Universiteit Utrecht

Faculteit Bètawetenschappen

Bachelorscriptie

Bayesiaanse Survival Analyse

Auteur:
C.C.E. Smans
5621216

Begeleider:
dr. M.C.J. Bootsma

15 juni 2018

Inhoudsopgave

1 Inleiding **3**

2 Eenvoudig model met twee toestanden **5**
2.1 Modelbeschrijving 5
2.2 Resultaten 8
2.2.1 Constante baseline hazard 8
2.2.2 Stuksgewijs constante baseline hazard 13
2.2.3 Weibull verdeling 17
2.3 Discussie 19

3 Multistate model **21**
3.1 Modelbeschrijving 21
3.2 Resultaten 23
3.3 Discussie 26

4 Conclusie **27**

1 Inleiding

Vaak voorkomende vragen in het ziekenhuis zijn vragen als 'hoe lang duurt het tot de patiënt genezen is?' of 'hoe lang heeft de patiënt nog te leven?'. Het is onmogelijk om een precies antwoord te geven op deze vragen, maar op basis van eerdere gebeurtenissen is het wel mogelijk om een schatting te doen. De schattingen kunnen worden gedaan met behulp van de survival analyse. Met de survival analyse wordt de tijd tot een bepaalde gebeurtenis plaatsvindt geanalyseerd. Deze gebeurtenis kan overlijden zijn, vandaar de naam survival analyse, maar het kan ook een besmetting, hartaanval of juist genezing zijn. De gegevens worden verzameld door het volgen van een groep patiënten voor een bepaalde periode. Tijdens deze periode kan het zijn dat bij een patiënt de gebeurtenis plaatsvindt, maar het is natuurlijk ook goed mogelijk dat dit tijdens de observatieperiode niet het geval is.

Idealiter is van elke patiënt de tijd waarop de gebeurtenis bij hem of haar plaatsvindt bekend. Als dit niet zo is dan is er sprake van *censoring*. Dit betekent dat de verkregen data niet compleet is. Het kan zo zijn dat de gebeurtenis helemaal niet plaats heeft gevonden tijdens de periode, maar het kan ook zo zijn dat het precieze tijdstip niet is waargenomen. Dit kan bijvoorbeeld zo zijn bij een besmetting. Een besmetting wordt (bijna) nooit precies op het moment van besmetting waargenomen, maar enige tijd later. In dit geval kun je dus wel zeggen in welke tijdsperiode de gebeurtenis heeft plaatsgevonden, maar niet het precieze tijdstip. Naast censoring is er nog een verschijnsel wat de schattingen van de survival analyse bemoeilijkt en dat is *transmissie*. Dit betekent dat patiënten elkaars toestand kunnen beïnvloeden. Een voorbeeld hiervan is dat een besmette patiënt zijn besmetting aan een gezonde patiënt kan doorgeven. Dit zou kunnen gebeuren als deze patiënten bij elkaar op de kamer komen te liggen.

In een model waarbij er geen sprake is van censoring en transmissie is het mogelijk om schattingen te doen met standaard modellen die gebaseerd zijn op counting processes. Het is dan namelijk bekend hoeveel mensen er op welk tijdstip er precies in een bepaalde toestand zitten en deze aantallen kunnen geteld worden. In een model zonder transmissie maar waar wel sprake is van censoring kunnen nog steeds standaardmodellen worden gebruikt. In dit geval is het niet mogelijk om precies te weten op welk tijdstip er hoeveel mensen in een bepaalde toestand zitten, maar is het wel mogelijk om

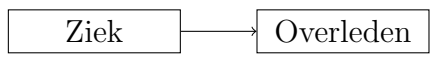
met verwachtingsmodellen hier een goede schatting van te maken. Zodra patiënten elkaars toestand kunnen gaan beïnvloeden is dit niet meer mogelijk. Er zijn nu te veel onzekerheden om precies te kunnen tellen hoeveel mensen er telkens in een bepaalde toestand zitten en het kost te veel werk en tijd om de benodigde integraties voor het schatten van de parameters uit te voeren. In deze gevallen is het handig om gebruik te maken van Bayesiaanse statistiek. In de Bayesiaanse statistiek zijn er namelijk numerieke methoden om dit soort problemen op te vangen en de parameters goed te kunnen schatten, zoals bijvoorbeeld *Markov Chain Monte Carlo*. Deze methode wordt gebruikt voor het berekenen van multi-dimensionale integralen, waar we bij de complexere modellen niet omheen kunnen.

In deze scriptie zullen we twee modellen bekijken zonder censoring en transmissie. Het doel is om te kijken wat voor effect Bayesiaanse schattingen hebben op de overgangssnelheden. Eerst bekijken we in hoofdstuk 2 een eenvoudig survival model met twee mogelijke toestanden. Op deze manier raken we gewend aan de Bayesiaanse methode. Vervolgens introduceren we in hoofdstuk 3 een multistate-model waarbij er bij patiënten meerdere gebeurtenissen achter elkaar kunnen plaatsvinden. Op deze manier kunnen we laten zien hoe de Bayesiaanse statistiek werkt als er meerdere gebeurtenissen tegelijk moeten worden geschat. In beide gevallen zouden we eenvoudige schattingen kunnen doen die gebaseerd zijn op counting processes, maar wij zijn juist geïnteresseerd in hoe de Bayesiaanse methode werkt bij de survival analyse.

2 Eenvoudig model met twee toestanden

2.1 Modelbeschrijving

Om kennis te maken met de toepassing van Bayesiaanse statistiek op de survival analyse kijken we eerst naar een relatief eenvoudig model. We nemen aan dat er geen sprake is van censoring en dat er geen transmissie plaatsvindt. We zullen de overgang van n patiënten bekijken van 'ziek' naar 'overleden'. Dit is geïllustreerd in Figuur 1.



Figuur 1: Model met twee toestanden.

Elke patiënt i heeft een eigen tijdstip waarop hij of zij zal overlijden, dit noemen we de *survivaltijd*. We definiëren de stochastische variabele T die deze survivaltijd weergeeft. We nemen aan dat T kansdichtheid $f(t)$ en kansverdeling $F(t) = P(T < t)$ heeft. De kansverdeling geeft de kans dat een gebeurtenis voor tijdstip t heeft plaatsgevonden. Het is makkelijker om te werken met de *survival functie* $S(t)$, het complement van de kansverdeling. Dit geeft de kans dat een gebeurtenis nog niet heeft plaatsgevonden op tijdstip t:

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(x)dx. \tag{1}$$

De verdeling van de survivaltijden kan ook worden gegeven met behulp van de hazard functie. Dit is een functie die aangeeft met welke snelheid, ook wel *rate* genoemd, een gebeurtenis zal plaatsvinden op een bepaald tijdstip. De hazard functie is gedefinieerd als

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

Hierbij is in de teller de voorwaardelijke kans te zien dat een gebeurtenis voorkomt in tijdsinterval $[t, t + dt)$, gegeven dat dit nog niet is gebeurd tot dan toe. Dit kunnen we ook schrijven als

$$\frac{P(t \leq T < t + dt \wedge T \geq t)}{P(T \geq t)} = \frac{P(t \leq T < t + dt)}{P(T \geq t)} = \frac{f(t)dt}{S(t)}.$$

Als we dit invullen in de definitie van de hazard functie en daarbij dus delen door dt en de limiet nemen dan krijgen we als resultaat

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{2}$$

We weten uit vergelijking (1) dat $f(t)$ de afgeleide is van $-S(t)$. We kunnen vergelijking (2) dus schrijven als

$$\lambda(t) = -\frac{d}{dt} \log S(t).$$

Nu kunnen we de survival functie als volgt uitdrukken in termen van de hazard functie:

$$S(t) = e^{-\int_0^t \lambda(x)dx}. \tag{3}$$

De uitdrukking in de exponent wordt de *cumulatieve hazard* genoemd; de som van alle risico's waar je mee te maken krijgt tot tijdstip t.

