Taut String methoden voor Current Status

Sander Boersma

23 augustus 2011

Inhoudsopgave

1 Inleiding

In de statistiek bestudeert men vaak een bepaalde populatie, bijvoorbeeld een partij gloeilampen. Een vraag die men zich dan kan stellen, is: Welk deel van deze partij gloeilampen gaat langer dan 10 jaar mee? Algemener is men geïnteresseerd in de verdelingsfunctie van de levensduur van een gloeilamp. De data die men nodig heeft om iets over deze verdelingsfunctie te kunnen zeggen, wordt verkregen door een steekproef te nemen uit de populatie. In het voorbeeld van de gloeilampen, zal nagegaan worden hoe lang een aantal gloeilampen zijn meegegaan. Op basis van deze data kan de verdelingsfunctie geschat worden.
In de ideale situatie zijn die data beschikbaar. We spreken dan van volledige data. In dat geval zijn er meerdere schattingsmethoden om een schatter te vinden voor de gezochte verdelingsfunctie. Het is echter mogelijk dat die data niet volledig beschikbaar zijn, met andere woorden: de data zijn gecensureerd. Vaak zijn er dan wel andere data beschikbaar. Zulke data worden ook wel onvolledige data genoemd. De data zijn bijvoorbeeld gecensureerd bij het schatten van een verdelingsfunctie van de leeftijd waarop je een bepaalde ziekte krijgt. Het exacte tijdstip waarop iemand de ziekte heeft gekregen, is namelijk vaak moeilijk of zelfs niet vast te stellen.
Er zijn verschillende soorten censurering. Bij rechts-censurering is de exacte waarde slechts tot een bepaalde grenswaarde bekend. Als de waarde groter is dan de grenswaarde, weten we wel dat de waarde groter is, maar de exacte waarde is gecensureerd. Bij links-censurering is het andersom: De exacte waarde is nu alleen vanaf een grenswaarde bekend. Is de waarde kleiner dan die grenswaarde, dan weten we wel dat de waarde kleiner is, de exacte waarde is echter gecensureerd. Verder is er bij interval-censurering alleen bekend in welk interval de waarde ligt. De exacte waarde weten we niet.
Wij gaan ons bezig houden met het Current Status model. Bij dit model, dat vaak wordt toegepast in de medische statistiek, zijn we geïnteresseerd in de verdelingsfunctie van de leeftijd waarop iemand een bepaalde ziekte krijgt. De data is dan dus gecensureerd. We beperken ons hier tot interval-censurering met twee intervallen. Dat wil zeggen, het is alleen bekend of een waarde links of rechts van een grenswaarde ligt. Met behulp van deze onvolledige data is het toch mogelijk om een schatting te geven voor de verdelingsfunctie waarin we geïnteresseerd zijn. Voor het schatten van zo'n verdelingsfunctie zijn er verschillende methoden. Een vrij bekende methode is de Maximum Likelihood methode. Een iets minder bekende en nieuwere methode is de Taut String methode. Dit is een niet-parametrische methode, vooral bekend geworden door Davies en Kovac (zie [1] en [2]).
We beginnen met een korte uitleg over het Current Status censureringsmodel. Vervolgens bekijken we een aantal schattingsmethoden die reeds zijn toegepast op dit model. De methoden die we behandelen, zijn de parametrische en de niet-parametrische Maximum Likelihood methoden. We leggen van deze twee methoden eerst het principe uit, daarna bekijken we de methoden in de context van het Current Status model. In hoofdstuk 4 behandelen we de Taut String methode in het algemeen, waarbij we bovendien zowel een algoritme als een voorbeeld geven. Als laatste passen we deze Taut String methode toe op verdelingsfuncties. Eerst met volledige data, daarna binnen het Current Status model. Dit gebeurt in hoofdstuk 5.

2 Het Current Status model

In dit hoofdstuk leggen we uit wat het Current Status model inhoudt. Hierbij wordt ingegaan op de onvolledige data die in dit model van belang zijn. Vervolgens wordt een wiskundige beschrijving van deze beschikbare data gegeven. We gaan er daarbij dus vanuit dat we geïnteresseerd zijn in de verdelingsfunctie van de leeftijd waarop iemand een bepaalde ziekte krijgt.

2.1 Censureringsmodel

Omdat de data uit de verdelingsfunctie gecensureerd zijn, kunnen we daar geen gebruik van maken. Er is echter wel andere data beschikbaar. Ondanks dat we niet weten op welke leeftijd een persoon de ziekte heeft gekregen, kan namelijk wel vastgesteld worden of de persoon de ziekte op een zekere leeftijd al heeft gehad of niet. Dit kan bijvoorbeeld door na te gaan of de persoon al antilichamen tegen die ziekte in zijn bloed heeft. Deze data is niet volledig: Als de persoon de ziekte al heeft gehad, weten we niet precies wanneer. Dit kan op elk willekeurig moment vóór de controle gebeurd zijn. Als de persoon de ziekte nog niet heeft gehad, weten we niet wanneer de persoon de ziekte zal krijgen. Bovendien is het mogelijk dat de persoon de ziekte helemaal niet krijgt. Merk op dat dit interval censurering is, waarbij de grenswaarde de leeftijd van onderzoek is. Toch kan aan de hand van deze data een schatting worden gegeven voor de verdelingsfunctie waar we in geïnteresseerd zijn. Daarbij is het wel van belang dat we data hebben van personen van verschillende leeftijden.
Per persoon zijn nu dus twee gegevens van belang: de leeftijd waarop de persoon is onderzocht, en het gegeven of de persoon op dat moment de ziekte al heeft gehad of niet.

2.2 Wiskundige beschrijving

Er zijn twee leeftijden van belang. Laat X_i de leeftijd zijn waarop persoon i de ziekte krijgt, en T_i de leeftijd waarop persoon i is onderzocht. Deze leeftijden zijn per persoon onafhankelijk van elkaar, oftewel X_i en T_i zijn onafhankelijk. Als persoon i de ziekte al heeft gehad, dan is de leeftijd waarop hij de ziekte heeft gekregen lager dan de leeftijd waarop hij is onderzocht, oftewel $X_i \leq T_i$. Als persoon i op leeftijd T_i de ziekte nog niet heeft gehad, geldt uiteraard $X_i > T_i$. Dit gegeven verwerken we in Δ_i:

$$\Delta_i = \begin{cases} 1, & \text{als } X_i \leq T_i, \\ \\ 0, & \text{als } X_i > T_i. \end{cases}$$

Als de persoon de ziekte op leeftijd T_i heeft gehad, geldt $X_i \leq T_i$ en dus $\Delta_i = 1$. Heeft de persoon de ziekte nog niet gehad, dan geldt $X_i > T_i$ en dus $\Delta_i = 0$. We kunnen Δ_i dus zien als de huidige toestand (Engels: "current status") van persoon i op leeftijd T_i. Vandaar dat het model het Current Status model heet.
Merk op dat de leeftijd X_i onbekend is (want de volledige data is gecensureerd). Door te controleren of de persoon al antilichamen tegen de ziekte in zijn bloed heeft of niet, weten we wel dat $X_i \leq T_i$ respectievelijk $X_i > T_i$. Δ_i wordt dus eigenlijk als volgt bepaald:

$$\Delta_i = \begin{cases} 1, & \text{als de persoon wel antilichamen in zijn bloed heeft,} \\ \\ 0, & \text{als de persoon geen antilichamen in zijn bloed heeft.} \end{cases}$$

Stel dat van n personen is onderzocht of ze de ziekte al gehad hebben of niet. Dan observeren we de data-paren $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$, waarbij T_i het tijdstip is waarop persoon i is onderzocht, en Δ_i de huidige toestand van persoon i op tijdstip T_i. Omdat X_i en T_i onafhankelijk zijn, zijn ook de paren (T_i, Δ_i) en (T_j, Δ_j) met $i \neq j$ onafhankelijk.
Met deze data-paren kunnen we een schatting geven voor de verdeling van X.

3 Maximum Likelihood methoden

Op basis van de hiervoor beschreven data-paren, kan binnen het Current Status model de onbekende verdelingsfunctie op verschillende manieren worden geschat. Hieronder worden twee veel gebruikte methoden behandeld. We beginnen met de parametrische Maximum Likelihood methode. Vervolgens behandelen we de niet-parametrische Maximum Likelihood methode. Bij beide methoden wordt eerst het principe, aan de hand van ongecensureerde data, uitgelegd. Daarna wordt die methode toegepast op het Current Status model.

