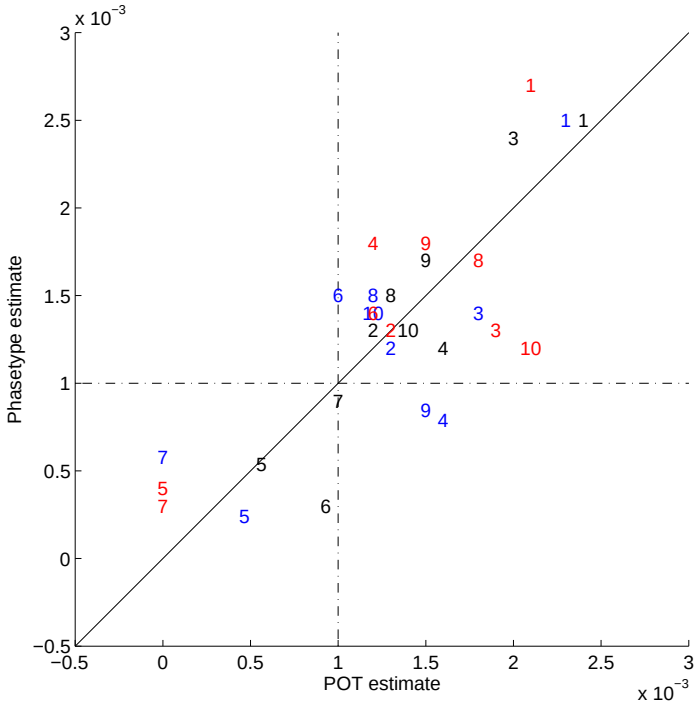
I fallet med lognormalfördelningen presterade POT-metoden något bättre än fastypsmetoden för de två lägre trösklarna. Resultatet för lognormalfördelningen ses i figur 7.

Värt att nämna är att för $u = 97,5$-percentilen förekommer fyra fall där F_1^* har en parameteruppsättning som ger ett ändligt stöd. I tre av dessa fall var POT-skattningen sämre än fastypsskattningen. Samma tendens märks för $u = 99$-percentilen. I fem fall har då F_1^* ändligt stöd och tillhörande POT-skattningar var alla sämre än fastypsskattningarna.

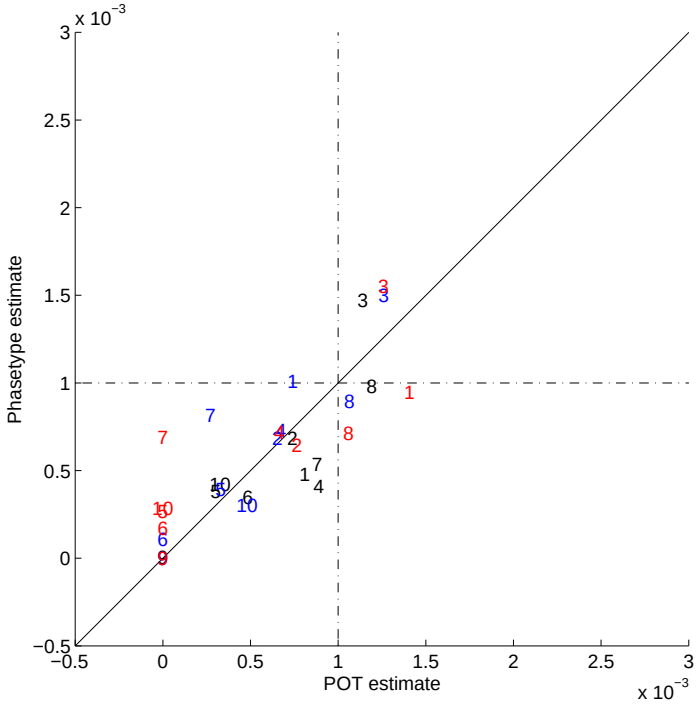
3.3.3 Paretofördelning

Föga förvånande fås här att POT-metoden över lag är något bättre. Även här verkar POT-metoden prestera bäst för det lägsta tröskelvärdet. Resultatet för paretofördelningen ses i figur 8.

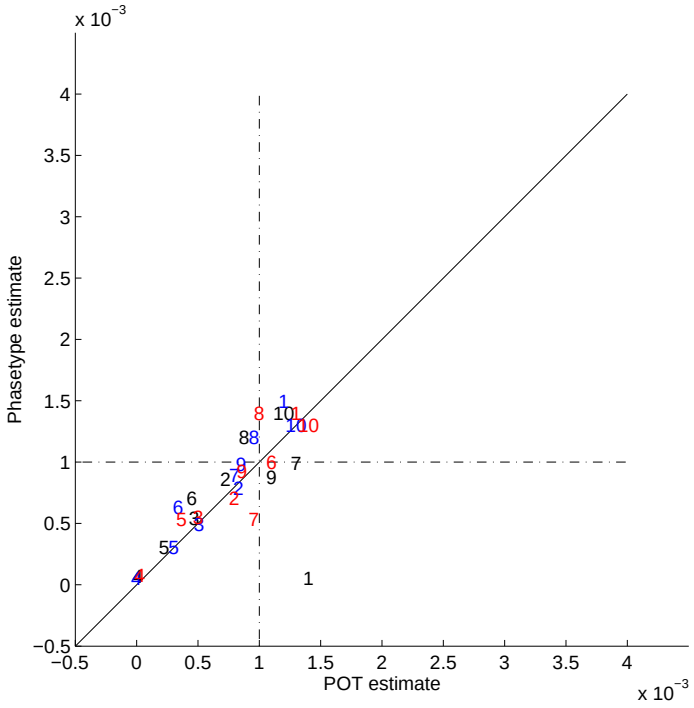
För $u = 97,5$-kvantilen fås ett fall då F_1^* har ändligt stöd, tillhörande POT-skattning är också sämre än motsvarande fastypsskattning. Även i gruppen som hör till $u = 99$-percentilen blir POT-skattningen sämre i alla fall av ändligt stöd.



