1.12 Hoogdimensionale statistiek

Prof. dr. A.W. (Aad) van der Vaart
Mathematical Institute
Universiteit Leiden

www.math.leidenuniv.nl/~avdvaart

- Lezing gehouden voor de Koninklijke Maatschappij voor Natuurkunde 'Diligentia' te 's-Gravenhage op 12 maart 2018.
- Van deze lezing is helaas geen video opname beschikbaar.

Samenvatting van de lezing:
In 2013 vierden we de 300ste verjaardag van 'Ars Conjectandi', een boek van Jacob Bernoulli, dat wordt gezien als het begin van de kansrekening en statistiek als wetenschappelijke discipline. Ongeveer vijftien jaar geleden zouden we volgens de statisticus Brad Efron een nieuw tijdperk van hoog-dimensionale statistiek zijn ingegaan, het derde tijdperk van de statistiek in die 300 jaar. Sinds enkele jaren horen we veel over 'big data' en sommigen voorspellen dat 'slimme algoritmen' binnenkort alle problemen zullen oplossen, niet alleen die van ons dagelijks leven, maar ook die bij het beantwoorden van vraagstellingen waar vroeger statistische analyse onontbeerlijk leek.
Tijdens de lezing werden enkele draden van deze ontwikkelingen opgepakt, aan de hand van voorbeelden zoals: de rechtvaardigheid van data-algoritmen in de rechtspraak, het causaliteitsprobleem in epidemiologisch onderzoek, en de soms gebrekkige repliceerbaarheid van wetenschappelijk onderzoek.

In de ochtend van 19 januari 2018 berichtten de Nederlandse media over een medische doorbraak: met behulp van een eenvoudige bloedtest is het mogelijk kanker vast te stellen. Het zou een ontdekking zijn met verreikende consequenties, maar helaas werd het bericht in de middag al weer afgezwakt tot "opwindend, maar geen doorbraak". De laatste jaren lezen we dit soort berichten met regelmaat. Het gaat in veel gevallen om het analyseren van data aangaande de expressie van veel (of alle) genen, of daarvan afgeleide producten als proteïnen of metabolen. De doorbraak zou hier zijn dat het analyseren van meerdere typen data tegelijk de diagnose enorm kan verbeteren. Helaas was de nieuwe techniek niet onderscheidend genoeg om echt bruikbaar te zijn.

Deze data zijn hoog-dimensionaal in de zin dat bij één patiënt tot wel een miljoen getallen worden verzameld. De experimentele technieken die dit mogelijk maken hebben de laatste 20 jaar een enorme ontwikkeling doorgemaakt. Waar eerst een meting van één gen of marker een aanzienlijke inspanning vereiste, is het inmiddels mogelijk grote aantallen in één keer te meten. In principe geeft dit een schat aan informatie. De belofte is dat met de huidige krachtige computers die informatie ook bruikbaar wordt. De massieve data over één persoon impliceert dan een medische behandeling die is afgestemd op die persoon. Dat dit echter nog niet zo makkelijk is, blijkt wel uit het feit dat rond 2000 werd gedacht dat deze *personalised medicine* in 15 jaar wel zou arriveren. Er is inderdaad veel bereikt, maar gepersonaliseerde behandelingen zijn in 2018 nog steeds toekomstmuziek. Biologische processen zijn buitengewoon ingewikkeld, en metingen over die processen leveren zowel signaal als ruis. Om dat signaal te extraheren zijn nieuwe statistische methoden en algoritmen noodzakelijk, en meer data en kennis van biologische processen.

Moderne medische data zijn voorbeelden van *big data*. De laatste term heeft de laatste jaren een

hype-status gekregen. 'Big data' zijn overal en hebben talloze toepassingen. Niet allemaal positief. Men denke bij 'big data' vooral ook aan bedrijven als Google en Facebook, die data verzamelen om geld te verdienen met advertenties, en/of mensen te beïnvloeden.