3.1 Parametrische Maximum Likelihood methode

Een van de bekendste schattings-methoden is de parametrische Maximum Likelihood methode. Het principe leggen we uit aan de hand van volledige data. We nemen nu dus aan dat we een realisatie $x_1, \ldots, x_n$ van n onafhankelijke en identiek verdeelde stochastische variabelen $X_1, \ldots, X_n$ uit de onbekende verdeling hebben. Vervolgens kiezen we een parametrisch model voor X_i, dat wil zeggen: We kiezen een klasse van verdelingen waarvan we veronderstellen dat de echte verdeling daartoe behoort. Als het bekend is uit welke klasse de verdeling komt, is dit geen probleem. Weten we dat echter niet, dan moeten we een keuze doen, en kunnen we achteraf (met behulp van een niet-parametrische methode) controleren of die keuze een goede schatter oplevert of niet. Beschouw de gekozen klasse met de verzameling kansdichtheids-functies:

$$\mathscr{F} = \{f(\cdot|\theta_1, \ldots, \theta_k) : \ (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k\}$$

De kansdichtheids-functie $f(x|\theta_1, \ldots, \theta_k)$, met onbekende parameters $\theta_1, \ldots, \theta_k$, is nu dus bekend. Deze parameters gaan we vervolgens schatten. Dit doen we door de Likelihood-functie te maximaliseren:

$$L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = f(x_1, \ldots, x_n|\theta_1, \ldots, \theta_k) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k)$$

De laatste gelijkheid geldt, omdat volgens de aanname $x_1, \ldots, x_n$ realisaties zijn van de onafhankelijke en identiek verdeelde stochastische variabelen $X_1, \ldots, X_n$.
De functie L is gedefinieerd als product. Omdat de logaritme een strikt stijgende functie is, heeft de functie $l = \log(L)$ zijn maximum in hetzelfde punt als L. We kunnen dus ook in plaats van L de Log-Likelihood-functie l maximaliseren:

$$\begin{aligned} l(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) &= \log(L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n)) \\ &= \log(f(x_1, \ldots, x_n|\theta_1, \ldots, \theta_k)) \\ &= \log\left(\prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k)\right) \\ &= \sum_{i=1}^{n} \log(f(x_i|\theta_1, \ldots, \theta_k)) \end{aligned}$$

Vaak is het gemakkelijker om l te maximaliseren dan L.
Ter illustratie van de parametrische Maximum Likelihood methode nemen we nu aan dat de verdeling exponentieel is. Bij andere verdelingen (zoals bijvoorbeeld de normale verdeling) werkt de methode analoog.
We bekijken nu dus de verzameling:

$$\mathscr{F} = \{f(\cdot|\theta) : \ \theta > 0, \ f(x|\theta) = \theta \mathrm{e}^{-\theta x}, \ x \geq 0\}$$

We vullen nu dus $f(x|\theta) = \theta \mathrm{e}^{-\theta x}$ in de Log-Likelihood-functie in:

$$\begin{aligned} l(\theta|x_1, \ldots, x_n) &= \sum_{i=1}^{n} \log(f(x_i|\theta)) \\ &= \sum_{i=1}^{n} \log(\theta \mathrm{e}^{-\theta x_i}) \\ &= \sum_{i=1}^{n} (\log(\theta) - \theta x_i) \\ &= n \log(\theta) - \theta \sum_{i=1}^{n} x_i \end{aligned}$$

Het maximaliseren van deze functie, door de vergelijking $\frac{\partial}{\partial \theta} l(\theta) = 0$ op te lossen naar θ, levert:

$$\frac{\partial}{\partial \theta} l(\theta) = \frac{\partial}{\partial \theta} \left(n \log(\theta) - \theta \sum_{i=1}^{n} x_i \right) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

Oftewel:

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i}$$

Merk op dat $\frac{\partial}{\partial \theta} l(\theta) > 0$ als $\theta < \hat{\theta}$, en $\frac{\partial}{\partial \theta} l(\theta) < 0$ als $\theta > \hat{\theta}$. De functie l heeft dus inderdaad een maximum in het punt $\hat{\theta}$. Als we de gezochte kansdichtheids-functie willen schatten met een exponentiële verdeling, dan kunnen we als parameter de gevonden $\hat{\theta}$ nemen.
Wanneer blijkt dat deze exponentiële verdeling geen goede schatting geeft, kan met een andere klasse van verdelingen geprobeerd worden om een betere schatter te vinden. Het is sowieso nuttig om te controleren of de gevonden exponentiële verdeling redelijk is. Dit kan door de verdelingsfunctie te vergelijken met een niet-parametrische schatter zoals bijvoorbeeld de empirische verdelingsfunctie (zie paragraaf 3.2).

De parametrische Maximum Likelihood methode kan ook worden toegepast binnen het Current Status model.
De onvolledige data die dan beschikbaar zijn, zijn de data-paren $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$ (zie paragraaf 2.2). Stel dat we voor de T_i-waarden (de leeftijden van de personen op het moment dat ze onderzocht werden) de gerealiseerde waarden $t_1 < t_2 < \cdots < t_n$ hebben. Dan weten we dat geldt:

$$P(\Delta_i = \delta_i) = P(X_i \leq t_i)^{\delta_i} P(X_i > t_i)^{1-\delta_i}, \text{ met } \delta_i \in \{0, 1\},$$

want dan geldt namelijk: $P(\Delta_i = 0) = P(X_i > t_i)$ en $P(\Delta_i = 1) = P(X_i \leq t_i)$.
Als we aannemen dat F de verdelingsfunctie van X is, weten we bovendien dat $P_F(X_i \leq t_i) = F(t_i)$. Oftewel:

$$P_F(\Delta_i = \delta_i) = F(t_i)^{\delta_i} (1 - F(t_i))^{1-\delta_i}$$

Dan geldt dus ook:

$$\log(P_F(\Delta_i = \delta_i)) = \delta_i \log(F(t_i)) + (1 - \delta_i) \log(1 - F(t_i))$$

We kunnen nu, gegeven dat $T_i = t_i$ voor $i = 1, 2, \ldots, n$, de Likelihood-functie opstellen:

$$L(F) = P_F(\Delta_1 = \delta_1, \Delta_2 = \delta_2, \ldots, \Delta_n = \delta_n) = \prod_{i=1}^{n} P_F(\Delta_i = \delta_i),$$

want X_i en X_j met $i \neq j$ zijn onafhankelijk, en dus zijn ook Δ_i en Δ_j met $i \neq j$ onafhankelijk. Dan is de Log-Likelihood-functie:

$$\begin{aligned} l(F) &= \log(L(F)) = \log\left(\prod_{i=1}^{n} P_F(\Delta_i = \delta_i)\right) = \sum_{i=1}^{n} \log(P_F(\Delta_i = \delta_i)) \\ &= \sum_{i=1}^{n} \left(\delta_i \log(F(t_i)) + (1 - \delta_i) \log(1 - F(t_i))\right) \end{aligned}$$

Als we wederom aannemen dat X_i exponentieel verdeeld is, dan geldt dus op $(0, \infty)$: $F_\theta(x) = 1 - \mathrm{e}^{-\theta x}$, voor een zekere $\theta > 0$.
De Log-Likelihood-functie wordt dan:

$$\begin{aligned} l(F_\theta) &= \sum_{i=1}^{n} \left(\delta_i \log(1 - \mathrm{e}^{-\theta t_i}) + (1 - \delta_i) \log(1 - (1 - \mathrm{e}^{-\theta t_i}))\right) \\ &= \sum_{i=1}^{n} \left(\delta_i \log(1 - \mathrm{e}^{-\theta t_i}) - (1 - \delta_i)\theta t_i\right) \end{aligned}$$

Om deze functie te maximaliseren, willen we de vergelijking $\frac{d}{d\theta} l(F_\theta) = 0$ oplossen naar θ. Er geldt:

$$\begin{aligned} \frac{\partial}{\partial\theta} l(F_\theta) &= \sum_{i=1}^{n} \left(\frac{\delta_i(-\mathrm{e}^{-\theta t_i})(-t_i)}{1 - \mathrm{e}^{-\theta t_i}} - t_i + \delta_i t_i\right) \\ &= \sum_{i=1}^{n} \left(\frac{\delta_i t_i \mathrm{e}^{-\theta t_i}}{1 - \mathrm{e}^{-\theta t_i}} + \frac{\delta_i t_i (1 - \mathrm{e}^{-\theta t_i})}{1 - \mathrm{e}^{-\theta t_i}} - t_i\right) \\ &= \sum_{i=1}^{n} \left(\frac{\delta_i t_i}{1 - \mathrm{e}^{-\theta t_i}} - t_i\right) \end{aligned}$$

We kunnen de vergelijking $\frac{d}{d\theta} l(F_\theta) = 0$ niet analytisch oplossen. Met behulp van de methode van Newton kunnen we echter wel een numerieke oplossing vinden.