Het meest gebruikte model voor het analyseren van survival data is het *Cox Proportional-Hazards model*. Dit model onderzoekt de effecten van covariaten op de survivaltijden. Deze covariaten kunnen bijvoorbeeld leeftijd, geslacht, gewicht of medicijngebruik zijn. Om gebruik te kunnen maken van het Cox model moet er wel worden voldaan aan de voorwaarde dat de hazard ratio tijdsonafhankelijk is. Dit betekent dat als een patiënt op een bepaald tijdstip een twee keer zo groot risico heeft om te sterven dan een ander, dat deze patiënt op een ander tijdstip ook een risico heeft dat twee keer zo groot is als dat van de ander. We nemen in ons model aan dat aan deze voorwaarde is voldaan. Met behulp van het Cox model kunnen we de hazard functie gaan schatten. De hazard functie kan als volgt worden geschat:

$$\lambda(t) = \lambda_0(t) \cdot e^{\beta X}.$$

Hierin is X de matrix met covariaten, β de vector met effectgroottes en $\lambda_0(t)$ de nulmeting, dit noemen we de *baseline hazard*. Normaal gesproken wordt het Cox model gebruikt om de waarde van β te gaan bepalen. Wij nemen echter aan dat deze gegeven zijn, want wij zijn namelijk geïnteresseerd in het effect van het schatten van parameters van de baseline hazard $\lambda_0(t)$.

Bij het schatten van de parameters gaan we Bayesiaanse statistiek toepassen. We gebruiken eerst onze voorkennis over de verzamelde data voor het kiezen van de parameters van de baseline hazard, dit is de *prior*. Vervolgens stellen we onze voorkennis bij met behulp van de *likelihood* zodat we de *posterior* krijgen. De posterior is de ge-update schatting van onze parameter. Dit doen we met behulp van de theorie van Bayes:

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}. \tag{4}$$

Volgens Bayes is de likelihood vermenigvuldigd met de prior in verhouding met de posterior. De likelihood functie geeft aan hoe aannemelijk de data is voor gegeven parameters. De likelihood is dus een functie van de parameters. Omdat we er in dit model vanuit gaan dat er geen sprake is van censoring en alle patiënten onafhankelijk zijn van elkaar, kunnen we de likelihood van patiënt i definiëren als $L_i = f(t_i)$. Hier is t_i de survivaltijd van patiënt i. De algehele likelihoodfunctie wordt dan

$$L = \prod_{i=1}^{n} f(t_i), \tag{5}$$

waarin L en $f(t_i)$ functies zijn die afhankelijk zijn van gekozen parameters. Het product wordt genomen over alle geobserveerde patiënten. Uit vergelijking (2) volgt dat $f(t) = \lambda(t)S(t)$ dus we kunnen de likelihood schrijven als

$$L = \prod_{i=1}^{n} \lambda(t_i)S(t_i) = \prod_{i=1}^{n} e^{\int_0^{t_i} \lambda(x)dx}\lambda(t_i).$$

De baseline hazard kan verschillende vormen hebben. We zullen een constante hazard rate en een stuksgewijs constante hazard rate gaan bekijken. Een stuksgewijs constante hazard rate is een vaak gebruikte aanname in de survival analyse. Bij een ziekte van een aantal jaren is de kans op sterven op een bepaalde dag meestal niet extreem anders dan op een dag later. We nemen aan dat deze hazard rate dan constant is op dit tijdsinterval.
We zullen data genereren met behulp van de exponentiële verdeling en de Weibull verdeling. De Weibull verdeling wordt vaak gebruikt bij het modelleren van levensduur. Als het risico op overlijden afneemt in de tijd dan kiest men de shape-parameter kleiner dan 1. Hierdoor neemt de dichtheid af. Als het risico juist toeneemt in de tijd dan kiest men deze juist groter dan 1,

waardoor de dichtheid toeneemt. Als deze parameter gelijk is aan 1 dan is de Weibull verdeling gelijk aan de exponentiële verdeling en dit betekent dat het risico op overlijden constant is over tijd.

Voor elke dataset gaan we de hazard functie schatten met behulp van de hiervoor uitgelegde methode. We vergelijken de resultaten van verschillende priors met elkaar om te kunnen zeggen welke priors beter werken in de verschillende situaties. Een prior werkt beter dan een andere als het een grotere kans op een goede uitkomst geeft. We maken onder andere gebruik van normaal verdeelde en exponentiële priors waarbij we de verwachte waarde baseren op de werkelijke waarde. Hierdoor zijn deze priors zeer informatief. Door gebruik te maken van een informatieve prior wordt de posterior met name bepaald door de prior. Daarnaast zullen we ook een niet-informatieve exponentiële prior meenemen waarbij de verdeling bijna vlak is. Ook zullen we een prior van de vorm $1/x$ gebruiken. Dit is een oneigenlijke prior wat betekent dat de integraal oneindig is en het dus geen kansdichtheid is. Oneigenlijke priors worden vaak gebruikt als niet-informatieve priors. De posterior wordt dan met name bepaald door de likelihood omdat de prior aan elke mogelijke waarde dezelfde kans geeft. Een verdeling van de vorm $1/x$ heeft een sterke voorkeur voor kleinere waarden. Om deze reden zullen de parameters met deze prior waarschijnlijk lager worden geschat dan voor andere priors.

2.2 Resultaten

We gaan de modellen eerst uitvoeren op een dataset van 100 patiënten om te kijken of de priors inderdaad werken zoals we verwachten. Vervolgens bekijken we ook de situatie waarin we gegevens hebben van maar 10 patiënten, zodat we het effect van de prior op de posterior beter kunnen zien. Dit effect is immers groter bij een kleinere hoeveelheid gegevens.

2.2.1 Constante baseline hazard

We beginnen met een constante baseline hazard. Dit betekent dat we te maken hebben met $\lambda_0 = c$ voor een $c \in \mathbb{R}_+$. We concluderen uit (1) en (2) dat voor een constante hazard geldt dat $f(t) = c \cdot e^{-ct}$. We genereren hiervoor dus een dataset met behulp van de exponentiële verdeling. We kiezen hiervoor waarde $c = 1$.