Algoritmen, leren en regressie

Massieve data is op zich niet heel zinvol, zoals iedereen weet die wel eens een spreadsheet met vele duizenden regels heeft opengemaakt. *Algoritmen* zijn noodzakelijk om zinvolle informatie aan die data te onttrekken. Het eenvoudigste voorbeeld van een data-algoritme is een input-output functie. Gegeven input data x wil men een *output* y voorspellen. Denk aan x gelijk aan alle metingen aan het bloed van een patiënt en y een codering voor het hebben van kanker of niet. Wiskundig gezien is een algoritme niets anders dan een *functie*:

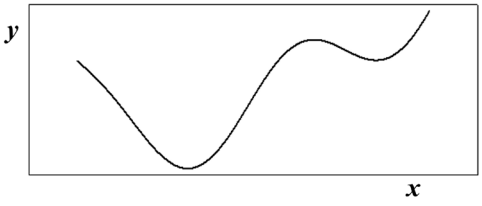
$$x \to y = f(x).$$

Voor een 1-dimensionale variabele x zou deze functie kunnen worden voorgesteld door een grafiek, zoals weergegeven in figuur 1, maar we denken hier natuurlijk aan een samengestelde input $x = (x_1, x_2, \ldots, x_p)$ bestaande uit de duizenden of miljoenen metingen aan één persoon of eenheid.

De functie f is niet op voorhand bekend; bij big data toepassingen is er meestal weinig theorie over haar vorm. De functie wordt daarom bepaald met behulp van voorbeelden, waarvoor zowel de input als de output bekend zijn. Men denke aan patiënten waarbij een bloedbepaling is verricht en tegelijkertijd kanker is vastgesteld of niet, bijvoorbeeld met behulp van een PET-scan. Gegeven een flink aantal van zulke voorbeelden $(x_1, y_1), \ldots, (x_n, y_n)$, bepaalt men een functie f die klopt voor deze voorbeelden: $y_i = f(x_i)$, voor iedere $i = 1, \ldots, n$. In de praktijk moet men zich hoeden voor het overfitten op de gegeven voorbeelden, en een functie zoeken die de output ongeveer, of meestal goed, voorspelt. Men zoekt bijvoorbeeld een functie die een kleinste-kwadraten criterium van de vorm

$$f \to \sum_{i=1}^{n} \left(y_i - f(x_i) \right)^2 + J(f)$$

minimaliseert door een gegeven aantal mogelijke functies f uit te proberen. Hierin is $J(f)$ een *penalty* die moet voorkomen dat f de gegeven voorbeelden te zeer interpoleert. In de statistiek wordt een dergelijke procedure sinds het einde van de 19de eeuw aangeduid met *regressie-analyse* (de output gaat terug op de input). Binnen de context van de big data wordt gezegd dat de functie f *geleerd* wordt, en men spreekt ook al wel van *zelf-lerende* systemen.

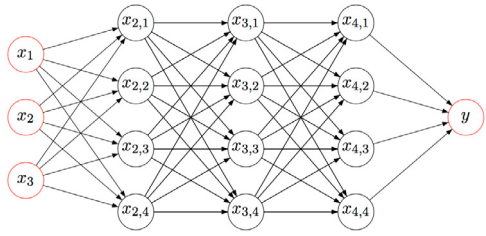


Figuur 1: voorbeeld van een grafiek van een 1-dimensionale functie $y = f(x)$.

Men beperkt zich altijd tot functies f van een bepaald type. *Fourier reeksen, splines, wavelets, support vector machines, Gaussian processes, random forests* wisselden in de afgelopen decennia in populariteit. In de laatste vijf jaar zijn *deep learning neural networks* de grote trend. Dit betreft functies die men kan voorstellen door een diagram zoals in figuur 2 is weergegeven.
De input vector $x = (x_1, \ldots, x_3)$, hier voor het gemak slechts 3-dimensionaal, staat links; de output y rechts. De tussenwaarden $x_{i,j}$ worden berekend, per laag van links naar rechts als

$$x_{i,j} = \psi \left(\sum_j \vartheta_{i,j} x_{i-1,j} \right),$$

waarbij ψ een vaste functie is, de zogenaamde



Figuur 2: een neuraal netwerk met 3-dimensionale input en 3 tussenlagen met elk 4 tussenwaarden.