3.1.1 Voorbeeld

We genereren nu zelf Current Status data. Met andere woorden, er worden 100 trekkingen $x_1, \ldots, x_{100}$ voor X gedaan, en 100 trekkingen $t_1, \ldots, t_{100}$ voor T. Hierbij worden de trekkingen voor X gedaan uit een exponentiële verdeling met parameter $\theta_X = 1$, voor T uit een exponentiële

verdeling met parameter $\theta_T = 2$. Op basis van X en T kan Δ bepaald worden, en dus weten we dan ook $\delta_1, \ldots, \delta_n$ (herinner: $\delta_i = 1$ als $x_i \leq t_i$ en $\delta_i = 0$ als $x_i > t_i$). Omdat we binnen het Current Status model werken, kennen we $x_1, \ldots, x_{100}$ eigenlijk niet. We maken dus ook uitsluitend gebruik van de data-paren $(t_1, \delta_1), \ldots, (t_n, \delta_n)$.

Met deze data-paren kunnen de functies $l(F_\theta)$ en $\frac{d}{d\theta} l(F_\theta)$ worden bepaald. Hieronder zijn beide functies in één figuur weergegeven:

Figuur 1: Een voorbeeld van zulke data: De functies $l(F_\theta)$ (rood, herschaald) en $\frac{d}{d\theta} l(F_\theta)$ (blauw).

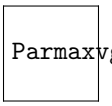
De methode van Newton zegt, dat het nulpunt $\tilde{\theta}$ van $\frac{d}{d\theta} l(F_\theta)$ benaderd kan worden door de volgende iteratie uit te voeren:

$$\theta_{n+1} = \theta_n - \frac{\frac{\partial}{\partial\theta} l(F_\theta)\big|_{\theta=\theta_k}}{\frac{\partial^2}{\partial\theta^2} l(F_\theta)\big|_{\theta=\theta_k}}.$$

Om er voor te zorgen dat de methode van Newton naar de $\tilde{\theta}$ convergeert waarin we geïnteresseerd zijn, moet θ_0 wel dicht genoeg bij die $\tilde{\theta}$ gekozen worden. (Als θ_0 toch te ver van $\tilde{\theta}$ ligt, kan demping er voor zorgen dat de methode van Newton toch naar $\tilde{\theta}$ convergeert.)
We weten dat:

$$\begin{aligned} \frac{\partial}{\partial\theta} l(F_\theta)\bigg|_{\theta=\theta_k} &= \sum_{i=1}^{n} \left(\frac{\delta_i t_i}{1 - \mathrm{e}^{-\theta_k t_i}} - t_i \right) \\ \frac{\partial^2}{\partial\theta^2} l(F_\theta)\bigg|_{\theta=\theta_k} &= \sum_{i=1}^{n} \left(-\frac{\delta_i t_i^2 \mathrm{e}^{-\theta_k t_i}}{(1 - \mathrm{e}^{-\theta_k t_i})^2} \right) \end{aligned}$$

De methode van Newton kan nu worden toegepast. Voor $\theta_0 = 0.1$ bleek de benadering binnen 10 stappen niet meer noemenswaardig te veranderen. Met de data uit het voorbeeld, bleek de benadering ongeveer gelijk te zijn aan 1.05. Dit is slechts één schatting voor θ_X. Ondanks dat deze schatting goed lijkt (herinner dat $\theta_X = 1$), zegt dit nog niks over hoe goed deze schattingsmethode is.
Door dit experiment te herhalen, dat wil zeggen voor verschillende trekkingen $\tilde{\theta}$ benaderen, komen we tot verschillende schatters. Hieronder is voor 4 experimenten de schatter weergegeven, samen met de exponentiële verdelingsfunctie met als parameter $\theta_X = 1$:



Figuur 2: De exponentiële verdelingsfunctie met parameter 1 (zwart) met 4 schatters.

Alle schatters liggen dicht bij de echte verdelingsfunctie. We kunnen dus informeel concluderen dat de schatters goed zijn.

Met zowel volledige als Current Status data kan de parametrische Maximum Likelihood schatter dus bepaald worden. Er is echter een belangrijk verschil tussen beide situaties. Dit illustreren we aan de hand van een voorbeeld:
Voor beide situaties schatten we 10,000 keer de parameter θ met de hiervoor beschreven methoden. Deze schattingen zijn in de onderstaande histogrammen verwerkt:

Figuur 3: Boven: Volledige data. Onder: Current Status data.

Hoewel de plaats van de piek en de vorm van beide histogrammen overeenkomen, is er ook een belangrijk verschil. Het histogram geconstrueerd met Current Status data is namelijk veel breder. Dit houdt in dat de geschatte waarden voor θ bij gelijke n grotere afwijkingen vertonen bij Current Status data, dan bij volledige data. Dit is de prijs die betaald moet worden voor gecensureerde data.

3.2 Niet-parametrische Maximum Likelihood methode

We kunnen de Maximum Likelihood methode ook toepassen binnen zogenaamde niet-parametrische modellen. Voordat we deze aanpak binnen het Current Status model uitwerken, bekijken we wederom de methode in het geval van ongecensureerde data $x_1, \ldots, x_n$. Als de onderliggende verdeling kansdichtheid f heeft, wordt de Log-Likelihood gegeven door

$$l(f) = \sum_{i=1}^{n} \log(f(x_i)).$$

Hier hangt l alleen af van de waarden van f in de punten $x_1, \ldots, x_n$. Een gevolg hiervan is het zogenaamde "spiking": Omdat f willekeurig groot in de punten $x_1, \ldots, x_n$ kan worden gekozen, zolang $\int_{-\infty}^{\infty} f(x)dx = 1$, kan ook l willekeurig groot worden. Dit is een probleem dat kan worden opgelost met behulp van discretisering. We beschouwen nu dus alle discrete verdelingen op $\{x_1, \ldots, x_n\}$. Daarbij is $f(x_i) = P(X = x_i) = p_i$, met $p_i \geq 0$. De Log-Likelihood-functie wordt dan:

$$l(f) = \sum_{i=1}^{n} \log(p_i).$$

Beschouw voor $\lambda > 0$ en $p = (p_1, \ldots, p_n)$

$$l(p) = \sum_{i=1}^{n} \log(p_i)$$

en

$$l_\lambda(p) = l(p) - \lambda\left(\sum_{i=1}^{n} p_i - 1\right).$$

Dan geldt voor elke $i \in \{1, \ldots, n\}$:

$$\frac{\partial}{\partial p_i} l_\lambda(p) = \frac{1}{p_i} - \lambda.$$

Als we de vergelijking $\frac{\partial}{\partial p_i} l_\lambda(p) = 0$ oplossen naar p_i, zien we dat l_λ wordt gemaximaliseerd door $\hat{p}_\lambda = (\hat{p}_{\lambda 1}, \ldots, \hat{p}_{\lambda n})$, met

$$\hat{p}_{\lambda i} = \frac{1}{\lambda}, \text{ voor elke } i.$$

Merk op dat we hier maximaliseren over $[0, \infty)^n$. Als we nu nemen $\lambda = n$, geldt:

$$\sum_{i=1}^{n} \hat{p}_{\lambda i} = 1.$$

Bovendien geldt dan voor alle $p \in [0, 1]^n$ met $\sum_{i=1}^{n} p_i = 1$:

$$l(p) = l_n(p) \leq l_n(\hat{p}_\lambda) = l(\hat{p}_\lambda).$$

Oftewel, l wordt gemaximaliseerd door $\hat{p}_\lambda$, met $\lambda = n$. Dat wil zeggen, l is maximaal voor:

$$p = (p_1, \ldots, p_n), \text{ met } p_i = \frac{1}{n} \text{ voor alle } i.$$

We hebben dan:

$$\hat{F}(x) = P(X \leq x) = \sum_{i=1}^{n} f(x_i) 1_{\{x_i \leq x\}} = \sum_{i=1}^{n} p_i 1_{\{x_i \leq x\}} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i \leq x\}}.$$

met

$$1_{\{x_i \leq x\}} = \begin{cases} 1, & \text{als } x_i \leq x \\ 0, & \text{als } x_i > x \end{cases}$$

Dit is precies de definitie van de empirische verdelingsfunctie F_n, dus er geldt $\hat{F} = F_n$.