Een constante hazard rate betekent dat de likelihood voor patiënt i er als volgt uit ziet:

$$L_i = c \cdot e^{-ct_i}.$$

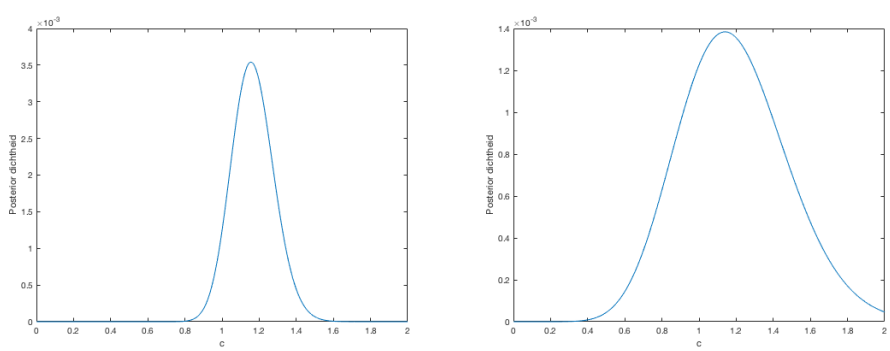
De algehele likelihood is het product over alle patiënten:

$$L = \prod_{i=1}^{n} c \cdot e^{-ct_i}.$$

Normaal verdeelde priors

We beginnen met een normaal verdeelde prior $pr(c)$ voor c met verwachtingswaarde 1 en standaardafwijking 1. De verwachtingswaarde hebben we gebaseerd op de werkelijke waarde. Dit betekent dat dit een zeer goede informatieve prior verdeling is. We merken op dat c niet negatief kan zijn en hier een probleem kan ontstaan bij het gebruiken van de normaalverdeling. Om deze reden maken we gebruik van de *truncated normaalverdeling*. Dit betekent dat de normaalverdeling afgeknot wordt op $x = 0$ en het overgebleven stuk opnieuw wordt genormaliseerd.

$$pr(c) \sim \mathcal{N}(1,\, 1)$$



(a) Dataset van 100 patiënten (b) Dataset van 10 patiënten

Figuur 2: De posterior verdelingen bij een constante baseline hazard met normaal verdeelde prior voor verschillende hoeveelheden data.

Met behulp van Bayes' theorie uit (4) berekenen we de verdeling van de posterior. Dit doen we door de likelihood te vermenigvuldigen met de prior en dit vervolgens te normaliseren. In Figuur 2a is de verdeling van de posterior te zien bij een dataset van 100 patiënten. Er is te zien dat de gemiddelde waarde iets groter is dan de werkelijke waarde 1. Dit betekent voor de hazard dat de snelheid waarmee een patiënt zal overlijden iets hoger wordt geschat dan dat deze in werkelijkheid is. De standaardafwijking is ook relatief klein als je het vergelijkt met de posterior verdeling in Figuur 2b. Dit is de verdeling van de posterior bij een dataset van 10 patiënten. Omdat er hier minder data is waar de likelihood op gebaseerd is, is het effect van de vermenigvuldiging met de prior beter zichtbaar. Er is te zien dat de verwachte waarde nu verder van de werkelijke waarde af ligt en ook de standaardafwijking groter is. Er kan dus met iets minder zekerheid worden gezegd dat de gemiddelde waarde ook echt de werkelijke waarde is.

In werkelijkheid zal het vaak niet mogelijk zijn om zo'n goed informatieve prior te gebruiken. Er is immers niet zo veel voorkennis dat de werkelijke waarde al meteen duidelijk is, anders zou het doen van schattingen overbodig zijn. We doen daarom ook schattingen met behulp van een normaal verdeelde prior die foutieve informatie bevat. We nemen hiervoor weer de truncated normale prior, maar dit keer met een willekeurige verwachte waarde die een stuk verder van de werkelijke waarde af ligt.

$$pr(c) \sim \mathcal{N}(5,\, 1)$$

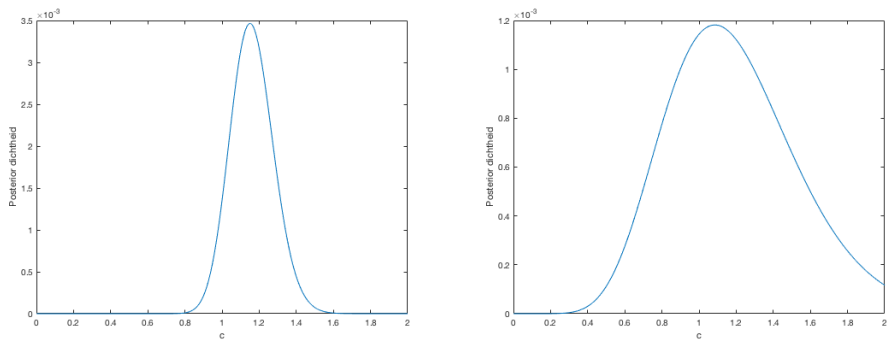
We maken gebruik van dezelfde dataset van 10 patiënten als in het vorige geval. Om de resultaten in dit geval te vergelijken met de resultaten van de goede informatieve prior bekijken we de gemiddelde gekwadrateerde fout van de resultaten van de kleine dataset, dit noemt men ook wel de *mean squared error*. We voeren de berekeningen 100 keer uit en berekenen het verschil tussen de maximale waarde van de verdeling en de werkelijke waarde. Dit noemen we de fout. Vervolgens berekenen we het gemiddelde van al deze gekwadrateerde fouten. Voor de goede informatieve prior vinden we een zeer kleine mean squared error van 0.0429. Voor de foutieve informatieve prior vinden we meteen een stuk grotere mean squared error van 0.1136. Dit betekent dat de normale verdeling goed werkt als er goede voorkennis is, maar zodra dit niet het geval is deze al snel een stuk minder goede resultaten geeft.

Exponentieel verdeelde priors

Vervolgens kiezen we een exponentieel verdeelde prior voor c met verwachtingswaarde 1. Ook in dit geval bekijken we dus eerst de resultaten voor een goede informatieve prior.

$$pr(c) \sim \exp(1)$$

We verkrijgen de posterior met behulp van theorie (4). De resulaten voor een dataset van 100 patiënten zijn te zien in Figuur 3a. Hier is dezelfde dataset gebruikt als bij de resultaten voor de normaal verdeelde priors. De resultaten van de kleinere dataset, ook dezelfde als in het eerste geval, zijn te zien in Figuur 3b. Beide verdelingen lijken niet veel af te wijken van de verdelingen die met de goed informatieve normaal verdeelde priors zijn berekend.



(a) Dataset van 100 patiënten (b) Dataset van 10 patiënten

Figuur 3: De posterior verdelingen bij een constante baseline hazard met exponentieel verdeelde prior voor verschillende hoeveelheden data.

Om toch te kunnen zeggen welke prior een beter resultaat geeft zullen we ook voor deze prior de mean squared error berekenen. Voor de exponentieel verdeelde prior vinden we een mean squared error van 0.0798. Dit betekent dat deze prior over het algemeen minder goede resultaten geeft dan de goed informatieve normaal verdeelde prior.