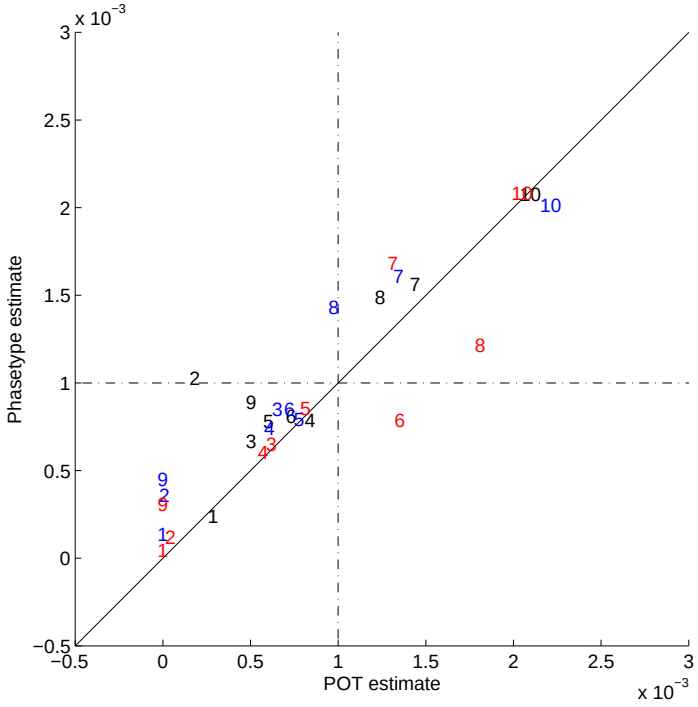
Figur 8: Resultatet från undersökningen då data simulerades från en paretofördelning. Svart siffra = 95-percentilen, blå siffra = 97.5-percentilen och röd siffra = 99-percentilen.



Figur 9: Resultatet från undersökningen då data simulerades från en kanonisk coxiansk fördelning. Svart siffra = 95-percentilen, blå siffra = 97.5-percentilen och röd siffra = 99-percentilen.



Figur 10: Resultatet från undersökningen då data simulerades från en t-fördelning. Svart siffra = 95-percentilen, blå siffra = 97.5-percentilen och röd siffra = 99-percentilen.



Figur 11: Resultatet från undersökningen då data simulerades från en exponentialfördelning. Svart siffra = 95-percentilen, blå siffra = 97.5-percentilen och röd siffra = 99-percentilen.

3.3.4 Kanonisk coxiansk fördelning av ordning 10

För att utvärdera hur bra de två metoderna presterar när data kommer från en coxiansk fördelning simulerar vi värden från en kanonisk coxiansk fördelning med generator

$$T = \begin{pmatrix} 10 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -9 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -8 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -7 & 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -6 & 5.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -5 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -4 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 1.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

För det lägsta värdet på u gav POT-metoden en bättre skattning av γ i sex fall. För samma tröskel erhölls en svansapproximation F_1^* med ändligt stöd, och motsvarande skattning blev $\hat{\gamma}_1 = 0$. För de två högre trösklarna gav POT-metoden bättre skattningar i tre fall. För $u = 0.025$-kvantilen fick fem av svansapproximationerna F_1^* ändligt stöd, tre av dessa gav skattningar av γ som blev noll. Motsvarande resultat för $u = 0.01$-kvantilen blev åtta fall av ändligt stöd, av vilka fem gav nollskattningar. Resultatet ses i figur 9.

För tröskeln $u = 0.01$-kvantilen genomförde vi även skattningar med två coxianska fördelningar av ordning 4 och av ordning 2. Vi gjorde denna undersökningen för 10 nya stickprov med samma fördelning, men nu med en stickprovsstorlek på 10000 istället för 1000. Vår tes var att då data kommer från en kanonisk coxiansk fördelning så skulle fastypsmetoden med en coxiansk fördelning av ordning 2 prestera lika bra eller bättre än fastypsmetoden med en coxiansk fördelning av ordning 4. En intuitiv förklaring till detta är att för en tidig tidpunkt så är det inte säkert att den kanonisk coxianska fördelningen nått sin kvasistationära fördelning. Då är det troligt att sannolikhetsmassan är koncentrerad till de två sista tillstånden (istället för endast det sista tillståndet, vilket är den kvasistationära fördelningen).