Ook deze methode kan worden toegepast op het Current Status model. We beschouwen weer de Log-Likehood functie die we bij de parametrische Maximum Likelihood methode hebben gevonden:

$$l(F) = \sum_{i=1}^{n} \left(\delta_i \log(F(t_i)) + (1 - \delta_i) \log(1 - F(t_i))\right).$$

Deze l maximaliseren we nu over een grotere klasse van verdelingsfuncties dan dat we deden bij de parametrische Maximum Likelihood methode. De klasse bevat nu namelijk alle verdelingsfuncties. Merk op dat de waarde van l alleen afhangt van de waarden van F in de punten t_i, met $i = 1, \ldots, n$. Tussen die punten kunnen we F constant nemen. Hier treedt het probleem "spiking" overigens niet op, omdat altijd moet gelden dat $0 \leq F(x) \leq 1$ voor alle x.

Uit [3] volgt dat we $\hat{F}$, de niet-parametrische Maximum Likelihood schatter voor F, op de volgende manier kunnen vinden:
Beschouw het diagram, bestaande uit de volgende $n+1$ punten in $\mathbb{R}^2$:

$$P_i = \left(i, \sum_{j=1}^{i} \delta_i \right), 1 \leq i \leq n, \text{ en } P_0 = (0,0).$$

Hierbij geldt dat de paren (t_i, δ_i) geordend zijn, zodanig dat: $t_1 < t_2 < \cdots < t_n$. Als we vervolgens de grootste convexe minorant van deze punten nemen, dan is $\hat{F}(t_i)$ gelijk aan de linker afgeleide van deze grootste convexe minorant in het punt i.
In de klasse van verdelingfuncties die we bekijken (constant tussen de punten t_i), weten we bovendien dat deze Maximum Likelihood schatter uniek is.

3.2.1 Voorbeeld

In dit voorbeeld wordt dezelfde soort data gebruikt als in het voorbeeld van paragraaf 3.1.1. In onderstaande figuur zijn de punten P_i, zoals hierboven gedefinieerd, en de bijbehorende grootste convexe minorant weergegeven.

Nonparmaxlikelihood.jpg

Figuur 4: De punten P_i, met grootste convexe minorant (rood)

We weten dan dat $\hat{F}(t_i)$ gelijk is aan de linker afgeleide van deze grootste convexe minorant in het punt i. De op deze manier verkregen $\hat{F}$ is in de figuur hieronder weergegeven. Bovendien is de werkelijke F (de exponentiële verdeling met parameter $\theta = 1$) in dezelfde figuur geplot.

Nonparschatter.jpg

Figuur 5: De schatter $\hat{F}$ (blauw), en de werkelijke F (zwart) met $n = 100$.

Hoewel de convergentiesnelheid bij volledige data van de orde $n^{\frac{1}{2}}$ is, is de convergentiesnelheid van deze schatter binnen het Current Status model van de orde $n^{\frac{1}{3}}$. Voor kleine steekproeven, wat bij ons het geval is ($n = 100$), is het dus verklaarbaar dat de schatting wat grof is. Voor $n = 10,000$ is de schatter al een stuk beter, zie onderstaande figuur:

Nonparschatter10000.jpg

Figuur 6: De schatter $\hat{F}$ (blauw), en de werkelijke F (zwart) met $n = 10,000$.

Een vergelijking tussen de niet-parametrische Maximum Likelihood schatter en de parametrische Maximum Likelihood schatter laat zien dat de parametrische schatter (voor gelijke n) beter is.

Toch is het vaak handiger om de niet-parametrische schatter te gebruiken. Dit is vooral het geval als het niet bekend is tot welke klasse van verdelingen de te schatten verdelingsfunctie behoort. Stel dat we de data nu uit een Gamma-verdeling trekken in plaats van uit een exponentiële verdeling. We kunnen dan wederom de Maximum Likelihood schatters bepalen. Hierbij kiezen we in het parametrische model weer voor een exponentiële verdeling. Zie hieronder de twee schatters met de echte verdeling:

Figuur 7: De echte verdelingsfunctie (zwart) met parametrische (rood) en niet-parametrische (blauw) Maximum Likelihood schatter ($n = 10,000$).

Omdat de echte verdeling geen exponentiële verdeling is, zien we dat de parametrische schatter nu minder goed is dan de niet-parametrische schatter.

4 De Taut String benader-methode

In dit hoofdstuk beschouwen we het principe achter een andere schattingsmethode, namelijk de Taut String methode. Met deze methode is het mogelijk om elke willekeurige functie, en dus ook verdelingsfuncties, te schatten. Voor de uitleg van de methode beschouwen we alle functies, in het volgende hoofdstuk zullen we ons beperken tot verdelingsfuncties.
De Taut String methode is een zogenaamde regulariseringsmethode. Dit houdt in dat de Taut String-benadering of schatting van een functie minder grillig is dan de functie zelf. De mate van grilligheid is met een parameter te regelen. Deze vermindering van grilligheid is vaak handig bij het bestuderen van de functie.
We beginnen met een intuïtieve beschrijving van de methode. We geven grafisch weer wat de methode precies inhoudt. Daarna gaan we deze intuïtie "concretiseren met wiskunde". Met andere woorden: we formuleren het wiskundige probleem waar de Taut String methode op is gebaseerd. Vervolgens geven we een algoritme om de oplossing van het wiskundige probleem, dat hiervóór geformuleerd is, te vinden. We eindigen met een voorbeeld waarin we de Taut String methode toepassen.

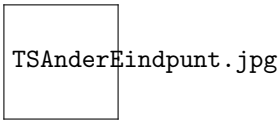
4.1 Het algemene principe

De functie g willen we schatten met behulp van de Taut String methode. Het enige wat deze methode eigenlijk nodig heeft, is een bovengrens en een ondergrens voor deze functie g. Als we ervan uitgaan dat we deze grenzen hebben, kunnen we op de volgende manier onze Taut String benadering vinden:
We weten dat de functie g tussen die grenzen blijft. We leggen nu een "touwtje" tussen deze grenzen. Dit touwtje kan al gezien worden als een soort schatting (het touwtje blijft namelijk tussen de bovengrens en de ondergrens van de gezochte functie). Vervolgens gaan we het touwtje strak trekken (we krijgen dan een "Taut String"), waarbij het touwtje de grenzen niet mag overschrijden.
We nemen aan dat het begin- en eindpunt van de boven- en ondergrens gelijk zijn. Is dit niet het geval, dan zal de schatter vanaf het voorlaatste punt óf horizontaal verder lopen óf, als de schatter daardoor niet tussen de boven- en ondergrens ligt, zo min mogelijk lineair stijgen of

dalen waardoor de schatter wel tussen de grenzen blijft.
Bij gelijk begin- en eindpunt, zit het touwtje dus "vast" op die punten. Na het strak trekken van het touwtje hebben we de Taut String benadering gevonden. Zie de figuur hieronder voor een voorbeeld:

Figuur 8: Links: De grenzen met een touwtje ertussen. Rechts: Het touwtje strak getrokken vanuit het begin- en eindpunt.

Als de grenzen niet hetzelfde eindpunt hebben, zal de schatter zich gedragen zoals hieronder is weergegeven. Dit geldt op een zelfde manier bij verschillende beginpunten.

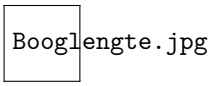


Figuur 9: De Taut String schatter bij verschillend eindpunt.

Intuïtief is de Taut String schatter nu gevonden: een strak touwtje, dat tussen de bovengrens en de ondergrens blijft van de functie g. Deze schatter zal op een aantal plaatsen langs die bovengrens of ondergrens lopen, en tussen die punten is de schatter lineair.
Dit principe moet nu worden omgezet in bruikbare wiskunde.