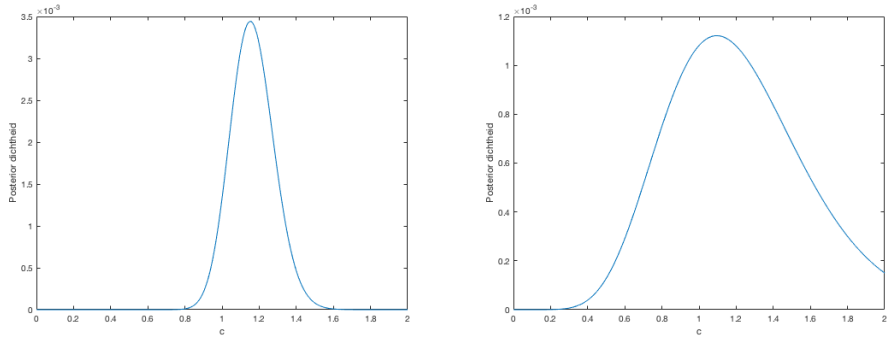
We bekijken ook nog de resultaten voor een niet-informatieve exponentieel verdeelde prior met een heel kleine verwachtingswaarde. Dit betekent voor de verdeling dat deze bijna vlak is en er dus voor elke mogelijke waarde een ongeveer gelijke kans is.

$$pr(c) \sim \exp(0.0001)$$

We berekenen de posterior door de likelihood te vermenigvuldigen met deze prior. We maken in dit geval weer gebruik van dezelfde dataset van 10 patiënten. We vinden voor deze situatie een mean squared error van 0.1013. Bij weinig goede voorkennis geeft deze exponentiële prior betere resultaten dan de normaal verdeelde prior. Dit is te verklaren door het feit dat er in dit geval een bijna gelijke kans is toegekend aan elke mogelijke waarde voor c. Dit is bij een normaal verdeling niet het geval.

Prior van de vorm $1/x$

Als laatste proberen we nu ook nog de oneigenlijke prior van de vorm $1/x$. Omdat deze prior aan elke mogelijke waarde voor c dezelfde kans toekent verwachten we dat deze soortgelijke resultaten zal geven als de prior met de vlakke exponentiële verdeling. De resultaten voor dezelfde datasets als in de vorige gevallen zijn te zien in Figuur 4a en Figuur 4b. We berekenen voor deze situatie een mean squared error van 0.0968. Dit betekent dat een prior van de vorm $1/x$ over het algemeen nog betere resultaten geeft dan de vlakke exponentiële prior. Dit is te verklaren door het feit dat elke mogelijke waarde van c nu precies dezelfde kans heeft.



(a) Dataset van 100 patiënten (b) Dataset van 10 patiënten

Figuur 4: De posterior verdelingen bij een constante baseline hazard met prior van de vorm $1/x$ voor verschillende hoeveelheden data.

2.2.2 Stuksgewijs constante baseline hazard

Als tweede nemen we aan dat de baseline hazard een stuksgewijs constante functie is op de intervallen $[0, t^*)$ en $[t^*, \infty)$. Deze is dus van de vorm

$$\lambda_0(t) = \begin{cases} c_1 & t < t^* \\ c_2 & t \geq t^* \end{cases}$$

met $c_1, c_2 \in \mathbb{R}_+$ en $t^* > 0$. We genereren hiervoor een dataset met behulp van de exponentiële verdeling voor waarden $c_1 = 1$ en $c_2 = 10$. Dit betekent dat het risico om te overlijden tot tijdstip 1 relatief klein is ten opzichte van het risico om te overlijden na dit tijdstip. We kiezen hiervoor omdat we door dit grote verschil wellicht een verschil kunnen opmerken in de resultaten. We merken op dat het gemiddelde van de gegenereerde overlevingstijden rond de waarde 1 ligt. Daarom kiezen we $t^* = 1$. Dit betekent dat de likelihood voor patiënt i er als volgt uit ziet:

$$L_i = \begin{cases} e^{-c_1 t_i} \cdot c_1 & t_i < 1 \\ e^{-c_1} \cdot e^{-c_2(t_i - 1)} \cdot c_2 & t_i \geq 1 \end{cases}$$

De algehele likelihood is weer het product over alle patiënten zoals gedefinieerd in (5).

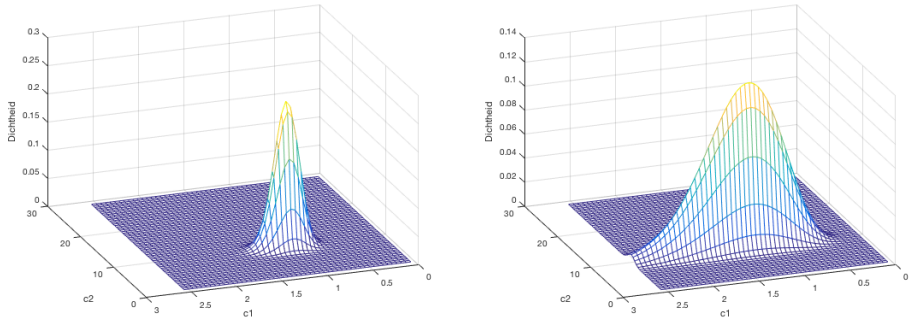
Normaal verdeelde prior

Voor deze situatie beginnen we ook met truncated normaal verdeelde priors $pr(c_1)$ en $pr(c_2)$ met verwachtingswaarden respectievelijk 1 en 10. Deze waarden zijn gebaseerd op onze werkelijk gekozen waarden voor de parameters, dit zijn dus ook zeer goede informatieve priors. Als standaardafwijking kiezen we 1.

$$pr(c_1) \sim \mathcal{N}(1,\, 1)$$
$$pr(c_2) \sim \mathcal{N}(10,\, 1)$$

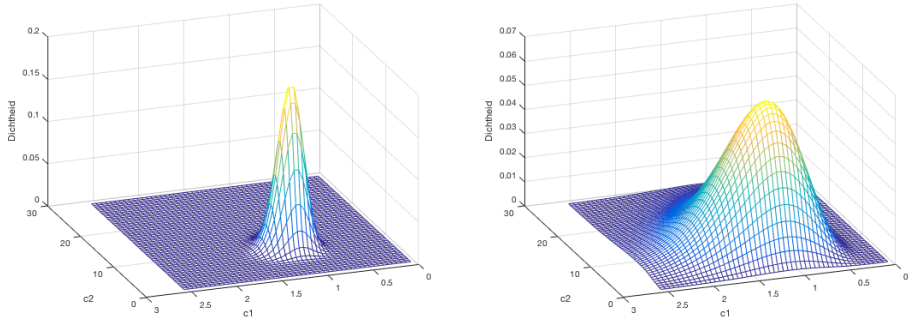
Gebruikmakend van theorie (4) verkrijgen we de dichtheid van de posterior. In Figuur 5a is het resultaat te zien voor een dataset van 100 patiënten. We zien een mooie verdeling van beide posteriors waarbij de gemiddelde waarden van c_1 en c_2 inderdaad dichtbij de werkelijke waarden, respectievelijk 1 en 10, zitten. In Figuur 5b zijn de resultaten te zien voor een dataset van 10 patiënten. De schatting voor c_2 blijkt nog steeds heel precies te zijn met

een kleine standaardafwijking. Voor c_1 is er een gemiddelde waarde die iets links ligt van de werkelijke waarde maar door de grote standaardafwijking nog steeds een grote kans heeft om wel een goede waarde aan te nemen. Voor deze situatie vergelijken we niet ook nog foutieve informatieve normale priors omdat we in de situatie met de constante baseline hazard al hebben gezien dat dit zoals verwacht een stuk minder goede resulaten oplevert.



(a) Dataset van 100 patiënten. (b) Dataset van 10 patiënten.

Figuur 5: De posterior verdelingen bij een stuksgewijs constante baseline hazard met normaal verdeelde priors voor verschillende hoeveelheden data.



(a) Dataset van 100 patiënten. (b) Dataset van 10 patiënten.

Figuur 6: De posterior verdelingen bij een stuksgewijs constante baseline hazard met exponentieel verdeelde priors.