Resultatet från simuleringarna blev att den coxianska fördelningen av ordning 4 gav en bättre skattning av γ i fem fall av tio. I tre fall blev den av ordning 2 bättre, och i resterande två fall var skattningarna lika bra. Det är ett otydligt resultat men är åtminstone ingen tydlig tendens för att en coxiansk svansapproximation av ordning 2 skulle vara sämre. För denna del av undersökningen visas inget spridningsdiagram.

3.3.5 T-fördelning med 6 frihetsgrader

För stickprov från en t-fördelning med 6 frihetsgrader gav fastypsmetoden bättre skattningar i en majoritet av fallen. För alla tre tröskelvärden hade hälften av svansapproximationerna F_1^* ändligt stöd. I majoriteten av tillhörande jämförelser var det fastypsmetoden som gav bäst skattning av $P(X > q_{0.001})$. Uppseendeväckande är att för inget av fallen då F_1^* fick ändligt stöd så var stödets övre punkt belägen nedanför $q_{0.001}$. Ingen av de skattningarna blev alltså noll så som de har blivit för andra fördelningar, vilket kan ses i figur 10.

3.3.6 Exponentialfördelning

Exponentialfördelningen är starkt förknippad med både fastypsfördelningar och den generaliserade paretofördelningen. I själva verket är den ett specialfall i båda dessa klasser av fördelningar.

För exponentialfördelningen visade sig fastypsmetoden prestera något bättre, framförallt för de två högre trösklarna. I majoriteten av alla fall resulterade parameteruppsättningen för F_1^* i ändligt stöd. I ett par av dessa fall ger POT-metoden skattningen $\hat{\gamma}_1 = 0$. Resultatet då data simulerades från en exponentialfördelning ses i figur 11.

4 Illustration av metoderna med nederbördsdata

Nedan demonstreras hur metoderna används på riktiga data. Vi vill undersöka vilka skattningar av kvantiler vi får för extrema nederbördsmängder. Låt Z vara den maximala dygnsnederbörden under en vecka och q tillhörande kvantil. Vi approximerar $q_{0.01} : P(Z > q_{0.01}) = 0.01$ med hjälp av en gp-fördelning och en coxiansk fördelning av ordning 4. Detta upprepas 25 gånger.

4.1 Beskrivning av datamängden

Datamängden kommer från SMHI[1] och består av ca 12 000 observationer av dygnsnederbörd. Den variabel vi är intresserade av är dygnsnederbörden för Jönköping. Data löper från 1965 till 1997.

Datavärdena anger mängden nederbörd från kl. 06 UTC till 06 UTC följande dag. Värdet -1.0 betyder 0 mm och värdet 0.0 betyder mindre än 0.1 mm.

4.2 Förbehandling av data

Båda metoderna förutsätter att data är oberoende. För att bli av med en del av beroendet som råder mellan på varandra följande observationer beräknas den maximala dygnsnederbörden för sju konsekutiva dygn, hädanefter kallat veckomax. Detta resulterar i 1713 observationer för veckomax. Eventuella trender över tid och variation mellan årstider förbises.

4.3 Tillvägagångssätt

Hädanefter antas att observationerna, dvs. veckomaxen, är oberoende. Datamängden delas slumpmässigt upp i 25 olika träningsmängder och testmängder med proportionerna 10/90. Detta för att kunna korsvalidera och få ett mått på hur bra de olika skattningarna är. Testmängden är stor för att kunna utvärdera hur bra kvantilskattningen blev. Observationerna i träningsmängden betecknas $y_i,\ i = 1, ..., n$, och i testmängden $z_j, j = 1, ..., m$, där n och m är antalet observationer i respektive mängd.

För varje uppdelning av data genomförs följande: Tröskeln sätts godtyckligt till 12.0 mm. Sannolikheten att överstiga värdet 12.0 skattas med den empiriska fördelningsskattningen $1 - \hat{F}_n(12) = \frac{\#\{y_{1i} > 12.0\}}{n}$. Värden över 12.0 behålls och translateras ner till 0 genom subtraktion med 12. Sedan skattas parametrarna i en gp-fördelning respektive en coxiansk fördelning av ordning 4 från träningsmängden.

Nu kan $q_{0.01}$ approximeras med $\hat{q}_{0.01} = F^{-1}(1 - \frac{0.01}{1-\hat{F}_n(12)})$, där F är den approximerande fördelningen för överskottet (den nyfikne kan se appendix för härledning). För att utvärdera

[1] http://www.smhi.se/klimatdata/meteorologi/dataserier-2.1102

hur bra denna är som skattning av $q_{0.01}$ skattas sannolikeheten för $P(Z > \hat{q}_{0.01})$ med den empiriska fördelningsskattningen från testmängden:

$$1 - \hat{F}_m(\hat{q}_{0.01}) = \frac{\#(z_j > \hat{q}_{0.01} + 12.0)}{m}.$$

4.4 Resultat

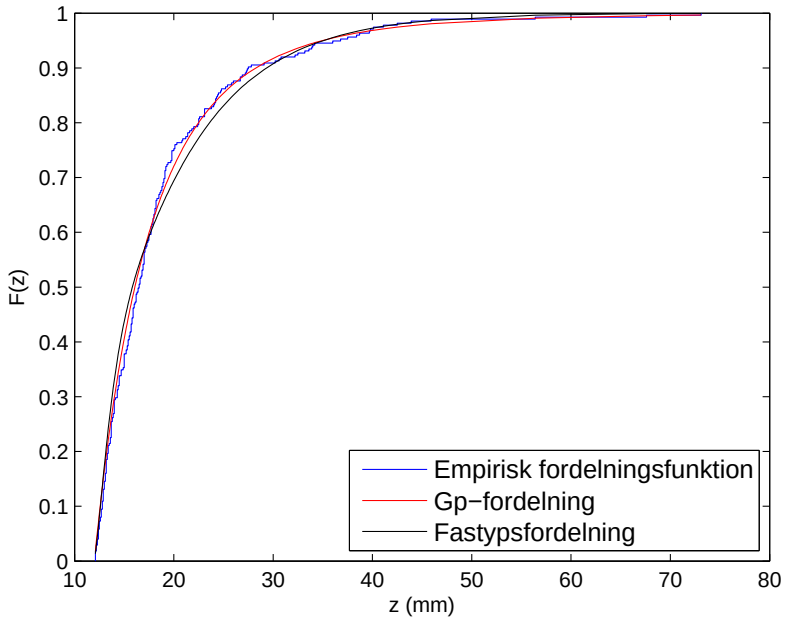
Både POT-metoden och fastypsmetoden skattar 0.01-kvantilen rätt bra. Resultaten visar inte på någon tydlig skillnad mellan metoderna (se tabell 1). Värt att nämnas är att POT-metoden fick ändligt stöd i 13 fall av 25. Att undersökningen på nederbördsdatan är liten samt att "facit" saknas gör att undersökningen blir svår att utvärdera. Den bör därför ej ses som en utvärdering av metoderna utan som en demonstration av hur metoderna kan användas.

	POT-metoden	Fastypsmetoden
Medelvärde	0.015279	0.014682
Standardavvikelse	0.007089	0.007356
Stickprovsstorlek	25	25

Tabell 1: Summering av hur bra de två metoderna är på att skatta 0.01-kvantilen, i de 25 skattningarna.

4.4.1 Resultat i en av de 25 skattningarna

Här redovisas närmare hur skattningarna från de två metoderna blir för en av de 25 uppdelningarna av data. Figur 12 visar hur väl de två metodernas skattade fördelningsfunktioner sammanfaller med den emperiska fördelningsfunktionen för testmängden, som ses så verkar båda approximationerna bli rätt bra.



Figur 12: De skattade fördelningsfunktionerna för överskottet samt den empiriska fördelningsfunktionen för en testmängd på 275 observationer.

Tabell 2 visar hur väl medelvärdet och variansen av överskottsvärdena i testmängden överensstämmer med de teoretiska motsvarigheterna för de båda approximerande fördelningarna.

	Testmängd	POT-metoden	Fastypsmetoden
Medelvärde	18.9425	18.9508	18.9035
Varians	69.4021	85.1044	63.2604

Tabell 2: Medelvärde och varians för överskottet av testmängden och de båda svansapproximationerna.

5 Slutsatser för undersökningen med simuleringar

Som vi ser i tabellerna 3 och 4 var metoderna ungefär lika bra för de två lägre värdena på u. För det högsta värdet på u så var fastypsmetoden överlag bättre för alla fördelningar. Tabell 3 visar dock att det för de flesta kombinationer av stickprovsfördelning och tröskelvärde var förhållandevis jämnt i antalet fall som respektive metod gav bäst skattning av γ.

Tröskel[1]	Normal	Lognormal	Pareto	t	Kanonisk cox	Exponential	Summa
95	4	6	8	4	6	4	32
97.5	6	7	5	4	3	2	27
99	2	3	4	4	3	2	18
Summa	10	16	17	12	12	8	

Tabell 3: Antalet vinster för POT-metoden. Vinst innebär att för givet stickprov av överskottsvärden så gav POT-metoden den närmsta skattningen. [1]Anger percentil för tröskeln.

Tröskel[1]	Normal	Lognormal	Pareto	t	Kanonisk cox	Exponential
95	fastyp	POT	POT	fastyp	POT	fastyp
97.5	POT	POT	lika	fastyp	fastyp	fastyp
99	fastyp	fastyp	fastyp	fastyp	fastyp	fastyp

Tabell 4: Den metod som för given fördelning och tröskel gav bäst skattning av γ i flest fall. [1]Anger percentil för tröskeln.

Resultaten är varken entydiga eller särskilt tydliga. Vi hade hoppats att vi skulle kunna se en korrelation mellan hur bra metoderna presterade och hur högt u var belägen, men vi har, förmodligen p.g.a. för små simuleringar, inte kunnat se något som tyder på sådana mönster.