4.2 Het wiskundige probleem

Het strak trekken van het touwtje tussen de grenzen, kunnen we opvatten als het verkleinen van de afstand die je aflegt als je vanaf het beginpunt langs het touwtje naar het eindpunt loopt. Dit wordt ook wel de booglengte genoemd. Deze afstand kan wiskundig worden beschreven. Bekijk de volgende figuur:



Figuur 10: Benadering van de booglengte met Δs

We benaderen nu de booglengte met Δs. Uit de stelling van Pythagoras volgt dat: $\Delta s = \sqrt{(\Delta x)^2 + (\Delta y)^2}$. Bovendien kan Δy benaderd worden door een uitdrukking in Δx, en wel als volgt:
Neem aan dat de functie, zeg g, tweemaal continu differentieerbaar is, dat wil zeggen g' en g'' bestaan en zijn continu. Dit is in het bijzonder het geval bij continue verdelingsfuncties. Het Taylorpolynoom van orde 1 van g rond het punt x is dan:

$$g(x + \Delta x) = g(x) + \Delta x g'(x) + \frac{1}{2}(\Delta x)^2 g''(\xi) \text{ voor een } \xi \in \mathbb{R}$$

Dit geeft:

$$\underbrace{g(x+\Delta x)-g(x)}_{\Delta y} = \Delta x g'(x) + \frac{1}{2}(\Delta x)^2 g''(\xi)$$

Nu kan dus ook Δs benaderd worden door een uitdrukking in Δx:

$$\begin{aligned}\Delta s &= \sqrt{(\Delta x)^2+(\Delta y)^2} = \sqrt{(\Delta x)^2\left(1+\left(\frac{\Delta y}{\Delta x}\right)^2\right)} = \sqrt{1+\left(\frac{\Delta y}{\Delta x}\right)^2}\Delta x \\ &= \sqrt{1+\left(g'(x)+\frac{1}{2}(\Delta x)g''(\xi)\right)^2}\Delta x = \sqrt{1+g'(x)^2+r_{\Delta x}(x)}\Delta x, \\ &\quad \text{waarbij de restterm } r_{\Delta x}(x)\to 0 \text{ als } \Delta x\to 0.\end{aligned}$$

Als we het interval $[a,b]$ opdelen in n kleine stukjes, met grenswaarden $x_i = a + i\frac{b-a}{n}$ voor $i = 0,\ldots,n$, dan geldt dat $\Delta x_i = x_i - x_{i-1} = \frac{b-a}{n} = \Delta x$. Vervolgens tellen we alle stukjes Δs_i op. De booglengte kan dan dus benaderd kan worden met:

$$A_n = \sum_{i=1}^{n}\Delta s_i = \sum_{i=1}^{n}\sqrt{1+g'(x_i)^2+r_{\Delta x}(x)}\Delta x$$

Omdat voor $n\to\infty$ geldt dat $\Delta x\to 0$ en dus $r_{\Delta x}(x)\to 0$, geldt voor

$$B_n = \sum_{i=1}^{n}\sqrt{1+g'(x_i)^2}\Delta x$$

dat:

$$\lim_{n\to\infty} B_n = \lim_{n\to\infty} A_n.$$

We nemen nu aan, dat de functie $x\mapsto\sqrt{1+g'(x)^2}$ Riemann integreerbaar is. Dan is B_n de bijbehorende Riemann-som van de integraal $\int_a^b\sqrt{1+g'(x)^2}dx$. Als nu $\Delta x\to 0$, oftewel $n\to\infty$, volgt dat:

$$\lim_{n\to\infty}\Delta s_i = \lim_{n\to\infty} A_n = \lim_{n\to\infty} B_n = \int_a^b\sqrt{1+g'(x)^2}dx$$

De afstand die je aflegt als je van a naar b langs g loopt is dus gelijk aan: $\int_a^b\sqrt{1+g'(x)^2}dx$.

Deze afstand minimaliseren we, door een functie G te zoeken, die de uitdrukking $\int_a^b\sqrt{1+G'(x)^2}dx$ minimaliseert. Zonder verdere voorwaarden, wordt deze uitdrukking geminimaliseerd door de functie $G\equiv c$, voor een $c\in\mathbb{R}$ (dan geldt namelijk $G'(x)^2 = 0$ en dus minimaal). Omdat deze oplossing niet interessant is, moeten er voorwaarden aan de functie G worden opgelegd.
Een eerste eis is dat moet gelden $G(a) = g(a)$ en $G(b) = g(b)$. De benadering heeft dan in ieder geval hetzelfde begin- en eindpunt als de gezochte functie. Stellen we daarnaast geen andere voorwaarden, dan zal de benadering een rechte lijn van het punt $g(a)$ naar het punt $g(b)$ zijn.
Een tweede eis is dat de benadering tussen twee vooraf bepaalde grenzen blijft. Om er voor

te zorgen dat de benadering niet te ver van de echte functie af ligt, stellen we als eis, dat G hooguit een supremum-afstand α van g af ligt. Oftewel:

$$\max_{a \leq x \leq b} |G(x) - g(x)| \leq \alpha.$$

Een korte samenvatting:
De functie g benaderen we met een andere functie G, waarbij G de functie is die de volgende uitdrukking minimaliseert:

$$\int_a^b \sqrt{1 + G'(t)^2} dt,$$

onder de voorwaarden:

$$G(a) = g(a), G(b) = g(b)$$
$$\max_{a \leq t \leq b} |G(t) - g(t)| \leq \alpha.$$

Merk op, dat het volgende geldt:

- Voor $\alpha = 0$, moet gelden: $\max_{a \leq t \leq b} |G(t) - g(t)| \leq 0$, oftewel $\max_{a \leq t \leq b} |G(t) - g(t)| = 0$. Dit betekent dus dat de functie G gelijk moet zijn aan g. We moeten dan wel nagaan of de afgeleide van G wel bestaat. Overigens wordt in de praktijk zelden de waarde $\alpha = 0$ gebruikt.
- Voor α relatief groot, mag de benadering G heel erg afwijken van de functie g. Als α maar groot genoeg is, zal G een rechte lijn zijn van $(a, g(a))$ naar $(b, g(b))$.
- Deze α kan dus worden gezien als een parameter die de mate van grilligheid van de benadering G beïnvloedt. Dit wordt ook wel een regulariseringsparameter genoemd.

4.3 Een algoritme

Het wiskundige probleem dat hiervoor is geformuleerd, is lastig op te lossen. Er is echter een algoritme, bedacht door P.L. Davies en A. Kovac (zie de Appendix van [2]), waarmee we precies díe functie vinden, die de booglengte minimaliseert onder de genoemde voorwaarden.
Dat algoritme gaan we nu gebruiken om de Taut String benadering te vinden. Daarbij gaan we er van uit dat we $n + 1$ data-punten hebben, en dat de bovengrens u en de ondergrens l als functie van x, met $x = 0, 1, \ldots, n$, bekend zijn. Bovendien nemen we aan, dat het begin- en eindpunt van deze grenzen gelijk zijn, oftewel: $u(0) = l(0)$ en $u(n) = l(n)$.
Het algoritme:

1. Bepaal stapsgewijs de grootste convexe minorant (GCM) van de bovengrens en de kleinste concave majorant (LCM) van de ondergrens. Dat wil zeggen: we bepalen eerst de GCM en LCM van de eerste 2 data-punten, daarna van de eerste 3 data-punten, enzovoorts.
2. Controleer na elk extra punt of de rechter afgeleide in 0 van de GCM van de bovengrens groter is dan de rechter afgeleide in 0 van de LCM van de ondergrens. Als dit het geval is, voegen we een extra punt toe en bepalen we vervolgens weer de GCM en de LCM (m.a.w. we gaan terug naar Stap 1). Is dit echter niet het geval, dan gaan we naar de volgende stap.

3. Bepaal de meest linkse knoop die op dat moment ofwel in de GCM van de bovengrens zit ofwel in de LCM van de ondergrens. Daarbij laten we het beginpunt buiten beschouwing. Dit punt behoort nu tot de Taut String.

4. Verplaats het beginpunt naar dat punt, en herhaal de vorige stappen.

5. Stop als het eindpunt is bereikt.

Een bepaald aantal punten behoort nu tot de Taut String. De Taut String-schatter is dan gelijk aan de lineaire interpolatie tussen die punten.
Dit algoritme hebben we verwerkt in een programma in R (zie de Appendix 7.1 voor de code) dat, gegeven de boven- en ondersgrens, de Taut String-schatter bepaalt. In het vermelde programma staan de grenzen gegeven waar we in het voorbeeld mee werken.

4.4 Een voorbeeld

We gaan nu een simpel voorbeeld bekijken, waarin het Taut String algoritme wordt toegepast. Hierbij hebben we zelf een bovengrens en ondergrens gekozen. Deze grenzen zijn een sinus, waar in elk punt een ander willekeurig en relatief klein positief respectievelijk negatief getal bij wordt opgeteld (zie Appendix 7.1 voor details). Verder zijn er 6 data-punten, waarvan het beginpunt en het eindpunt beide gelijk zijn aan 0. Zie de figuur:

Figuur 11: De bovengrens (stippellijn) en de ondergrens (doorgetrokken lijn)

Bij stap 1 van het algoritme bepalen we de GCM van de bovengrens en de LCM van de ondergrens van de eerste 2 data-punten. Vervolgens doen we de controle van stap 2 van het algoritme. Omdat de ongelijkheid stand houdt (zie de figuur hieronder), wordt het 3e data-punt toegevoegd. Merk overigens op dat de ongelijkheid bij slechts 2 data-punten altijd stand houdt. Nu wordt de GCM en de LCM van de eerste 3 data-punten bepaald. We zien nu dat de ongelijkheid niet meer stand houdt. Bij stap 3 voegen we het meest linkse punt ongelijk aan het beginpunt dat ofwel in de GCM ofwel in de LCM zit, toe aan onze Taut String. Dit blijkt het 2e punt van de LCM van de ondergrens te zijn. Zie de figuur hieronder voor de grafische weergave van de genomen stappen.