Exponentieel verdeelde priors

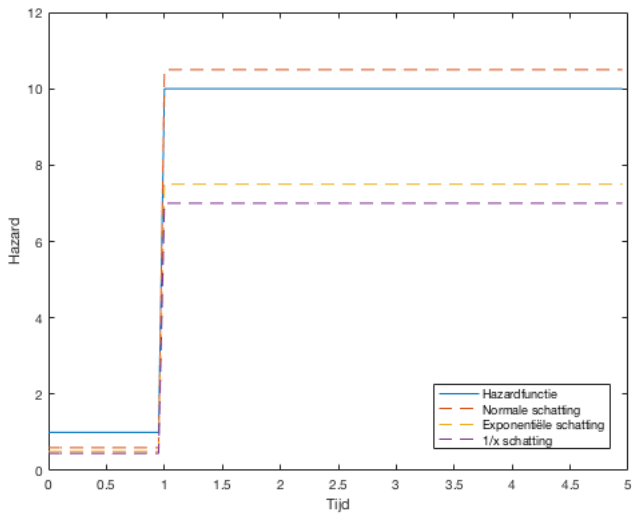
Vervolgens proberen we exponentieel verdeelde priors $pr(c_1)$ en $pr(c_2)$ met dezelfde verwachtingswaarden 1 en 10. Dit zijn dus ook weer goede informatieve priors.

$$pr(c_1) \sim \exp(1)$$
$$pr(c_2) \sim \exp(10)$$

Ook voor deze priors verkrijgen we de dichtheid voor de posterior door middel van theorie (4) van Bayes. Voor dezelfde datasets als bij de normaal verdeelde priors zijn de resultaten te zien in Figuur 6a en Figuur 6b. We zien dat de verdeling van de posteriors anders is dan bij de normaal verdeelde prior, vooral de betrouwbaarheid voor de waarde van c_2 is een stuk kleiner. We kunnen niet direct conclusies trekken over welke prior betere resultaten geeft. Om te kijken of de exponentiële prior beter of minder goed werkt dan een normale prior vergelijken we de mean squared errors met elkaar. We berekenen voor de normaal verdeelde prior mean squared errors van 0.0166 voor c_1 en 0.4200 voor c_2. Voor de informatieve exponentiële prior berekenen we voor c_1 en c_2 mean squared errors van 0.0161 en 2.6250. We merken op dat de mean squared error voor c_2 een stuk groter is dan andere berekende fouten. Dit kan te verklaren zijn door te kijken in de likelihood functie. Bij een kleine fout voor beide constanten is te zien dat de fout voor c_1 wordt vermenigvuldigd met een $t_i < 1$ terwijl de fout voor c_2 wordt vermenigvuldigd met $t_i - 1$. Zodra $t_i > 2$ is de factor waarmee deze fout wordt vermenigvulgd dus groter waardoor er ook een grotere gemiddelde gekwadrateerde fout uit komt. Het vergelijken van de resultaten met de eerdere resultaten laat zien dat de normale prior gemiddeld een betere schatting doet voor de parameters. Dit kunnen we ook zien in Figuur 7. De werkelijke hazard functie en de geschatte hazard functies op basis van de gevonden schattingen voor de parameters zijn geplot. Er is duidelijk te zien dat de hazard functie op basis van de goed informatieve normale prior een betere schatting geeft voor de werkelijke hazard functie. Voor beide priors die we nu hebben vergeleken is een hele goede voorkennis nodig, dit is in werkelijkheid bijna nooit het geval. We zullen nu dus ook minder informatieve priors gaan bekijken.

Net als bij de constante baseline hazard bekijken we de resultaten voor vlak exponentieel verdeelde priors. We nemen voor de priors voor beide constanten een verwachte waarde van 0.0001. We berekenen de dichtheid van de posterior door de likelihood met de priors te vermenigvuldigen. Voor beide

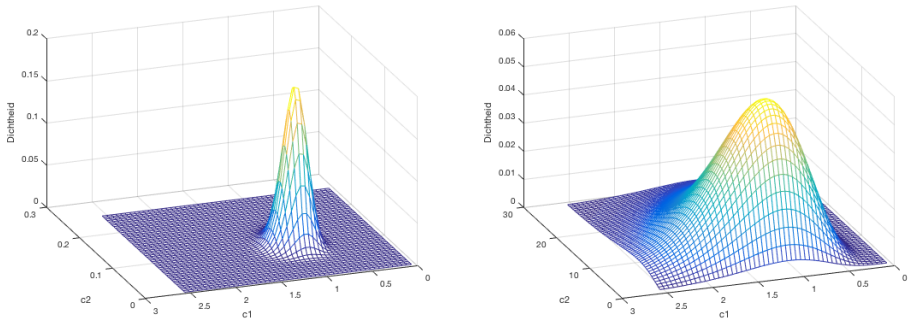
constanten vergelijken we de mean squared errors met de vorige situaties. Voor c_1 en c_2 berekenen we mean squared errors van 0.0173 en 2.9675. Dit geeft dus over het algemeen iets minder goede resultaten dan de informatieve normaal en exponentieel verdeelde priors.



Figuur 7: De werkelijke hazard functie en de hazard functies op basis van de schattingen met informatieve normaal verdeelde priors, informatieve exponentieel verdeelde priors en priors van de vorm 1/x.

Prior van de vorm $1/x$

Als laatste nemen we priors van de vorm $1/x$. We vermenigvuldigen deze met de likelihood en verkrijgen op deze manier de posterior verdeling. De resultaten voor een grote en een kleine dataset zijn te zien in Figuur 8a en Figuur 8b. Ondanks het feit dat we niet-informatieve priors gebruiken merken we op dat de verdelingen sterk lijken op de verdelingen met de informatieve exponentiële priors. Om te kijken of deze ook echt even goed werken berekenen we ook de mean squared error voor deze situatie. Dit levert ons een mean squared error op van 0.0164 voor c_1 en van 2.8350 voor c_2. Priors van de vorm $1/x$ geven dus betere resultaten dan niet-informatieve exponentiële priors. Dit is te verklaren door het feit dat nu elke mogelijke waarde voor de



(a) Dataset van 100 patiënten. (b) Dataset van 10 patiënten.

Figuur 8: De posterior verdelingen bij een stuksgewijs constante baseline hazard en priors van de vorm $1/x$.

constanten een even grote kans krijgt. De hazard rate op basis van de prior van de vorm $1/x$ is ook te zien in Figuur 7.