activatiefunctie van het netwerk (bijvoorbeeld de *RELU functie* die een negatieve waarde op 0 zet en een positieve waarde onveranderd laat). De $\vartheta_{i,j}$ zijn onbekende gewichten, voor iedere pijl in het diagram één, die geleerd moeten worden uit de voorbeelden, bijvoorbeeld met behulp van de kleinste-kwadraten methode als eerder uiteengezet. In moderne toepassingen kunnen er honderdduizenden inputs zijn en moeten er miljoenen onbekenden $\vartheta_{i,j}$ worden bepaald. Het resulterende algoritme kan opmerkelijk goed werken om de output van nieuwe inputs x te bepalen. Vooral bij het analyseren van foto's is in de laatste 10 jaar veel vooruitgang geboekt. (Een foto kan worden omgezet in een numerieke input vector, door de pixels te representeren door hun kleur en intensiteit.) Waarom dit zo goed werkt, wordt op dit moment niet goed begrepen. De klassieke regel dat een overdaad aan vrije parameters tot problemen leidt bij het aanpassen aan de data, blijkt niet op te gaan. Eén mogelijke verklaring is dat het kleinste-kwadraten criterium niet werkelijk wordt geminimaliseerd, maar de gebruikte iteratieve methoden op een goed moment worden gestopt.

Leermethoden behoren tot de gereedschapskist van zowel de statistiek als de kunstmatige intelligentie. Toepassingen vinden we op vele terreinen: medische diagnostiek, het voorspellen van verzekeringsclaims (met het doel klanten niet te willen verzekeren of een hogere premie te berekenen), het voorspellen van misdaad door de politie, belastingfraude, robotica. Algoritmes worden thans zelfs gebruikt door rechters voor het voorspellen van recidive. Sommigen voorspellen dat computers door al dat 'leren' straks slimmer zullen zijn dan mensen. De mogelijkheden zijn inderdaad groot en de successen indrukwekkend. Zoals eerder uitgelegd blijft leren in de basis echter kleinste kwadraten. Neurale netwerken met één laag werden in de jaren 1950 al toegepast onder de titel *logistische regressie*, op soortgelijke problemen. Nieuw is de beschikbaarheid van zowel veel meer data als veel krachtiger computers. De toekomst zal leren of dit inderdaad leidt tot een geheel nieuwe situatie (de *singularity*).

Duidelijk is wel dat lang niet alle mogelijke toepassingen gewenst zijn. Niet alleen zou men sommige outputs niet moeten willen voorspellen, ook zijn de algoritmes blind. Stel bijvoorbeeld dat je in een wijk woont, met veel mensen die de verzekering oplichten. Je persoonlijke input vector x zal dan de postcode gemeen hebben met veel oplichters, en het algoritme van een verzekeringsmaatschappij zal je als een risico aanmerken. Uiteraard ben je goudeerlijk, maar je lijdt onder de buren. Dit gebeurt omdat het voor het kleinste-kwadraten criterium niet uitmaakt, welke input variabelen worden gebruikt. Als de voorspelling maar goed is. Een neuraal netwerk is daarbij ook nog zo complex, dat niet goed is te achterhalen welke aspecten van de input x, en hoe, worden gebruikt. Dergelijke problemen zijn nog serieuzer als algoritmen worden gebruikt in de rechtspraak. In de Verenigde Staten heeft dit al geleid tot een verhitte discussie over het gebruik van ras. Men zou kunnen denken dat racistische algoritmen kunnen worden voorkomen door huidskleur niet als input in het algoritme op te nemen. Echter hangt huidskleur met tal van andere variabelen samen, en die zouden in de black box die het geleerde algoritme is, toch tot racisme kunnen leiden.

