

Gesprek van Piet en Robbert

Robbert— Zo Piet, je begon je loopbaan als columnist bij ons door veel over corona te schrijven. Hoe staat het daar nu mee?

Piet— Tja, ik ben de afgelopen jaren eigenlijk enigszins tegen mijn zin met het probleem van het schatten van de verdelingsfunctie van de incubatietijd bezig geweest.

R.— Tegen je zin?

P.— Er zijn verschillende redenen.

- (1) Je moet je altijd afvragen: “Voor wie schrijf ik eigenlijk?”. Als het voor de medische statistici is, heb ik eigenlijk het gevoel dat ik het net zo goed kan laten. Aan de andere kant zou ik erg graag willen dat ze nu eens op zouden houden met het domme fitten met parametrische verdelingen. En dat zal op een gegeven moment ook wel gebeuren, maar ik denk dat ik dat mogelijk niet meer zal meemaken.
- (2) De waarnemingen zelf zijn zodanig dat we niet al te veel kunnen verwachten. Meestal is de tijd van symptomatisch worden (dat is iets je tot op zekere hoogte kunt waarnemen, i.t.t. tot de infectietijd) gecensureerd: we weten alleen de dag waarop het gebeurde of erger nog, we hebben alleen een tijdsinterval van meer dagen voor dit tijdstip.

Als we nu ons target wat verleggen en bijvoorbeeld zeggen: “Laten we ons beperken tot het schatten van integralen van de verdelingsfunctie over een dag of verschillende dagen.”, zijn we in een iets betere positie. Over die kwestie heb ik het in [7].

- (3) In [6] heb ik de asymptotische verdeling afgeleid voor de situatie dat we de tijd van symptomatisch worden wel met grotere precisie kunnen waarnemen. Dat was bijzonder lastig (vond ik), maar had enige overeenkomst met iets dat ik al eerder in [3] had gedaan. Je kunt het afleiden uit een fixed point karakterisering van de oplossing, maar het

vervelende is dat die fixed point karakterisering zelf weer stochastische termen bevat die van de oplossing afhangen. Je moet dan laten zien dat die stochastische termen dichter bij deterministische termen zitten dan de oplossing zelf. Hier heb je iets voor nodig dat wel “smooth functional theory” wordt genoemd. Dat is een vrij subtiële kwestie, met Cauchy-Schwarz en zo kom je er niet.

R.— Maar ik neem aan dat dat toch allemaal wel in de kringen van de medische statistici bekend is?

P.— Ja, hier raak je een gevoelig punt. Ik durf tegenwoordig niet meer vrijuit hier over te spreken, uit angst dat er gedreigd wordt met gerechtelijke stappen. Eigenlijk ben ik niet eens zo bevreesd voor die gerechtelijke stappen, als wel voor de tijd en energie die mij dat allemaal gaat kosten. Met die gerechtelijke stappen ben ik tijdens het schrijven van columns voor jullie tijdschrift toch al bedreigd (indirect, via jullie).

Als ik mijn grote voorbeeld Youp van 't Hek zie schrijven over “het domme Dilan-gansje” (de politieke leider van de VVD dus), denk ik bij mezelf: “Ohhh, stel je voor dat ik op soortgelijke wijze over mijn tegenstanders onder de medische statistici zou schrijven.”.

R.— Kun je niet iets concreter worden?

P.— Ja, nu probeer je me dus toch weer te verleiden om de gevarenzone te betreden. Maar vooruit dan maar, misschien dient het toch een goed doel.

In deel D van het supplement van [2] staat één bladzijde over de in hun artikel zo belangrijke niet-parametrische maximum likelihood schatter van de verdelingsfunctie van de incubatietijd. Midden op de pagina zeggen ze over de betrouwbaarheidsintervallen: “We used the confidence intervals as provided by the survfit function (dat komt uit een R package), which are known to be inconsistent.”. Dat vond ik al heel moeilijk te begrijpen, hoor, dat ze betrouwbaarheidsintervallen gebruiken waarvan ze zeggen dat het bekend is dat ze inconsistent zijn.