2.2.3 Weibull verdeling

Tot nu toe hebben we steeds een dataset gegenereerd met een exponentiële verdeling. Een vaker gebruikte manier voor het modelleren van survival-data is echter de Weibull verdeling. Deze zullen we nu gaan gebruiken. We zullen hiervoor echter geen parametrische schattingen gaan doen, maar semi-parametrische. Dit betekent dat we niet de parameters van de Weibull verdeling gaan terugschatten, maar dat we aannemen dat de hazard functie stuksgewijs constant is. We gaan dus een parametrische functie benaderen met een stuksgewijs constante functie. Om het voor nu nog eenvoudig te houden kiezen we voor een stuksgewijs constante functie op twee intervallen. Het zou natuurlijk realistischer zijn om de tijd in meerdere intervallen te verdelen. We genereren eerst een dataset van 10 patiënten met behulp van de Weibull verdeling. Dit betekent dat de hazard functie van de gegenereerde data van de vorm

$$\lambda(t) = \frac{k}{\theta}\left(\frac{t}{\theta}\right)^{k-1}$$

is. We kiezen shape-parameter $k = 1.5$ en scale-parameter $\theta = 1$. De waarde van k geeft aan dat het risico op overlijden toeneemt in de tijd. We merken op dat het gemiddelde van de gegenereerde overlevingstijden rond de

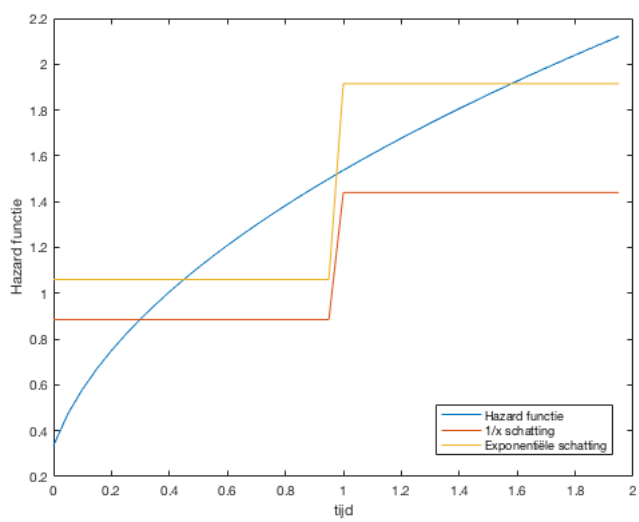
waarde 1 ligt. We kiezen dus bij het schatten van de hazard functie voor een stuksgewijs constante functie op tijdsintervallen $[0, 1)$ en $[1, \infty)$. Omdat we aannemen dat de hazard functie van de data stuksgewijs constant is, ziet de likelihood voor patiënt i er als volgt uit:

$$L_i = \begin{cases} e^{-c_1 t_i} \cdot c_1 & t_i < 1 \\ e^{-c_1} \cdot e^{-c_2(t_i - 1)} \cdot c_2 & t_i \geq 1 \end{cases}$$

De algehele likelihood is het product over alle 10 patiënten. Omdat te goede informatieve normale en exponentiële verdelingen niet zo realistisch zijn zullen we deze in dit geval niet meer meenemen. We kiezen voor het schatten van de constanten c_1 en c_2 de niet-informatieve exponentiële priors met verwachte waarde 0.0001 en de priors van de vorm $1/x$.

Niet-informatieve priors

De resultaten voor de dichtheden van de posteriors voor een dataset van 10 patiënten zijn te zien in Figuur 9. Om deze te krijgen hebben we weer gebruik gemaakt van theorie (4). We merken op dat de prior van de vorm $1/x$ lagere



Figuur 9: Schattingen van de Weibull hazard functie met stuksgewijs constante hazard functies met niet-informatieve exponentiële priors en priors van de vorm $1/x$.

schattingen geeft dan de exponentiële prior. Zoals we eerder al bespraken heeft een prior van de vorm $1/x$ meer voorkeur voor kleinere waarden. Dit is dus ook duidelijk te zien in de resultaten. Om een conclusie te kunnen trekken over welke prior over het algemeen de beste schattingen geeft vergelijken we mean squared errors met elkaar. In dit geval hebben we de mean squared error berekend door de waarde van de stuksgewijze schatting om de 0.05 tijdseenheden te vergelijken met de werkelijke hazard functie. Deze verschillen hebben we gekwadrateerd en hiervan hebben we het gemiddelde genomen. Dit hebben we voor beide priors voor dezelfde 100 datasets berekend. Het gemiddelde van al deze berekeningen noemen we de mean squared error. We vinden voor voor de prior van de vorm $1/x$ een mean squared error van 0.6058. Voor de exponentiële prior vinden we een mean squared error van 0.4925. Hieruit kunnen we dus concluderen dat de een niet-informatieve exponentiële prior met een zeer kleine verwachte waarde over het algemeen de beste resultaten geeft.

2.3 Discussie

We hebben gezien dat er bij de verschillende hazard rates voor verschillende priors wisselende resultaten uit zijn gekomen. Door middel van de gemiddelde kwadratische fouten hebben we gekeken welke priors de beste resultaten gaven. We kunnen concluderen dat een prior met een normaalverdeling gecentreerd rondom de werkelijke waarde de beste resultaten geeft. Dit is echter geen realistische prior omdat hiervoor zeer goede voorkennis nodig is. Om deze reden hebben we daarom ook minder informatieve priors met elkaar vergeleken. De verschillen tussen de exponentiële prior en de prior van de vorm $1/x$ zijn niet heel groot. In het geval dat we een stuksgewijze hazard functie ook stuksgewijs gingen terugschatten gaf een prior van de vorm $1/x$ de beste resultaten. Bij het schatten van een Weibull hazard kunnen concluderen dat een niet-informatieve exponentiële prior het algemeen de beste resultaten geeft. In dit geval heeft elke mogelijke waarde voor een parameter ongeveer dezelfde kans.

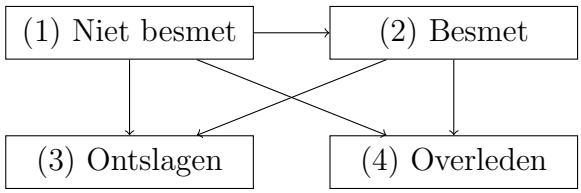
Het survival model dat we hier hebben opgesteld is een zeer vereenvoudigd model van de werkelijkheid. We nemen aan dat er geen sprake is van censoring. We gaan er dus vanuit dat alle patiënten binnen de observatietijd zullen overlijden en dat we dit op het precieze tijdstip waarnemen. In werkelijkheid zullen er altijd gecensoreerde observaties zijn en deze observaties

hebben effect op de likelihood functie. Dit heeft uiteraard ook effect op de schattingen van de parameters. Ook gaan we ervan uit dat patiënten elkaar onderling niet kunnen beïnvloeden. Dit betekent dat alle patiënten onafhankelijk zijn van elkaar. Dit is ook geen realistische aanname omdat patiënten in een ziekenhuis vaak samen op een kamer liggen of op andere manieren met elkaar in aanraking komen.

3 Multistate model

3.1 Modelbeschrijving

We zullen nu een model gaan bekijken waarbij er meer dan twee toestanden zijn waar de patiënten zich in kunnen bevinden. We bekijken de overgangen van n patiënten die zich in 4 verschillende toestanden kunnen bevinden: niet besmet en in het ziekenhuis, besmet en in het ziekenhuis, ontslagen uit het ziekenhuis of overleden. Deze situatie is geschetst in Figuur 10.



Figuur 10: Multistate model met 4 toestanden.