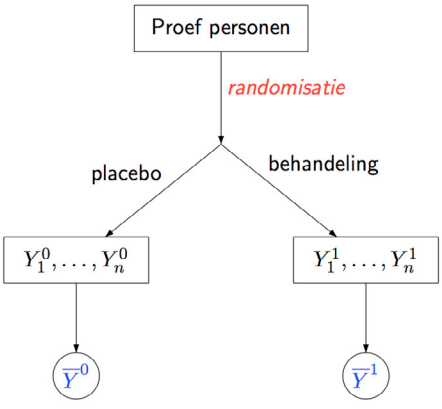
Correlatie is geen causatie

Deze bedenkingen hangen nauw samen met *causaliteit*. Als het zo was dat een bepaalde postcode *oorzaak* was van verzekeringsfraude, dan was het wellicht te billijken dat de premie in dit postcodegebied hoger zou zijn. Dit lijkt in dit geval echter uitgesloten: er is geen causale relatie tussen postcode en fraude.

Causale relaties zijn de basis van het wetenschappelijk kennen. Voorspellingsalgoritmen kunnen gebruikmaken van causale verklaringen, maar in veel gevallen zijn ze gebaseerd op *correlatie* en niet op causatie. En correlatie is geen causatie, leert iedere wetenschapper: twee variabelen kunnen samenhangen (gelijktijdig groot of klein zijn, of andersom, in het geval van negatieve samenhang), zonder dat de ene variabele oorzaak is van de andere. Filosofen van alle tijden, waaronder

Aristoteles, Kant en Hume, hebben over de diepe betekenis van causaliteit geschreven. Statistici hanteren een eenvoudige operationele definitie van causaliteit, gebaseerd op *randomisatie*, zie figuur 3. In het eenvoudigste geval is men geïnteresseerd in het effect y van een behandeling die men kan geven of niet. Ieder persoon in een gegeven steekproef van proefpersonen uit een populatie onder studie wordt ofwel behandeld ofwel niet, waarbij de keuze wordt bepaald door een munt op te werpen, onafhankelijk voor de verschillende personen. Na de behandeling verzamelt men de effecten, zeg $y_1^0,\ldots,y_n^0$ voor de onbehandelde groep en $y_1^1,\ldots,y_n^1$ voor de behandelde groep. Wordt een verschil tussen de twee groepen waargenomen, dan zijn er twee mogelijkheden: het verschil is per toeval ontstaan, door de toevallige indeling van de personen in de twee groepen, of de behandeling is de *oorzaak* van het verschil. Nu is de kern van de statistische wetenschap om toeval te kwantificeren. Als het verschil zogenaamd *statistisch significant* is, dan kan het toeval met grote waarschijnlijkheid worden uitgesloten, en is de causaliteit van de behandeling aangetoond.

Nu lenen veel vragen zich niet voor het opzetten van een gerandomiseerd experiment. Berucht zijn bijna alle onderzoeken over voeding, maar veel sociaal-economisch onderzoek heeft soortgelijke beperkingen. Men kan moeilijk proefpersonen verplichten voor lange termijn veel rode wijn te drinken, of niet, of veroordeelden randomiseren over taakstraf of gevangenisstraf. De data bestaat dan uit het passief waarnemen van de wereld zoals die is, en is *observationeel* in plaats van experimenteel. Veel big data onderzoek is observationeel, en het feit dat het veel data betreft, is niet zonder meer behulpzaam. De grootte van de data ondervangt immers niet dat in de echte wereld alles met alles samenhangt, en een waargenomen effect allerlei oorzaken kan hebben. Zo zullen ernstig zieke patiënten eerder een zwaardere behandeling krijgen, en er later slechter aan toe zijn, dan minder zieke patiënten. Men moet dan niet de fout maken te concluderen dat de zwaardere behandeling slecht was.



Figuur 3: Diagram van een veel gebruikte methode om causaliteit te testen.

Statistici proberen die andere mogelijke oorzaken (zogenaamde *confounding variables*) te controleren, door ze in een model op te nemen. Sinds de jaren 1980 is hiermee veel vooruitgang geboekt, en de grote datasets van vandaag geven nieuwe mogelijkheden om deze modellering realistischer te maken. We lichten een tipje van de sluier op aan de hand van het voorgaande voorbeeld.