Figuur 12: De GCM (rood) van de bovengrens en de LCM (blauw) van de ondergrens

Na het uitvoeren van stap 4, het verplaatsen van het beginpunt naar het punt dat we zojuist aan onze Taut String hebben toegevoegd, beginnen we weer bij stap 1. De rest van de te nemen stappen is grafisch weergegeven in onderstaande figuur.

Figuur 13: De GCM (rood) van de bovengrens en de LCM (blauw) van de ondergrens en de Taut String-schatter (groen) tot dat moment

Nu is bekend welke punten tot de Taut String behoren. Na lineaire interpolatie tussen die punten hebben we onze Taut String-schatter voor de gezochte functie gevonden. Zie hieronder nog een keer de uiteindelijke Taut String-schatter:

Figuur 14: De Taut String-schatter (groen)

5 De Taut String methode voor verdelingsfuncties

Nu we weten wat de Taut String methode is, kunnen we deze methode gaan gebruiken voor het schatten van verdelingsfuncties. We hebben nu dus altijd te maken met monotoon stijgende functies, die in het beginpunt waarde 0 en in het eindpunt waarde 1 hebben. Omdat de monotone stijging zich vertaalt in een monotone stijging van de boven- en ondergrens, zal de Taut String schatter ook altijd monotoon stijgend zijn. Als bovendien beide grenzen in het beginpunt waarde 0 en in het eindpunt waarde 1 hebben, zal de schatter een verdelingsfunctie zijn.
We gaan eerst in het geval van volledige data de Taut String schatter voor de verdelingsfunctie bepalen. Daarna bepalen we de Taut String schatter binnen het Current Status model.

5.1 Volledige data

We willen van een stochast X de verdelingsfunctie F met de Taut String methode schatten. De data is nu volledig. Dit betekent dat de data $x_1, \ldots, x_n$ beschikbaar zijn. We kunnen dan de empirische verdelingsfunctie F_n opstellen:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i \leq x\}},$$

met

$$1_{\{x_i \leq x\}} = \begin{cases} 1, & \text{als } x_i \leq x \\ 0, & \text{als } x_i > x \end{cases}$$

Met behulp van deze empirische verdelingsfunctie kunnen we een bovengrens en een ondergrens vinden. Voor $u > 0$ weten we namelijk hoe

$$\begin{aligned} P(\sup_x |F_n(x) - F(x)| \leq u) &= P(-u \leq F_n(x) - F(x) \leq u \ \forall x) \\ &= P(F_n(x) - u \leq F(x) \leq F_n(x) + u \ \forall x) \end{aligned}$$

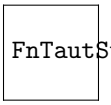
zich gedragen. Dit is de kans dat de werkelijke verdelingsfunctie F tussen de waarden $F_n + u$ en $F_n - u$ (de empirische verdelingsfunctie F_n met een getal u er bij opgeteld respectievelijk van afgetrokken) blijft.
Stel, we hebben de empirische verdelingsfunctie bij onze data geconstrueerd. Door bij deze functie een bepaald getal u op te tellen en af te trekken, kan een bovengrens en een ondergrens worden geconstrueerd. Zie de figuur hieronder:

Figuur 15: Links: Empirische verdelingsfunctie. Rechts: Met bovengrens (rood) en ondergrens (blauw)

Nu we een boven- en ondergrens hebben gevonden, kunnen we het Taut String algoritme toepassen. Dit leidt tot de volgende Taut String schatter:

Figuur 16: Links: Met F_n gevonden grenzen. Rechts: Taut String-schatter (groen)

Hieronder is de Taut String-schatter nog een keer geplot, met de echte verdelingsfunctie (exponentiële verdeling met parameter 1) in dezelfde figuur geplot:



Figuur 17: De Taut String schatter (groen) en de echte verdelingsfunctie (zwart)

In dit voorbeeld is een willekeurige u gekozen. Daarbij is geen rekening gehouden met de kans dat de werkelijke verdelingsfunctie tussen de grenzen $F_n + u$ en $F_n - u$ blijft. Omdat we het stochastische gedrag van $\|F_n - F\|_\infty$ kennen, kunnen we bepalen voor welke u de verdelingsfunctie in bijvoorbeeld 95% van de gevallen tussen deze grenzen blijft. We kunnen dus een 95%-betrouwbaarheidsband voor F construeren.

We bepalen de u met behulp van Monte Carlo simulatie. Laat D_n de maximale afstand

tussen de empirische verdelingsfunctie F_n en de echte verdelingsfunctie F zijn. Dit maximum wordt, voor gegeven $x_1, \ldots, x_n$, aangenomen in één van deze data-punten. De afstand tussen zowel de linker- als de rechter-limiet van F_n in zo'n punt en de waarde van F in dat punt is daarbij van belang. Er geldt dus:

$$D_n = \max_{1 \leq i \leq n} \{|F_n(x_{i-1}) - F(x_i)|, |F_n(x_i) - F(x_i)|\}, \text{ met } F_n(x_0) = 0.$$

Genereer nu voor $y_1, \ldots, y_n$ uit $U[0,1]$, de uniforme verdeling op $[0,1]$, de data $x_1, \ldots, x_n$ door $x_i = F^{-1}(y_i)$ voor alle i. Omdat F een monotoon stijgende functie is, geldt:

$$F_n(x_i) = \frac{\#\{j : x_j \leq x_i\}}{n} = \frac{\#\{j : F(x_j) \leq F(x_i)\}}{n} = F_n^U(F(x_i)) = F_n^U(y_i).$$

Als de x_i's gesorteerd zijn, en dus ook de y_i's, geldt dan:

$$\begin{aligned} D_n &= \max_{1 \leq i \leq n} \{|F_n^U(y_{i-1}) - F(F^{-1}(y_i))|, |F_n^U(y_i) - F(F^{-1}(y_i))|\} \\ &= \max_{1 \leq i \leq n} \left\{ \left| \frac{i-1}{n} - y_i \right|, \left| \frac{i}{n} - y_i \right| \right\}. \end{aligned}$$

Voor het bepalen van D_n zijn dus alleen trekkingen uit de uniforme verdeling nodig. De verdelingsfunctie F hoeven we niet te kennen.

We hebben 100,000 keer een trekking gedaan, met $n = 25$, uit de uniforme verdeling. Voor elke trekking berekenen we D_n, om vervolgens het 95%-kwantiel (onze u) daarvan te bepalen. Dit kwantiel blijkt ongeveer gelijk te zijn aan $u = 0.264$ (Zie Appendix 7.2 voor de code van het programma in R).

In een artikel van Marsaglia, Tsang en Wang (zie [4]) is de verdeling van D_n bestudeerd, waarbij de volgende limiet-vorm is gevonden:

$$L(x) = \lim_{n \to \infty} P(\sqrt{n} D_n \leq x) = \frac{\sqrt{2\pi}}{x} \sum_{i=1}^{\infty} \mathrm{e}^{-\frac{(2i-1)^2 \pi^2}{8x^2}}.$$

Op deze manier kan ook het 95%-kwantiel bepaald worden. Dit bleek gelijk te zijn aan $u^* = 0.272$. Het op deze manier gevonden kwantiel ligt in de buurt van de eerder gevonden $u = 0.264$. Deze u kunnen we nu gebruiken om de Taut String methode toe te passen.

Met behulp van de originele data-set $x_1, \ldots, x_n$ kan de empirische verdelingsfunctie F_n worden opgesteld. We weten nu dat de echte verdelingsfunctie F in ongeveer 95% van de gevallen tussen de grenzen $F_n + u$ en $F_n - u$ blijft, met $u = 0.264$. Zie hieronder de empirische verdelingsfunctie van de dataset, met rechts de bijbehorende 95%-betrouwbaarheidsband:

FnGoedeU.jpg

Figuur 18: Links: De empirische verdelingsfunctie. Rechts: De bijbehorende 95%-betrouwbaarheidsband.

Met deze boven- en ondergrens voor F kunnen we het Taut String algoritme toepassen. De schatter die we daarmee vinden, is in de figuur hieronder (links) weergegeven. In de rechter figuur is de Taut String schatter nog een keer weergegeven, met in dezelfde figuur de echte verdelingsfunctie (de exponentiële verdeling met parameter 1).