De kans dat een persoon van de ene toestand naar een andere toestand gaat hangt nu niet alleen maar van deze twee toestanden af, maar ook van de overige toestanden. Om deze situatie te modelleren definiëren we $S(t)$ nu als de toestand waarin een individu zich bevindt op tijdstip t. Voor elke overgang (v, w) met $v, w \in \{1, 2, 3, 4\}$ en $v \neq w$ definiëren we de overgangsintensiteit

$$q_{vw}(t) = \lim_{dt \to 0} \frac{P(S(t + dt) = w | S(t) = v)}{dt}.$$

Vervolgens definiëren we $q_{vv} = -\sum_{w \neq v} q_{vw}$. De overgangsintensiteit, deze noemden we eerder al de *hazard*, is de snelheid waarmee een patiënt overgaat naar een andere toestand. Al deze overgangsintensiteiten kunnen we samenvoegen in een overgangsmatrix, oftwel de *transition rate matrix*. De overgangsmatrix voor ons model ziet er als volgt uit:

$$Q(t) = \begin{pmatrix} -(q_{12} + q_{13} + q_{14}) & q_{12} & q_{13} & q_{14} \\ 0 & -(q_{23} + q_{24}) & q_{23} & q_{24} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

We definiëren voor patiënt i de overgangstijd $t_{i,vw}$ van toestand v naar w. Met deze gegevens kunnen we de likelihood opstellen voor patiënt i voor een

overgang (v, w):

$$L_i = e^{q_{vv} t_{i,vw}} \cdot q_{vw} \tag{6}$$

De totale likelihood van het model is het product over alle overgangen en alle patiënten.

Bij een simpele likelihood is de posterior makkelijk te berekenen met behulp van de theorie van Bayes in (4). Dit hebben we gezien in de situatie met twee toestanden. Nu hebben we echter een complexere likelihood die afhangt van meerdere parameters door de verschillende overgangsintensiteiten. Als we ook nog censoring en transmissie zouden toevoegen aan ons model dan zou de likelihood alleen maar complexer worden. Het normaliseren van de vermenigvuldiging van de likelihood en de priors kan nu niet meer door het uitrekenen van een enkele of dubbele integraal. Door de vele parameters moet dit genormaliseerd worden met een multidimensionale integraal. In dit geval kunnen we gebruik maken van de *Markov Chain Monte Carlo* (MCMC) methoden om de posterior te berekenen. Deze methoden kunnen multidimensionale integralen berekenen door middel van sampling. Bij MCMC methoden worden er verschillende sets van parameterwaarden gebruikt, dit zijn de *samples*. Dit is het 'Monte Carlo' deel van MCMC. Voor een eerste sample wordt uitgerekend hoeveel van de data wordt verklaard door deze sample, gegeven de prior verdelingen. Vervolgens wordt er in de buurt van deze sample een willekeurige volgende set parameters gekozen waarvoor ook wordt uitgerekend hoe goed deze de data verklaard. Als hieruit blijkt dat deze sample de data beter verklaard dan de voorgaande, dan zal deze worden toegelaten met een bepaalde kans. Deze kans is afhankelijk van hoe veel beter de nieuwe sample is. Het is belangrijk om een goed gebied te definiëren waaruit een volgende sample wordt gekozen. Als dit gebied namelijk te klein wordt gekozen dan zal het heel erg lang duren voordat er een goede schatting uitkomt. Bij een te groot gebied is de kans juist heel groot dat er na het in de buurt komen van een goede waarde weer uit het gebied wordt weggegaan en het erg lang duurt voordat er een update wordt toegelaten. De samples die toegelaten worden vormen samen de *chain*. Dit is het 'Markov Chain' deel van MCMC. Elke sample in de chain is afhankelijk van de laatste sample, maar niet van alle voorgaande samples. De chain zal uiteindelijk convergeren naar een gebied rondom een bepaalde waarde. Als dit gebied wordt geplot in een histogram kan hier een goede schatting voor de posterior verdeling uit worden gehaald.

We gaan deze methode toepassen op ons multistate model. Omdat we eerder al hebben laten zien hoe dit voor minder realistische exponentiële en stuksgewijs exponentiële data in zijn werk gaat zullen we deze situaties nu overslaan. We gaan alleen werken met de Weibull verdeling omdat deze verdeling een realistische weergave is van survival data. We maken gebruik van priors van de vorm $1/x$ voor alle parameters, omdat deze in het kleinere model goed bleken te werken als de werkelijke waarden van te voren niet bekend zijn.

3.2 Resultaten

We genereren eerst een dataset van 10 patiënten met behulp van de Weibull verdeling. Voor elke overgang q_{vw} kiezen we andere waarden voor θ en k. Voor de overgangen q_{13} en q_{24} willen we dat de snelheid op een overgang toeneemt in de tijd. Het risico om ontslagen te worden wordt immers veel groter als je langer niet besmet bent. Hetzelfde geldt voor het overlijden bij een besmetting; des te langer je besmet bent, des te sneller je zult te komen overlijden. We kiezen voor deze overgangen dus een $k < 1$, de rest van de waarden kiezen we willekeurig. In Tabel 1 zijn de gekozen waarden voor θ en k te zien. We gaan, net als in het eenvoudige model met twee toestanden, de overgangsfuncties schatten als stuksgewijs constante functies. Omdat de gemiddelden van de gegenereerde datasets rond de waarde 1 liggen nemen we aan dat de overgangsfuncties stuksgewijs constant zijn op de intervallen $[0, 1)$ en $[1, \infty)$ voor constanten $c_{vw,1}, c_{vw,2} \in \mathbb{R}_+$. Dit betekent dat de likelihood voor patiënt i voor de overgang van toestand v naar w er als volgt uit ziet.

	k	θ
q_{12}	1.5	1
q_{13}	0.5	1
q_{14}	1.1	1.2
q_{23}	1.7	1.3
q_{24}	0.8	1.2

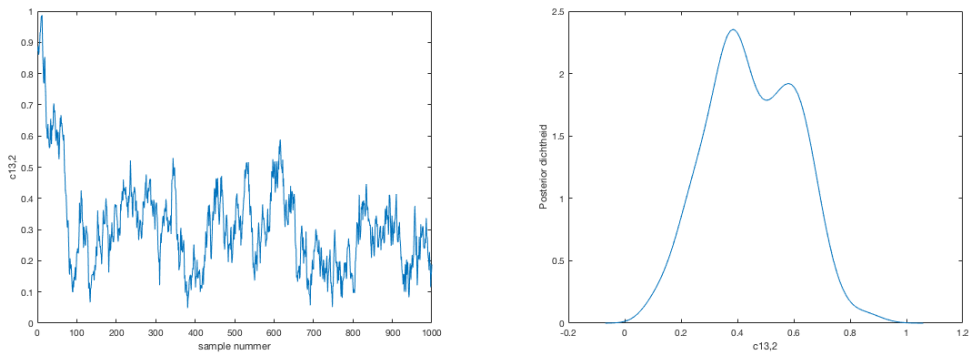
Tabel 1: Parameters k en θ

$$L_{vw,i} = \begin{cases} e^{-c_{vw,1}t_i} \cdot c_{vw,1} & t_i < 1 \\ e^{-c_{vw,1}} \cdot e^{-c_{vw,2}(t_i-1)} \cdot c_{vw,2} & t_i \geq 1 \end{cases}$$

De totale likelihood is het product over alle overgangen en alle patiënten en is dus een functie van alle tien parameters $c_{vw,i}$:

$$L = \prod_{i=1}^{10} \prod_{(v,w)} L_{vw,i}.$$

We gaan nu gebruik maken van de MCMC methode om goede schattingen voor al deze parameters te doen. In Figuur 11a is de Markov Chain voor constante $c_{13,2}$ te zien. Er is een convergentie te zien naar een gebied rond de 0.3. Voor het plotten van de posterior zijn de eerste 100 samples niet interessant, hier is de chain nog bezig met zoeken naar het goede gebied. Het plotten van de waarden voor de samples 100 tot en met 1000 geeft uiteindelijk de verdeling in Figuur 11b.