Een gegeven persoon zouden we in principe kunnen behandelen of niet, en we zouden dan een uitkomst y^1 (bij behandeling) of y^0 (bij geen behandeling) kunnen vaststellen. Het causaal effect voor die persoon is dan het verschil y^1-y^0 en het effect voor een gegeven populatie van individuen is het gemiddelde E(y^1-y^0) van de causale effecten (de E is het symbool voor 'expectation', E duidt een populatiegemiddelde aan). Helaas nemen we voor iedere persoon maar één effect y waar: $y=y^1$ indien behandeld en $y=y^0$ indien niet behandeld. Veronderstel nu dat we voor een grote verzameling personen paren (y,z) waarnemen, waarbij z de behandeling codeert: $z=1$ als de persoon behandeld is en $z=0$ als niet. Met deze notatie kunnen we de volgende wiskundige stelling formuleren:

STELLING Als de uitkomst y en de behandelingsindicator z stochastisch onafhankelijk zijn, dan geldt:

$$E\left(y^1-y^0\right)=E(y\mid z=1)-E(y\mid z=0).$$

Het begrip 'stochastisch onafhankelijk' heeft een precieze wiskundige betekenis in de kansrekening, maar ook een duidelijk intuïtieve interpretatie. Het betekent dat kennis van z, de wetenschap of de persoon behandeld is, geen informatie geeft over de uitkomst y voor die persoon. Dat is bijvoorbeeld waar als tot al dan niet behandeling is besloten op grond van het gooien van een munt, zoals bij het gerandomiseerd experiment dat we eerder bespraken. De stelling is derhalve toepasbaar op een gerandomiseerd experiment. Zij zegt dat het gemiddeld causaal effect (het linkerlid van de vergelijking) gelijk is aan het gemiddelde van de effecten van de behandelde mensen, namelijk $E(y \mid z=1)$, min het gemiddelde effect van de niet-behandelde mensen, namelijk $E(y \mid z=0)$. De verticale streep in deze notaties geeft aan dat de expectation/het gemiddelde moet worden beperkt tot de mensen met $z = 1$ of $z = 0$.

De stelling zegt derhalve dat in een gerandomiseerd experiment het causaal effect kan worden berekend door de gemiddelde effecten van de behandelde en onbehandelde personen te vergelijken, precies zoals we al betoogden. De stelling is niet meer dan een rechtvaardiging van die eerdere redenatie. Het mooie is dat de stelling een denkkader geeft dat geschikt is voor uitbreiding naar observationele data.

Op observationele data is de stelling niet toepasbaar, omdat de voorwaarde niet is vervuld: het effect y hangt samen met de behandelingsindicator z. Een dokter besloot bijvoorbeeld de ziekste mensen juist wel te behandelen, of de rechter gaf taakstraf aan de veroordeelden van de lichtste vergrijpen of die in zijn visie de minste kans op recidive hadden. We veronderstellen nu dat we extra data x tot onze beschikking hebben om voor deze mogelijkheden te corrigeren. Preciezer, we veronderstellen dat x alle informatie bevat die de dokter of rechter gebruikt heeft om tot behandeling te besluiten, of niet, en die samenhangt met het effect van de behandeling. In dat geval is aan de voorwaarde van de volgende uitbreiding van de voorgaande stelling voldaan.

STELLING Als de uitkomst y en de behandelingsindicator z stochastisch onafhankelijk zijn, gegeven x, dan geldt:

$$E\left(y^1 - y^0\right) = E_x\left(E(y \mid x, z=1) - E(y \mid x, z=0)\right).$$

Wederom staat links het causaal effect, en rechts een manier om dit uit te rekenen, op basis van de verzamelde data (y, z, x). De uitdrukkingen $E(y \mid x, z = 1)$ en $E(y \mid x, z = 0)$ zijn gemiddelde effecten in de populaties van personen met een gegeven waarde x die behandeld ($z = 1$) zijn of niet ($z = 0$). Het verschil van deze gemiddelden wordt tenslotte gemiddeld over de waarden van x (aangegeven door het symbool E_x).