Vad vi istället har sett är att när vi i POT-metoden maximum likelihood-skattar parametrarna till den generaliserade paretofördelningen, så är det påfallande ofta som vi får parameteruppsättningar som ger ändligt stöd för den approximerande fördelningen F_1^*. För vissa fördelningar är det vanligare, speciellt för normalfördelningen, exponentialfördelningen och t-fördelningen. Enda gången då vi får skattningar av $P(X > q_{0.001})$ som är 0, är då POT-metoden ger en approximation av svansen F_1^* som har ett ändligt stöd med övre punkten mindre än $q_{0.001}$. Det får betraktas som en otroligt dålig skattning. Det är förvisso inte ovanligt att vi överskattar $q_{0.001}$ med en faktor 2, eller ibland till och med 3, vilket alltså rent avståndsmässigt är en sämre skattning än 0. Men eftersom vi vill skatta sannolikheten för extrema händelser, händelser som i tillämpningar ofta är farliga eller otroligt kostsamma, får det ses som mycket sämre med en skattning på 0 än en skattning på 0.003.

Det är inte ovanligt att båda metoderna ger skattningar av $P(X > q_{0.001})$ som är flera gånger högre än 0.001. Det gör att vi får varianser som inte skiljer sig märkbart åt mellan metoderna. Tendensen hos POT-metoden att ge skattningen noll syns alltså inte i variansen

på grund av det faktum att båda metoderna ger stora positiva felskattningar.

5.1 Ändligt stöd i generaliserade paretofördelningen

I flera fall ger POT-metoden en approximation av svansen som har ändligt stöd. Som redan nämnts är detta ett problem eftersom vårt uttalade mål är att skatta sannolikheter för extrema värden. Vi vill alltså inte få en övre gräns för stödet på våra approximerande fördelningar. Resultatet från undersökningen ger anledning att tro att vissa av fördelningarna är mer känsliga för det här än andra. En naturlig fråga är alltså om det finns kvalitéer hos fördelningarna som gör dem mer utsatta för ändligt stöd vid skattning av parametrarna i gp-fördelningen.

Alla fördelningar i vår undersökning har omskalade överskottsfördelningar som konvergerar i fördelning mot betingningsstabila fördelningar, dvs. gp-fördelningar. Alltså konvergerar felintensiteten för dessa omskalade överskottsfördelningar mot felintensiteter hos gp-fördelningarna. Men i familjen av generaliserade paretofördelningar kan endast tre kvalitativa beteenden hos felintensiteterna förekomma: konstant eller avtagande eller växande. De två första svarar mot gp-fördelningar med oändligt stöd. Den sista svarar mot en gp-fördelning med ändligt stöd. Detta inses genom att beräkna felintensiteten till gp-fördelningen:

Låt $X \sim gp(\mu, \sigma, \beta)$. Då får X felintensitet

$$\begin{aligned} h(x) = \frac{f(x)}{1-F(x)} &= \frac{\frac{1}{\sigma}(1+\frac{\beta(x-\mu)}{\sigma})^{(-1/\beta-1)}}{(1+\frac{\beta(x-\mu)}{\sigma})^{-1/\beta}} \\ &= \frac{1}{\sigma}(1+\frac{\beta(x-\mu)}{\sigma})^{-1} = [z = \frac{x-\mu}{\sigma}] \\ &= \frac{1}{\sigma(1+\beta z)}. \end{aligned}$$

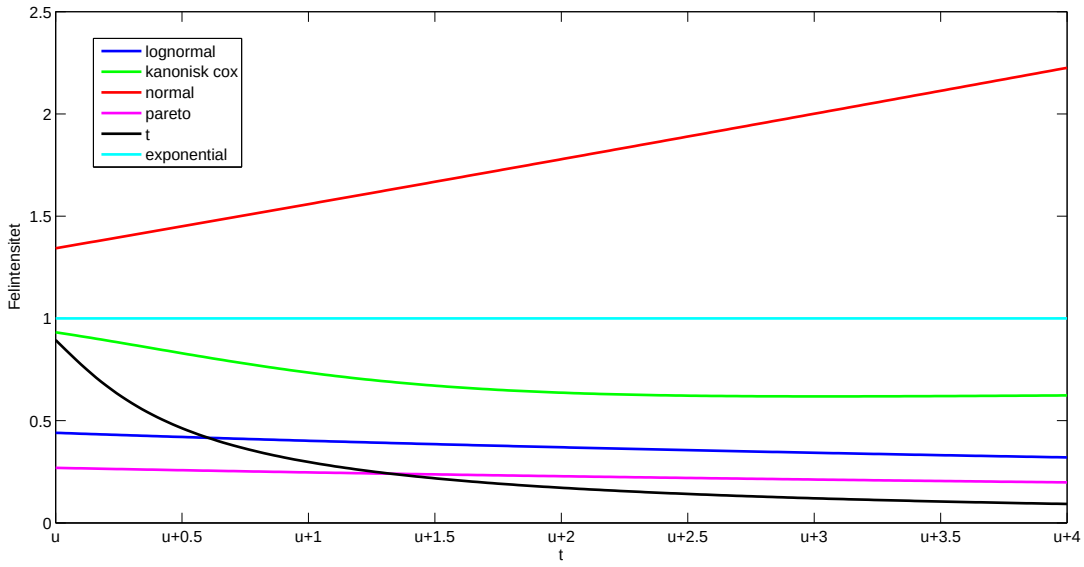
Derivering ger $h'(z) = -\frac{\beta}{\sigma(1+\beta z)^2}$. Gp-fördelningen har således ökande felintensitet då $\beta < 0$, konstant felintensitet då $\beta = 0$, och avtagande då $\beta > 0$. Enligt definition (2.12) av gp-fördelning stämmer alltså påståendet.