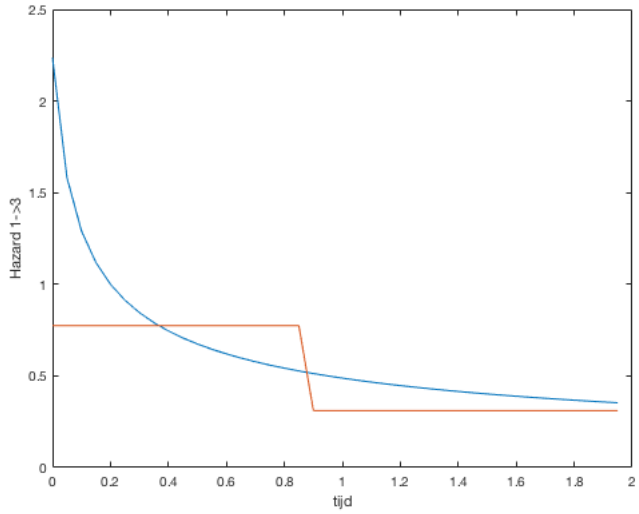
(a) De Markov Chain van constante $c_{13,2}$.

(b) De dichtheid van de posterior voor $c_{13,2}$.

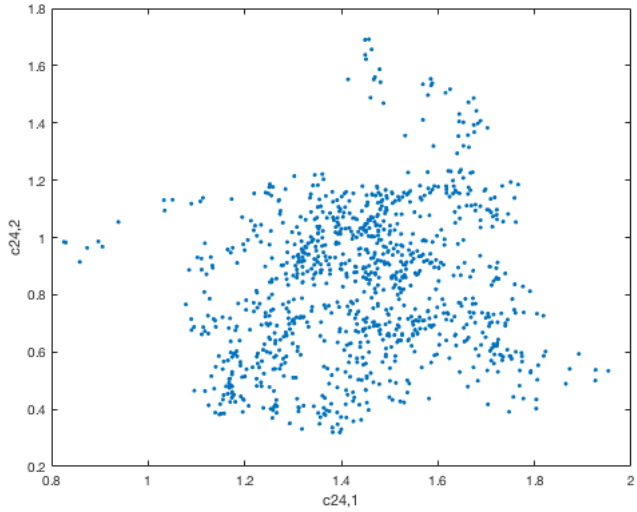
Figuur 11: MCMC voor de tweede constante $c_{13,2}$ van de overgang (1) naar (3).

We kunnen voor elke constante de gemiddelde waarde vinden en hiermee de hazard functie plotten. Voor q_{13} is het resultaat te zien in Figuur 12. Het valt op dat de waarden van de constanten vrij laag worden geschat. Hier hebben we in de eerdere resultaten al een verklaring voor gegeven. Een verdeling van de vorm $1/x$ heeft namelijk voorkeur voor kleinere waarden.

Het is ook interessant om te kijken naar de correlatie tussen verschillende parameters. Is het bijvoorbeeld zo dat als het risico om te overlijden bij een besmetting groot is in de eerste periode, dat dit ook zo is in de tweede periode? Of dat het risico in de tweede periode juist klein is? Hiervoor kijken we naar de gezamenlijke samples van $c_{24,1}$ en $c_{24,2}$. Dit is te zien in Figuur 13. We concluderen hieruit dat er geen correlatie is tussen deze twee parameters. Dit resultaat zullen we bespreken in de discussie.



Figuur 12: De Weibull en de stuksgewijs constante hazard functie voor q_{13}.



Figuur 13: De gezamenlijke samples van $c_{24,1}$ en $c_{24,2}$, duidelijk niet gecorreleerd.

3.3 Discussie

We hebben gezien dat we met behulp van MCMC methoden de overgangsintensiteiten voor een multistate model goed kunnen schatten.
We hebben echter wel een zeer vereenvoudigd model van de werkelijkheid gebruikt. We hebben net als in het model met twee toestanden aangenomen dat we geen gecensoreerde data hebben. We gaan ervan uit dat bij alle patiënten binnen de observatietijd overgangen hebben plaatsgevonden en dat we de precieze overgangstijd weten. Ook nemen we aan dat er geen transmissie kan plaatsvinden en dat de patiënten dus onafhankelijk zijn van elkaar. Beiden komen in de realiteit wel bijna altijd voor en zouden voor een realistischer model dus wel moeten worden meegenomen.
Ten tweede hebben we voor het model willekeurige waarden voor de parameters gekozen. In werkelijkheid zouden deze geheel anders kunnen zijn waardoor de overgangsfuncties andere vormen hebben. De keuze van de waarden voor de parameters heeft ook effect op de correlatie tussen de verschillende parameters. In werkelijkheid zullen er altijd parameters zijn die met elkaar gecorreleerd zijn. Er is immers altijd een verband tussen het risico om te overlijden als je nog maar kort besmet bent of als je al langer besmet bent. Dit verband hebben wij nu niet kunnen vinden maar zal in een realistischer model waarschijnlijk wel aanwezig zijn.
Het is dus ook nog interessant om te kijken naar de situatie waarin de priors voor de parameters afhankelijk zijn van elkaar. Dit hebben we in deze scriptie niet behandeld. In dit geval is het bijvoorbeeld zo dat de prior voor de eerste parameter c_1 exponentieel verdeeld is en de tweede parameter c_2 normaal verdeeld met verwachte waarde de prior voor c_1.

4 Conclusie

In deze scriptie hebben we gekeken naar de invloed van verschillende prior verdelingen op de schattingen voor parameters in het survival model. In de resultaten voor het model met twee toestanden hebben we gezien dat een prior met een normaalverdeling gecentreerd rondom de werkelijke waarde over het algemeen de beste resultaten oplevert. Als we kijken naar niet-informatieve priors dan levert de prior van de vorm $1/x$ de beste resultaten. We hebben in het multistate model vervolgens gebruik gemaakt van normaal verdeelde priors. Hier hebben we gezien dat de parameters van de overgangsintensiteiten goed geschat kunnen worden met behulp van Markov Chain Monte Carlo methoden.

De modellen die we hebben gebruikt zijn zeer vereenvoudigde modellen van de werkelijkheid. Dit hebben we in de discussies van hoofdstuk 2 en hoofdstuk 3 al besproken. We hebben laten zien hoe complexere modellen geschat zouden kunnen worden met behulp van de Bayesiaanse statistiek. Dit maakt het mogelijk om op deze manier schattingen te doen voor meer realistische modellen waarbij censoring en transmissie wel worden meegenomen.

Referenties

[1] Jackson C. (2010) *Multi-State Modelling with R: The msm Package.* `https://cran.r-project.org/web/packages/msm/vignettes/msm-manual.pdf`

[2] Meira-Machado L, Una-Ávarez J, Cadarso-Suárez C, Andersen PK (2009). *Multi-state models for the analysis of time to event data.* Stat Methods Med Res 18.

[3] Rodriguez G. (2005) *Lecture Notes on Survival Analysis.* `http://data.princeton.edu/wws509/notes/c7.pdf`

[4] Van Ravenzwaaij D., Cassey P., Brown S. (2016). *A simple introduction to Markov Chain Monte-Carlo sampling.* Psychonomic Bulletin and Review.

[5] Zhang D. (2005) *Lecture Notes on Analysis of Survival Data.* `https://www4.stat.ncsu.edu/ dzhang2/st745/chap3.pdf`