Dit werkt alleen als de variabele x goed is gekozen, wat alleen mogelijk is op basis van inzicht in de behandeling. Zo'n x is vaak een hoog-dimensionale vector met veel verschillende waarden, hetgeen tot het probleem leidt dat de voorwaardelijke verwachtingen $E(y \mid x, z = 1)$ en $E(y \mid x, z = 0)$ moeten worden bepaald op grond van relatief kleine subpopulaties. Een eerste oplossing van dit probleem is dat de stelling waar blijft als x aan de rechterkant wordt vervangen door de zogenaamde *propensity score*

$$\pi\left(x\right) = P(z = 1 \mid x).$$

Dit is de kans dat een persoon met waarde x wordt behandeld ($z = 1$). Substitutie van $\pi(x)$ voor x maakt dat het in principe voldoende is de populatie te stratificeren naar de waarden van $\pi(x)$. In de praktijk is deze kans op behandeling niet bekend, maar kan wel uit de beschikbare data worden bepaald. De functie $x \longrightarrow \pi(x)$ is immers niets anders dan een algoritme dat uit de beschikbare data $(x^1, z^1), \ldots, (x^n, z^n)$ kan worden geleerd. Een deep-learning netwerk is één van de mogelijkheden.

Een efficiënte implementatie van deze correctiemethoden grijpt terug op semiparametrische statistische theorie ontwikkeld in de jaren 1990. Praktische implementatie vereist verder onderzoek; wellicht is de aanpak over 15 jaar standaard? Naast de propensity score worden ook andere ideeën

benut, zoals bijvoorbeeld instrumentele variabelen, die vooral in de econometrie populair zijn.

Bayesiaanse statistiek

In de statistiek zet men vaak twee paradigma's tegenover elkaar. De oudste is de Bayesiaanse statistiek, die teruggaat op Thomas Bayes (1702-1761) (figuur 4), een Engelse dominee en amateur wiskundige. Bayes' ideeën werden na zijn dood herontdekt en gegeneraliseerd door grote wiskundigen als Laplace en Gauss, en waren algemeen geaccepteerd als de theoretische basis voor statistisch redeneren tot het begin van de 20ste eeuw, toen met name Ronald Fisher de methode bekritiseerde en voor het gebruik van "maximum likelihood" pleitte. In de laatste decennia zien we een revival van de Bayesiaanse methode, deels omdat de praktische implementatie gemakkelijker is geworden door de rekenkracht van computers, deels omdat de methode voordelen lijkt te bieden, wellicht in het bijzonder voor hoog-dimensionale data.

Statistiek betreft het omgaan en kwantificeren van onzekerheid, en de methode van Bayes is even eenvoudig als elegant. Voordat data verzameld is, wordt de kennis over een fenomeen samengevat door middel van een kansverdeling over de mogelijke toestanden, de zogenaamde *a-priori verdeling*. Nadat de data is verkregen, wordt deze kansverdeling aangepast tot een nieuwe kansverdeling, de *a-posteriori verdeling*. Bayes ontwikkelde voor het eerste de regel die deze update beschrijft.

Figuur 4: Thomas Bayes (1702-1761).

Het voorbeeld dat Bayes zelf behandelde kan dit verduidelijken. Veronderstel dat een onbekende fractie ϑ van een populatie een bepaalde eigenschap bezit. Voordat we de populatie waarnemen zouden we kunnen redeneren dat ϑ volledig onbekend is, en deze parameter daarom kunnen beschouwen als een willekeurig getal tussen 0 en 1. In termen van kansen volgt ϑ de homogene verdeling: de kans dat ϑ in een bepaald interval ligt is gelijk aan de lengte van dit interval.