Ze gaan dan verder met te zeggen: “The NPMLE does not have $\sqrt{n}$ convergence rate and the central limit theorem does not hold.” De NPMLE is de “nonparametric maximum likelihood estimator”. En ze verwijzen hierbij in een noot naar naar [1]. Dat ze hier in een noot naar verwijzen en niet via de gewone bibliografie die ze ook hebben, is op zichzelf al eigenaardig. “Waarom?”, vraag je je af.

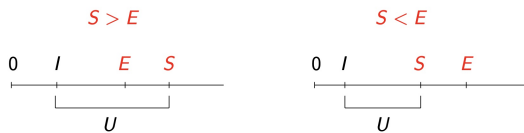
Maar die verwijzing is op zichzelf ook al interessant.

R.— Waarom vind je dat?

P.— Omdat het artikel [1] over iets totaal anders gaat. Dat artikel gaat over een schatter die in de econometrie gebruikt wordt: “Manski’s maximum score estimator”, waarvoor al in 1990 door Kim en Pollard theorie werd ontwikkeld ([8]). Ik zie met geen mogelijkheid hoe hier de verdelingstheorie en in het bijzonder de bewering die ze doen over de NPMLE uit kan worden afgeleid.

R.— Nou ja, misschien begrijp je de reuzensprong die de auteurs hier maken van [1] naar het gedrag van de NPMLE niet?

P.— OK, om het mogelijk voor jou wat interessanter te maken, zal ik nu even ingaan op de problemen die hier spelen. Zowel bij het uitrekenen van het zg. *effectieve reproductiegetal* R_e (“is het groter of kleiner dan 1?”, herinner je je nog Jaap van Dissel op de televisie?) als bij het schatten van de verdeling van de incubatietijd hebben we te maken met een infectietijd die we niet kunnen waarnemen. En bij het schatten van de incubatietijdverdeling hebben we ook nog te maken met een deconvolutie probleem, omdat we de incubatietijd moeten losweken uit de tijd van symptomatisch worden (som van de infectietijd en incubatietijd), die we tot op zekere hoogte wel kunnen waarnemen. De situatie wordt schematisch weergegeven in Figuur 1, die ik even haal uit [7].



FIGUUR 1. E is het einde van de “exposure time”, S het tijdstip van symptomatisch worden, I de infectietijd en U de (lengte van de) incubatie tijd. We kunnen alleen E en S waarnemen. E kan zowel links als rechts van S liggen

Als we aannemen dat, gegeven de lengte van de blootstellingstijd (“exposure time”) E , de dichtheid van de infectietijd I uniform is op $[0, E]$, en als F de verdelingsfunctie van de incubatietijd is, wordt de dichtheid van (E, S) in het model gegeven door

$$\begin{aligned} \phi_F(e, s) &= e^{-1} \int_{u \in [(s-e)_+, s]} dF(u) \\ &= e^{-1} \{F(s) - F(s-e)\}, \end{aligned} \quad (1)$$

t.o.v. het product van de maat dF_E van de (lengte van de) blootstellingstijd E en Lebesgue maat op $\mathbb{R}$.

In dit zeer eenvoudige model is de maximum likelihood schatter van de verdelingsfunctie F van de incubatietijd, *die* verdelingsfunctie $\hat{F}_n$ die de uitdrukking (“log likelihood”)

$$(2) \quad \ell(F) = \sum_{i=1}^n \log \phi_F(E_i, S_i)$$

maximaliseert als functie van F over waarnemingen (E_i, S_i) , $i = 1, \dots, n$ van de blootstellingstijden (ik zeg eigenlijk liever “exposure times”) E_i en tijdstip van symptomatisch worden S_i , waarbij ϕ_F is gedefinieerd als in (1). Ik heb het hier trouwens al eerder in jullie tijdschrift over gehad.

R.— Het lijkt me niet zo heel eenvoudig, dat maximaliseren als functie van de verdelingsfuncties F .