I figur 13 ser vi felintensiteterna till de omskalade överskottsfördelningarna över 0.05-kvantilen. I endast ett fall fås en ökande felintensitet, nämligen hos normalfördelningen. I fallet med exponentialfördelningen är felintensiteten konstant. Felintensiteten för den omskalade kanonisk coxianska fördelningen avtar mot en konstant. De resterande tre fördelningarna; lognormal-, t- och paretofördelningen, har avtagande felintensiteter. Fördelningarna med en avtagande felintensitet för den omskalade överskottsfördelningen tenderar att ha färre fall av svansapproximationer F_1^* med ändligt stöd, se tabell 5.

Asymptotiskt har alla fördelningar i vår undersökning med undantag för paretofördelningen omskalade överskottsfördelningar med felintensiteter som går mot en konstant. Paretofördelningen är i sig en gp-fördelning och behåller därför sin avtagande felintensitet även efter omskalning och betingning.

Tröskel[1]	Normal	Lognormal	Pareto	t	Kanonisk cox	Exponential	Summa
95	8	0	0	5	1	6	20
97.5	9	4	1	5	5	6	30
99	7	5	4	5	8	10	39
Summa	24	9	5	15	14	22	

Tabell 5: Antalet fall av svansapproximationer F_1^* med ändligt stöd i undersökningen per fördelning och tröskel. [1]Anger percentil för tröskeln.



Figur 13: Felintensiteter för de omskalade överskottsfördelningarna. Observera att u är respektive fördelnings 95-percentil, varför $[u, u+4]$ är olika intervall för varje fördelning.

Konvergensen är i regel långsam och för vissa fördelningar och parameteruppsättningar är den mycket långsam, se figur 5 i avsnitt 2.4.5. POT-metoden är som många andra viktiga metoder i statistik ett resultat av konvergens i fördelning. I verkligheten är man hänvisad till approximation, som kan lida kraftigt av den långsamma konvergensen. För normalfördelningen kommer t.ex. den omskalade intensiteten över 0.05-kvantilen att vara tydligt ökande. Det är av den anledningen inte konstigt att maximum likelihood-skattningarna ger en gp-fördelning med ökande felintensitet och därmed ett ändligt stöd. I tabell 5 ser vi att normalfördelningen hade absolut flest förekomster av svansskattningar F_1^* med ändligt stöd. Strax därefter kommer exponentialfördelningen. Lognormal- och paretofördelningarna, som båda har avtagande felintensitet för den omskalade överskottsfördelningen, har minst antal av svansapproximationer F_1^* med ändligt stöd. T-fördelningen, som också har avtagande felintensitet, sticker ut eftersom den, trots avtagande felintensitet, har ganska många fall av ändligt stöd för F_1^*.

Exponentialfördelningen har en omskalad överskottsfördelning med konstant felintensitet. En undersökning av de maximum likelihood-skattade parametrarna till F_1^* för fördelningen visar att formparametern oftast har en skattning som ligger nära noll för den lägsta tröskeln. Oftast kommer då första nollskilda siffran i andra decimalen. Dessa skattningar ligger alltså nära det asymptotiska resultatet för överskottsfördelningarna, då formparametern är noll. Att skattningen av formparametern aldrig resulterar i exakt noll beror förmodligen på slumpen vi är utsatta för i och med stickproven. En annan förklaring skulle kunna vara MATLABs numeriska skattningsalgoritm. Även en liten avvikelse nedåt från nollan gör att stödet för F_1^* blir ändligt. För de högre tröskelvärdena tenderar formparameterns skattningar att ligga längre från noll.

De fördelningar som har omskalade överkottsfördelningar över 0.05-kvantilen med konstant eller växande felintensitet får alltså oftare svansapproximationer F_1^* med ändligt stöd. Den fördelning som sticker ut är t-fördelningen, som får något fler fall med svansapproximation med ändligt stöd, trots en omskalad överskottsfördelning som har en avtagande felintensitet. Möjligen kan detta förklaras med vår undersöknings begränsade underlag.

6 Diskussion

I kapitlen ovan har grunden för extremvärdesteori och teorin bakom fastypsfördelningar presenterats. Båda områdena har resultat som kan användes för att motivera approximationer av svansarna över en tröskel hos fördelningar. I ett senare kapitel har en empirisk undersökning gjorts i syfte att utvärdera hur bra gp-fördelningar respektive fastypsfördelningar av den coxianska klassen är på att approximera svansar över en tröskel. I båda metoderna har värden över tröskeln använts för att skatta parametrar hos fördelningarna. Den ursprungliga tesen var att för en låg tröskel är konvergensen mot gp-fördelningen ännu inte uppnådd. Då borde approximationsegenskaperna hos de coxianska fördelningarna göra dessa mer lämpliga för att skatta sannolikheten för extrema värden. Som redan påpekats var resultatet inte helt i linje med vår tes och inte heller entydigt på något vis. Detta kan dock ha sin förklaring i undersökningens begränsade omfattning, framförallt gällande antalet simuleringar.