$$\prod(d\vartheta) = d\vartheta.$$

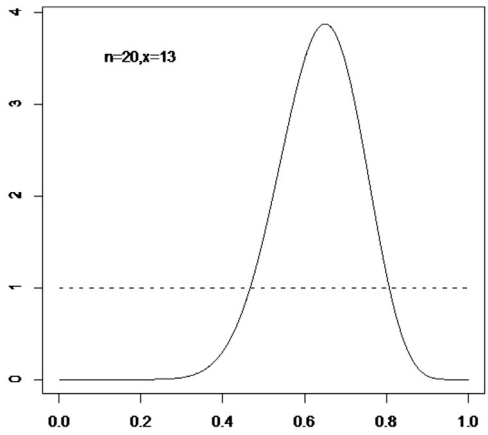
Stel dat we x mensen met de eigenschap vinden in een steekproef ter grootte n uit de populatie. Als x groot is (ten opzichte van n) zullen we onze a-priori verdeling willen aanpassen tot een verdeling die meer kans toekent aan grotere waarden van ϑ. De kans dat bij gegeven ϑ de data x wordt verkregen, volgt de binomiale kansverdeling:

$$p_\vartheta(x) = \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x}.$$

De regel van Bayes zegt nu dat de kans dat ϑ behoort tot het interval $(\vartheta,\vartheta+d\vartheta)$ gegeven de data x, gelijk is aan

$$\prod(d\vartheta \mid x) = p_\vartheta(x) \prod(d\vartheta).$$

De oorspronkelijke gewichten $\Pi(d\vartheta)$ worden aangepast door ze te herwegen met de kans $p_\vartheta(x)$ op de waargenomen waarde. De methode wordt geïllustreerd in figuur 5 waarin de horizontale lijn de



Figuur 5: homogene en a-posteriori kansverdeling.

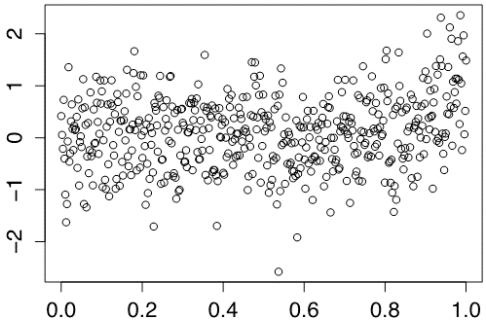
homogene kansdichtheid weergeeft, en de curve de dichtheid van de a-posteriori in het speciale geval dat $n = 20$ en $x = 13$. De a-posteriori heeft duidelijk meer massa in de buurt van de fractie x/n. De a-posteriori curve wordt op twee manieren gebruikt: ten eerste geeft de top van de curve de meest waarschijnlijke waarde ϑ; ten tweede geeft de spreiding van de massa een kwantificering van onze onzekerheid: waarden tussen ongeveer 0,5 en 0,8 zijn op grond van deze data niet uit te sluiten; waarden kleiner dan 0,2 wel.

De bezwaren van Fisher betroffen zowel de keuze van de a-priori verdeling (waarom homogeen?) als het voorgaande woordgebruik. In werkelijkheid is de parameter ϑ immers een vast getal. Alleen de data x zijn het resultaat van een kansexperiment, namelijk het nemen van een steekproef, maar welk kansexperiment rechtvaardigt, dat we in kansen spreken over ϑ?