Figuur 19: Links: De Taut String schatter (groen). Rechts: De echte verdelingsfunctie (zwart).

We kunnen concluderen dat de schatter niet echt goed is (vergelijk met de parametrische Maximum Likelihood schatter in figuur 2 op pagina 8). De 95%-betrouwbaardheidsband die we construeren is voor $n = 25$ te groot om een goede Taut String schatter te vinden. Dit heeft te maken met het feit dat de empirische verdelingsfunctie (de niet-parametrische Maximum Likelihood schatter) voor kleine n in het algemeen niet nauwkeurig is.

We kunnen op twee manieren tot een betere schatter proberen te komen:

1. Voor grotere n is de empirische verdelingsfunctie nauwkeuriger, waardoor we een kleinere betrouwbaarheidsband krijgen. (In de praktijk ligt de n, het aantal data-punten, uiteraard vast.)
2. In plaats van een 95%-betrouwbaarheidsband te nemen, kunnen we ook met bijvoorbeeld een 85%-betrouwbaarheidsband werken.

We beginnen met het eerste punt:
Voor verschillende n bepalen we u zodanig dat we een 95%-betrouwbaarheidsband hebben. Ook u^*, het kwantiel gevonden met de limiet-vorm van de kansverdeling van D_n, is ter controle in onderstaande tabel weergegeven:

n	25	50	100	200
u	0.264	0.189	0.134	0.095
u^*	0.272	0.192	0.136	0.096

Bij meer data-punten, oftewel een grotere n, wordt de betrouwbaarheidsband inderdaad kleiner. Merk verder op dat u en u^* steeds dichter bij elkaar liggen. Hieronder is voor de gegeven n de Taut String schatter met de echte verdelingsfunctie geplot:

Figuur 20: Boven: $n = 25$ (links) en $n = 50$ (rechts). Onder: $n = 100$ (links) en $n = 200$ (rechts).

We zien dat de Taut String schatter beter wordt naarmate n toeneemt. Bij $n = 200$ kunnen we zelfs zeggen dat de schatter redelijk nauwkeurig is. Voor kleine waarden van n is de schatter

echter minder goed.

Het is ook mogelijk dat we op een andere manier een betere schatter krijgen. We kunnen bijvoorbeeld een andere betrouwbaarheidsband nemen. Hierbij nemen we een constante n, en wel $n = 200$. Hieronder staan voor enkele betrouwbaardheidsbanden de bijbehorende u:

Betrouwbaarheidsband	95%	85%	75%	65%
u	0.095	0.080	0.071	0.065
u^*	0.096	0.080	0.072	0.066

Hieronder is voor de gegeven betrouwbaarheidsbanden de Taut String schatter met de echte verdelingsfunctie geplot:

Figuur 21: Boven: 95% (links) en 85% (rechts). Onder: 75% (links) en 65% (rechts).

In de bovenstaande figuur treedt zeker een verbetering op naarmate de betrouwbaarheidsband kleiner wordt. De reden hiervan is het feit dat de empirische verdelingsfunctie voor $n = 200$ een redelijk goede schatter is. Met een kleinere betrouwbaarheidsband kan dus een betere schatter gevonden worden. Dit geldt vaak niet voor kleinere n, zoals bijvoorbeeld bij $n = 25$, omdat de empirische verdelingsfunctie in dat geval een minder goede schatter is. Het is echter wel mogelijk dat een verbetering optreedt.
Met volledige data kunnen we dus een Taut String schatter vinden. Voor een nauwkeurige schatter moet wel gelden $n \geq 200$.

5.2 Current Status data

Binnen het Current Status model zijn de data $x_1, \ldots, x_n$ voor ons gecensureerd. We kunnen dus niet de empirische verdelingsfunctie F_n opstellen. Bovendien is het door de vaste T_i-waarden ook niet mogelijk om een herschaling uit te voeren, zoals bij volledige data wel kon. Dat wil zeggen, we kunnen nu niet D_n (en dus u) bepalen op basis van de uniforme verdeling. De methode met volledige data, met grenzen $F_n + u$ en $F_n - u$ voor een 95%-betrouwbaarheidsband, werkt dus niet met Current Status data. We moeten dus een andere manier vinden om deze grenzen te vinden.

De data die we nu tot onze beschikking hebben, zijn de data-paren $(t_1, \delta_1), \ldots, (t_n, \delta_n)$. We kunnen met die data-paren de niet-parametrische Maximum Likelihood schatter vinden (zie paragraaf 3.2). We bekijken nu niet de schatter die constant is tussen de punten t_i, maar de schatter die lineair geïnterpoleerd is tussen die punten. We noemen deze schatter $\hat{F}_n$.

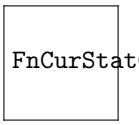
Figuur 22: De schatter constant (blauw) en geïnterpoleerd (rood, $\hat{F}_n$) tussen de punten t_i

Beschouw nu $\|\hat{F}_n^* - \hat{F}_n\|_\infty$ in plaats van $\|F_n - F\|_\infty$. Hierbij bepalen we $\hat{F}_n^*$ als volgt:
Voor $i = 1, \ldots, n$ bepalen we δ_i^* door een trekking te doen uit een Bernoulli-verdeling met parameter $\hat{F}_n(t_i)$. Dat houdt in dat δ_i^* met kans $\hat{F}_n(t_i)$ gelijk aan 1 is en met kans $1 - \hat{F}_n(t_i)$ gelijk aan 0. Dan is $\hat{F}_n^*$ de niet-parametrische Maximum Likelihood schatter, die constant is tussen de punten t_i, gebaseerd op de data-paren $(t_1, \delta_1^*), \ldots, (t_n, \delta_n^*)$.

We nemen aan dat $\hat{F}_n$ voor $n = 200$ een dusdanig goede schatter is, dat we D_n kunnen benaderen met C_n:

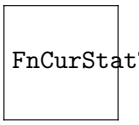
$$C_n = \max_{1 \le i \le n} \{|\hat{F}_n^*(x_{i-1}) - \hat{F}_n(x_i)|, |\hat{F}_n^*(x_i) - \hat{F}_n(x_i)|\}, \text{ met } \hat{F}_n^*(x_0) = 0.$$

We genereren nu 100,000 keer een $\hat{F}_n^*$, met $n = 200$, om zo C_n te berekenen. Hiervan bepalen we vervolgens het 95%-kwantiel. Voor een gegeven $\hat{F}_n$ blijkt deze gelijk te zijn aan $u = 0.302$. Dit kwantiel is echter afhankelijk van $\hat{F}_n$. Met andere woorden, voor verschillende $\hat{F}_n$ heeft het kwantiel een andere waarde. Deze waarden zijn maximaal ongeveer 0.4 als de data uit een exponentiële verdeling komt, en zelfs maximaal ongeveer 0.5 bij een normale verdeling. Voor $n = 25$ bij volledige data bleken deze waarden te groot te zijn om een goede schatter te vinden. In de figuur hieronder is de $\hat{F}_n$ voor $n = 200$ weergegeven, met daarnaast de bijbehorende 95%-betrouwbaarheidsband.



Figuur 23: Links: de lineair geïnterpoleerde niet-parametrische Maximum Likelihood schatter. Rechts: de bijbehorende 95%-betrouwbaarheidsband.

Als op deze grenzen het Taut String algoritme wordt toegepast, levert dit de volgende schatter op:



Figuur 24: De Taut String schatter bij de gegeven grenzen en $n = 200$.

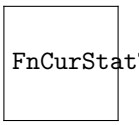
Zoals al bij volledige data bleek, is de geconstrueerde betrouwbaarheidsband ook in dit geval te groot om een goede schatter te vinden. De schatter is zelfs een rechte lijn, oftewel een uniforme

verdeling. Voor $n = 200$ levert deze methode dus geen goede schatter op. Om de Taut String methode toe te kunnen passen binnen het Current Status model, zal dus meer (onvolledige) data beschikbaar moeten zijn.

In onderstaande tabel zijn voor enkele waarden van n het bijbehorende 95%-kwantiel u van C_n (voor een gegeven $\hat{F}_n$) weergegeven:

n	200	400	1,000	5,000
u	0.302	0.227	0.189	0.104

Bij volledige data kon worden geconcludeerd dat bij $u = 0.095$ de Taut String schatter goed was. Binnen het Current Status model vinden we nu dat bij $n = 5,000$ voor een gegeven $\hat{F}_n$ geldt dat $u = 0.104$. Voor een gegeven $\hat{F}_n$ bij deze n, is de Taut String schatter in onderstaande figuur, samen met de echte verdelingsfunctie, weergegeven:



Figuur 25: De Taut String schatter bij de gegeven grenzen en $n = 5,000$.