För att på ett mer detaljerat sätt undersöka om hypoteserna håller kan en fortsättning vara att genomföra en större undersökning än den som redogjorts för i föregående kapitel. Ett mer naturligt sätt att planera undersökningen hade varit att basera undersökningen på samma sorts simuleringar, men istället för att para och jämföra avstånden till 0.001 fokusera på skillnader i medelvärden och standardfel mellan metoderna. För att detta ska vara genomförbart behöver man lösa det problem vi upplevt med implementeringen av programvaran EMpht, så att man kan anropa den godtyckligt många gånger i en körning. På så vis hade man fått ett betydligt större stickprov av skattningar att jobba med. Mer lättillgängliga statistiska metoder hade då kunnat användas för att utvärdera resultatet.

Som påpekats tidigare är en nackdel med EMpht-algoritmen att de resulterande skattningarna inte nödvändigtvis ger ett globalt maximum för likelihoodfunktionen, då den kan fastna i sadelpunkter eller lokala maxima. För att få bukt med detta hade man behövt strukturera körningarna av EMpht så att varje skattning av parametrarna i en coxiansk fördelning hade gjorts för många olika startfröer. Genom en sådan metod hade man haft bättre kontroll över när skattningarna verkligen ger en maximumpunkt, t.ex. hade en stor skillnad i skattningarna mellan olika fröer indikerat att skattningarna i flera av fallen troligen inte ger en maximumpunkt. För att göra detta hade även en studie av EM-algoritmen i allmänhet och tillämpningen för fastypsfördelningar i synnerhet varit nödvändig.

I arbetet med att utvärdera hur bra en coxiansk fördelning är på att skatta olika fördelningars svansarna över en tröskel har en ordning på fyra genomgångstillstånd valts. Valet var godtyckligt. I linje med tidigare förslag på förbättringar av undersökningen hade en närmare studie av de coxianska fördelningarna kunnat ge en bättre grund för valet av ordning hos den approximerande fördelningen.

Som visats i kapitlet om undersökningen ger de approximerande gp-fördelningarna ibland skattningen 0 av $P(X > q_{0.001})$. Det ligger i teorins natur att denna skattningen är en sämre skattning än, säg, 0.002. För att justera undersökningen hade en transformation av svanssannolikheter på lämpligt sett kunnat göra nollskattningar mer "kostsamma" än andra felskattningar.

Bland gp-fördelningarna finns inte någon fördelning som har en växande felintensitet i kombination med oändligt stöd. Som vi har sett ger normalfördelningen över en hög tröskel u en felintensitet som är linjärt växande. Detta trots att felintensiteten går mot en konstant då $u \to \infty$. Samtidigt ger de maximum likelihood-skattade parametrarna i en gp-fördelning ett ändligt stöd. Detta belyser varför det vore bra att ha en klass av fördelningar som även innehåller växande felintensiteter samtidigt som oändligt stöd.

Man skulle kunna tänka sig en ändamålsenlig modell som innehåller
i) klassen av gp-fördelningar,
ii) alla fördelningar med linjärt växande felintensitet,
iii) en fastypsfördelning av coxiansk typ av ordning k, för k fixt.

En sådan modell verkar efter denna undersökning lämpad att använda för att skatta svansen hos fördelningar, med målet att skatta sannolikheter för extrema värden. Det vore intressant att se en modell som innehåller åtminstone ett par av ovanstående klasser av fördelningar.

Ett mer övergripande problem när det gäller skattningen av sannolikheten för extrema värden är att "facit" saknas. Det är ett återkommande problem inom statistik, men kan anses vara olika allvarligt i olika sammanhang. I jämförelse med till exempel centrala gränsvärdessatsen, där resultaten gäller för alla fördelningar, är extremvärdesteorins resultat mindre generella. Alla fördelningar har inte en överskottsfördelning som konvergerar mot en gp-fördelning. Det är problematiskt eftersom man i verkligheten sällan har data som man vet kommer från en specifik teoretisk fördelning. Ett annat problem är hastigheten hos konvergensen, som varierar för olika fördelningar och parametrar. Hur rimligt antagandet om konvergens är beror därför i stor utsträckning på ursprungsfördelningen.

Dessa problem är viktiga att ha i åtanke vid användning av extremvärdesteori, som alltid bör användas med försiktighet.

A Appendix

A.1 Entydigheten hos den kvasistationära fördelningen $\boldsymbol{\nu} = (0, 0, ..., 1)$ för en kanonisk coxiansk fördelning

Låt Z vara en stokastisk variabel med kanonisk coxiansk fördelning. Antag att det utöver $\boldsymbol{\nu} = (0, 0, ..., 1)$ finns minst en till kvasistationärfördelning $(\nu_1, \nu_2, ..., \nu_p)$, där $\nu_i = P(Z_t = i | Z_t \neq 0) = P(Z_{t+u} = i | Z_{t+u} \neq 0)$ för $i = 1, ...p$.

Då gäller att

$$\nu_1 = \frac{\nu_1(e^{-\lambda_{11}t})}{e^{-(\nu_1\theta_1+...+\nu_p\theta_p)t}} \tag{8}$$

Nämnaren är sannolikheten att inte ha absorberats vid tidpunkt t. Täljaren är sannolikheten att hamna i tillstånd ett och att stanna där.