P.— Inderdaad Robbert! Nee, dat is helemaal niet zo triviaal. In de medische statistiek is het gebruikelijk om een of andere familie van verdelingen voor F te kiezen: Weibull, gamma, log-normaal, Erlang, ... en dan te maximaliseren over de parameters die zo’n verdeling karakteriseren. Het nadeel van deze werkwijze is dat als je denkt dat Weibull redelijk is, je er met de log-normale behoorlijk naast zit (en vice versa) en dat er geen dwingende reden is om voor één van deze families te kiezen. Daarom zie je dus in de medisch-statistische artikelen die verdelingen allemaal of bijna allemaal optreden, wat het aantal tabellen ook weer behoorlijk opdrijft.

Maar in [2] bekijken ze ook de volledig niet-parametrische schatter, waar deze parametrische veronderstellingen niet gemaakt worden. Tenminste, dat zeggen ze. Helaas zeggen ze niet hoe ze die uitrekenen.

Ik vrees dat hier mogelijk het ongelooflijk langzame EM (“Expectation Maximization”) algoritme is gebruikt (ik had dat algoritme in 2021 ook al besproken in [5]), wat tot gevolg heeft dat er in het algemeen te vroeg gestopt wordt met de iteraties. Daar is een uitdrukking voor: “EM with early stopping” (dan raak je dus volledig afhankelijk van de startwaarden van je iteratieve algoritme). Maar ik weet het niet, want ze zeggen het niet. Ik zou natuurlijk in de toepassing van de door hen gebruikte R packages kunnen gaan zitten kijken, maar daar heb ik echt geen zin in; het hoort toch eigenlijk in hun artikel te staan wat ze doen.

Maar belangrijker is dat we niets wisten over de verdelingseigenschappen van de schatter $\hat{F}_n$ die tenslotte aan het eind van (niet vroeg gestopte) iteraties tevoorschijn komt, waar ik me dus de afgelopen jaren, enigszins tegen mijn zin mee heb bezig gehouden.

R.— OK. Het is wel enigszins duidelijk waarom je aarzelingen had je hierin te storten. Maar misschien moet je in het algemeen toch niet te veel nadenken over het publiek waarvoor je schrijft.

Referenties

- [1] Jason Abrevaya and Jian Huang. On the bootstrap of the maximum score estimator. *Econometrica*, 73(4): 1175–1204, 2005.
- [2] Vera H. Arntzen, Marta Fiocco, Nils Leitzinger, and Ronald B. Geskus. Towards robust and accurate estimates of the incubation time distribution, with focus on upper tail probabilities and sars-cov-2 infection. *Statistics in Medicine*, 42:2341–2360, 2023. URL <https://onlinelibrary.wiley.com/doi/full/10.1002/sim.9726>.
- [3] P. Groeneboom. Lectures on inverse problems. In *Lectures on probability theory and statistics (Saint-Flour, 1994)*, volume 1648 of *Lecture Notes in Math.*, pages 67–164. Springer, Berlin, 1996. doi: 10.1007/BFb0095675. URL <http://dx.doi.org/10.1007/BFb0095675>.
- [4] Piet Groeneboom. En flagrant délit. <https://www.nieuwarchief.nl/serie5/pdf/naw5-2021-22-3-163.pdf>, 2021.
- [5] Piet Groeneboom. Estimation of the incubation time distribution for COVID-19. *Stat. Neerl.*, 75(2):161–179, 2021. URL <https://onlinelibrary.wiley.com/doi/epdf/10.1111/stan.12231>.
- [6] Piet Groeneboom. Nonparametric estimation of the incubation time distribution. <https://arxiv.org/abs/2205.04399>, 2022.
- [7] Piet Groeneboom. Estimation of the incubation time distribution in the singly and doubly interval censored model. <https://arxiv.org/abs/2310.04225>, 2023.
- [8] J.K. Kim and D. Pollard. Cube root asymptotics. *Ann. Statist.*, 18:191–219, 1990. ISSN 0090-5364. doi: 10.1214/aos/1176347498. URL <http://dx.doi.org/10.1214/aos/1176347498>.