Nu kunnen we, wederom informeel, wel concluderen dat de schatter goed is. Dit komt mede door het feit dat de niet-parametrische Maximum Likelihood schatter bij $n = 5,000$ goed is. Een groot voordeel van de Taut String schatter ten opzichte van de niet-parametrische Maximum Likelihood schatter is dat de Taut String schatter veel minder grillig is. Het is dus een eenvoudige functie, waar gemakkelijk mee is te rekenen.

Binnen het Current Status model kan dus ook, met de hiervoor beschreven methode, een goede Taut String schatter gevonden worden. Er is, in vergelijking met volledige data, echter wel veel meer data nodig voor een zelfde nauwkeurigheid.

6 Conclusie en aanbevelingen

Binnen het Current Status model kan met behulp van verschillende methoden een schatter worden gevonden voor een onbekende verdelingsfunctie. Twee veel gebruikte methoden zijn de parametrische en niet-parametrische Maximum Likelihood methode. Bij gecensureerde data zullen in het parametrische model de schattingen voor de parameter een grotere afwijking vertonen dan bij volledige data (zie figuur 3 op bladzijde 9). In het niet-parametrische model heeft de censurering van de data invloed op de convergentie-snelheid. Bij volledige data is deze namelijk van de orde $n^{\frac{1}{2}}$, bij gecensureerde data van de orde $n^{\frac{1}{3}}$.
Ook de Taut String methode is toe te passen binnen het Current Status model. Het enige dat nodig is om het algoritme, waarmee de Taut String schatter gevonden kan worden, te kunnen gebruiken, is een boven- en een ondergrens. Bij volledige data blijkt het vinden van deze grenzen relatief eenvoudig. Door met behulp van de empirische verdelingsfunctie een 95%-betrouwbaarheidsband te construeren, kan het algoritme worden toegepast. Als de data gecensureerd is, is het vinden van een boven- en ondergrens gecompliceerder. Door gebruik te

maken van de niet-parametrische Maximum Likelihood schatter, is het wel mogelijk om via een omweg een 95%-betrouwbaarheidsband te constueren. Om op deze manier tot een goede schatter te komen, is er echter wel veel meer data nodig dan bij volledige data.

Dit probleem zou opgelost kunnen worden, door een efficiëntere manier te vinden om met behulp van Current Status data een goede boven- en ondergrens te vinden voor de te schatten verdelingsfunctie. Uit het feit dat dit nog een open probleem is in de wiskunde, blijkt dat dit niet eenvoudig is. Het oplossen van dit probleem, zou er wel voor kunnen zorgen dat de Taut String methode ook voor kleinere hoeveelheden data een goede schatter is. Doordat de schatter een eenvoudige functie is, zal de Taut String methode erg bruikbaar kunnen zijn.

7 Appendix

7.1 Taut String algoritme in R

```
# Laden van "fdrtool"
library("fdrtool")

par(mfrow=c(1,1))

# We genereren een onder- (l) en bovengrens (u)
n<-5
t = (0:n)
u<-t
s<-0
u = sin((2*pi*t)/n)+abs(rnorm(n+1,0,0.2))
u[1]<-0
u[n+1]<-0
l = sin((2*pi*t)/n)-abs(rnorm(n+1,0,0.2))
l[1]<-0
l[n+1]<-0
umax<-max(u)
lmin<-min(l)

# We plotten de onder- en bovengrens
plot(t, u,ylim=c(lmin,umax), type="l", lty=3, lwd=2, main="De bovengrens en ondergrens")
points(t, u)
lines(t,l, lwd=2)
points(t,l)

# Onze Taut String (s)
s[1]<-u[1]
i<-0

for(j in 1:n)
{s[j+1]=NA}

upperslope<-2
lowerslope<-1

par(ask=TRUE)

par(mfrow=c(1,2))

# Stappen 1 en 2 voor het beginpunt
while(upperslope > lowerslope)
{
i = i+1
x = (0:i)
y = x
for(j in 1:length(x))
{y[j]<-u[j]}

plot(t, u,ylim=c(lmin,umax), type="l", lty=3, lwd=2)
```

```
points(t, u)
lines(t,l, lwd=2)
points(t,l)

# De grootste convexe minorant van de bovengrens (rood)
gg = gcmlcm(x,y)
lines(gg$x.knots, gg$y.knots, col=2, lwd=2)

upperslope = gg$slope.knots[1]
upperknot = gg$x.knots[2]

for(j in 1:length(x))
{y[j]<-l[j]}

# De kleinste concave majorant van de ondergrens (blauw)
ll = gcmlcm(x,y, type="lcm")
lines(ll$x.knots, ll$y.knots, col=4, lwd=2)

lowerslope = ll$slope.knots[1]
lowerknot = ll$x.knots[2]

title("\n\n\nGCM (red) bovengrens en LCM (blue) ondergrens", outer = TRUE)
}

# We voegen het meeste linkse punt toe aan onze Taut String
if(upperknot<=lowerknot) s[upperknot+1]<-u[upperknot+1] else if(upperknot>lowerknot)
  s[lowerknot+1]<-l[lowerknot+1] else if(upperslope==lowerslope) s[upperknot+1]<-u[upperknot+1]

par(mfrow=c(3,3))

# De stappen 1,2 en 3 voor de rest van de punten
while((upperknot==n)==FALSE || (lowerknot==n)==FALSE)
{

# We verplaatsen het beginpunt
a<-min(lowerknot,upperknot)
i<-a

upperslope<-2
lowerslope<-1

# Stappen 1 en 2 voor het nieuwe beginpunt
while(upperslope > lowerslope)
{

plot((0:n), u, ylim=c(lmin,umax), type="l", lty=3, lwd=2, main="GCM (red) and LCM (blue)")
points((0:n), u)
lines((0:n),l, lwd=2)
points((0:n),l)
ind<-!is.na(s)
lines((0:n)[ind],s[ind], col=3, lwd=2)
points((0:n),s,pch=19)
```

```
i = i+1
x = (a:i)
y = x

for(j in 1:length(x))
{y[j]<-u[j+a]}
y[1]<-s[a+1]

# De grootste convexe minorant van de bovengrens (rood)
gg = gcmlcm(x,y)
lines(gg$x.knots, gg$y.knots, col=2, lwd=2)

upperslope = gg$slope.knots[1]
upperknot = gg$x.knots[2]

for(j in 1:length(x))
{y[j]<-l[j+a]}
y[1]<-s[a+1]

# De kleinste concave majorant van de ondergrens (blauw)
ll = gcmlcm(x,y, type="lcm")
lines(ll$x.knots, ll$y.knots, col=4, lwd=2)

lowerslope = ll$slope.knots[1]
lowerknot = ll$x.knots[2]
}

# We voegen weer een punt toe aan onze Taut String
if(upperknot<=lowerknot) s[upperknot+1]<-u[upperknot+1] else if(upperknot>lowerknot)
  s[lowerknot+1]<-l[lowerknot+1] else if(upperslope==lowerslope) s[upperknot+1]<-u[upperknot+1]
}

We plotten onze bovengrens, ondergrens en de Taut String-schatter
x = (0:n)
plot(x, u, ylim=c(lmin,umax), type="l", lty=3, lwd=2, main="GCM (red) and LCM (blue)")
points(x, u)
lines(x,l, lwd=2)
points(x,l)
ind<-!is.na(s)
lines(x[ind],s[ind], col=3, lwd=2)
points(x,s,pch=19)
```

7.2 Bepaling van u m.b.v. Monte Carlo simulatie in R

```
B = 100000
Dn = 0
N = 25

el = (0:(N-1))/N
er = (1:N)/N
```

```
for(b in 1:B)
{

X = sort(runif(N))

vl = abs(el-X)
rl = abs(er-X)

Dn[b] = max(vl,rl)
}

quantile(Dn,0.95)
```

Referenties

[1] Extensions of smoothing via taut strings, van A. Kovac, Electronic Journal of Statistics, 2009, Vol. 3, blz 41-75.

[2] Local extremes, runs, strings and multiresolution, van P.L. Davies en A. Kovac, The Annals of Statistics, 2001, Vol. 29, Nr. 1, blz 1-65.

[3] Information bounds and nonparametric maximum likelihood estimation, van P. Groeneboom en J.A. Wellner, 1992.

[4] Evaluating Kolmogorov's Distribution, van G. Marsaglia, W.W. Tsang en J. Wang.