(8) ger

$$e^{-\lambda_{11}t} = e^{-(\nu_1\theta_1+...+\nu_p\theta_p)t}$$

Så

$$\lambda_{11} = \nu_1\theta_1 + ... + \nu_p\theta_p \tag{9}$$

Men $\lambda_{ii} > \theta_i$ och för en kanonisk coxiansk fördelning gäller $\lambda_{11} \geq \lambda_{22} \geq ... \geq \lambda_{pp}$, vilket ger att $\lambda_1 > \theta_1, ..., \theta_p$. Dessutom gäller att $\sum_{n=1}^{n} \nu_i = 1$. Så likheten (9) ger en motsägelse. Alltså kan den kvasistationära fördelningen inte ha någon sannolikhetsmassa i första tillståndet. Att detsamma gäller för tillstånd $2, ..., p-1$ visas med hjälp av induktion.

A.2 Mer detaljerade beräkningar för hur ett uttryck fås fram för felintensiteten $h(t)$:

Om $h'(0) = 0$ fås

$$0 = h'(0) = \frac{h'(t)}{h(t)^2} \implies h'(t) = 0 \implies h(t) = c$$

Eftersom $h(0) = 1$ fås därför att $h(t) = 1$ för alla t.

Om $h'(0) \neq 0$ innebär det att

$$h'(0) = \frac{h'(t)}{h(t)^2} = \frac{d}{dt}\frac{1}{h(t)}$$

Vi får att $1/h(t) = 1 + At$ för en konstant A ty

$$A = h'(0) = d/dt(1/h(t))$$

Och

$$\int Adt = \int \frac{d}{dt}\frac{1}{h(t)}dt$$

Så

$$At + d = 1/h(t)$$

Där $h(0) = 1$ ger

$$1/h(t) = At + 1$$

Vilket ger att

$$h(t) = \frac{1}{1 + At}$$

A.3 Härledning av skattningen $\hat{q}_{0.01}$

Låt Z vara veckomax med fördelningfunktionen F_z, X vara överskottet, dvs. $X = (Z|Z > 12)$, med fördelningsfunktionen F_x och F vara den approximerande funktionen till F_x.

Definitionen av betingad sannolikhet ger:

$P(Z < q_{0.01}|Z > 12) = \frac{P(12<Z<q_{0.01})}{P(Z>12)} = \frac{F_z(q_{0.01})-F_z(12)}{1-F_z(12)} = \frac{1-F_z(12)-0.01}{1-F_z(12)} = 1 - \frac{0.01}{1-F_z(12)}$

Vi får:

$q_{0.01} = F_x^{-1}(1 - \frac{0.01}{1-F_z(12)})$ (*)

Om vi nu i (*) ersätter F_x med F och $F_z(12)$ med den empiriska fördelningsskattningen $\hat{F}_n(12)$ så fås:

$q_{0.01} \approx F^{-1}(1 - \frac{0.01}{1-\hat{F}_n(12)})$

A.4 Programkod

Efter samråd med handledaren har vi kommit fram till att inte lägga in programkoden här eftersom den skulle ta mycket plats och vår bedömning är att den inte skulle tillföra så mycket.

Referenser

O. O. Aalen, Ö. Borgan, H.K. Gjessing (2008), 'Survival and Event History Analysis', *Springer*, New York.

S. Asmussen, O. Nerman, M. Olsson (1996), 'Fitting phase type distributions via the EM algorithm', *Scandinavian Journal of Statistics*, 23, 419-441.

M. Bladt (2005), 'A review on phase-type distributions and their use in risk theory', *Astin Bulletin*, Vol. 35, No. 1, 145-161.

E. Castillo (1988), 'Extreme Value Theory in Engineering', *Academic Press*, San Diego.

S. Coles (2001), 'An Introduction to Statistical Modeling of Extreme Values', *Springer*, London.

A. Cumani (1982), 'On the canonical represantation of homogeneous Markov processes modelling failure-time distributions', *Microelectron Reliab*, 22:583-602.

J. Enger, J. Grandell(2003), 'Markovprocesser och köteori', Stockholm.

M. Fackrell (2009),'Modeling healthcare systems with phase-type distributions', *Health Care Management Science*, 12:11-26.

M. Falk, J. Hüsler, R-D. Reiss (1994), 'Laws of Small Numbers: Extremes and Rare Events', *Birkhäsuser Verlag*, Basel.

M. Johnson, M. Taaffe (1988), 'The denseness of phase distributions', *Research Memorandum*, No.88-20.

CA. O'Cinneide (1989), 'On non-uniqueness of representations of phase-type distributions', *Commun Stat Stoch Models*, 5:247-259.

M. Olsson (1995), 'EM Estimation in Phase Type Models', doktorsavhandling vid Matematiska Vetenskaper, Chalmers/GU.

M. Olsson (1996), 'Estimation of phase type distributions from censored data', *Scandinavian Journal of Statistics*, 23, 443-460.

L. Råde, B. Westergren (2008), 'Beta - Mathematics Handbook', 5:e utgåvan, Studentlitteratur, Lund, s. 467.

D. Zholud (2011), 'Extreme Value Analysis of Huge Datasets', doktorsavhandling vid Matematiska Vetenskaper, Chalmers/GU.