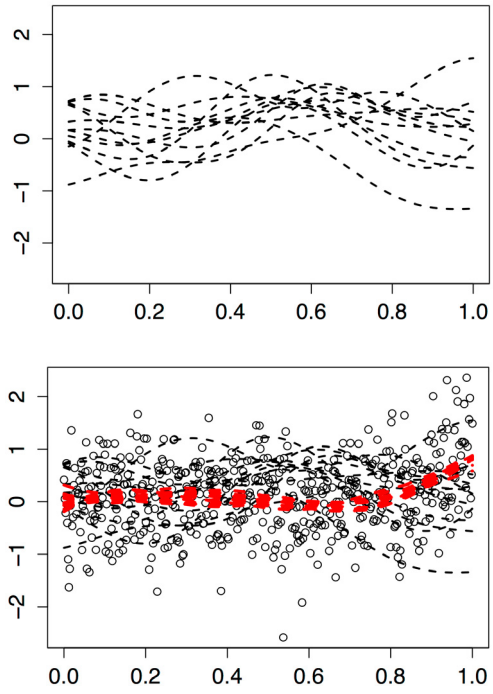
Het debat over de grondslagen van correct statistisch redeneren is 250 jaar na Bayes' overlijden nog steeds gaande, maar de elegantie van de Bayesiaanse methode lijkt in de 21ste eeuw nieuwe aanhangers te trekken. Daarbij komt dat het gebruik van a-priori kansen heel zinvol kan zijn voor complexe data en problemen, omdat zij toelaat bestaande informatie op een zachte manier aan de data toe te voegen. Zo kan men in een studie over de werking van genen inbouwen dat slechts een beperkt aantal van de 30.000 genen een bepaald genotype mede bepalen, of de interacties in het netwerk van genen relatief klein in aantal zijn, en gedeeltelijk overeenkomen met gegevens uit een database van eerder onderzoek.

Waar in het voorbeeld van Bayes de parameter een getal was, biedt de moderne kansrekening (die dateert van de 20ste eeuw) een formalisme dat ook toepasbaar is op complexe objecten, zoals functies. We geven een eenvoudig voorbeeld. Veronderstel dat de data bestaat uit paren $(x_1, y_1), \ldots, (x_n, y_n)$, waarvan we weten dat de y-waarde een functie $y = f(x)$ van de bijbehorende x-waarde is. We willen de functie afleiden uit de data, maar zijn ons bewust dat de waargenomen punten niet precies op de curve f liggen, maar onderhevig zijn aan ruis en toevallige fluctuaties. Figuur 6 geeft een extreem voorbeeld. Wat is de ware curve bij deze data?

De Bayesiaanse methode wordt geïllustreerd in figuur 7. Omdat de a-priori verdeling nu gewichten



Figuur 6: een waargenomen puntenreeks, onderhevig aan ruis en toevallige fluctuaties.



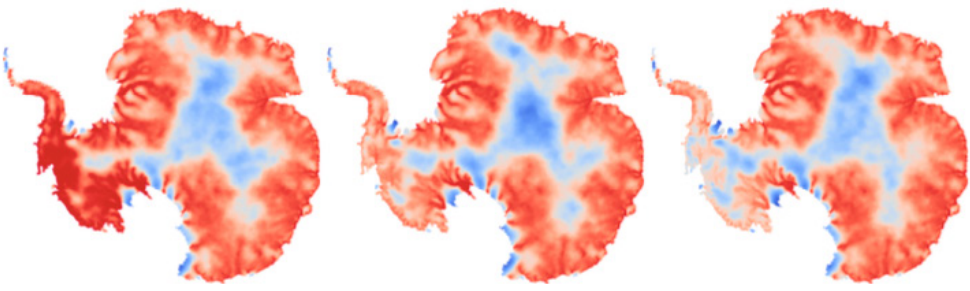
Figuur 7: (boven) trekkingen uit de a-priori verdeling en **(onder)** uit de a-posteriori verdeling (in rood).

geeft aan curves, is het niet mogelijk een dichtheidsfunctie voor de a-priori kansmassa te tonen, als in het voorbeeld van de binomiale verdeling. We kunnen echter wel een indruk van de a-priori verdeling geven door herhaald curves te trekken. De a-priori verdeling geeft iedere curve een gewicht, dat de kans bepaalt dat die curve wordt gekozen. Door trekkingen te herhalen zien we het soort curves dat veel kans krijgt onder de a-priori; zie het bovenste plaatje, waar de data niet getoond wordt (zoals dat hoort voor een echte a-priori). Nadat de data is verzameld, passen we de Regel van Bayes toe en verkrijgen de a-posteriori verdeling. Deze kunnen we op dezelfde manier visualiseren, door herhaald trekkingen te doen. Het idee is dat de verzameling mogelijke functies nu nieuwe gewichten heeft gekregen op grond van de waargenomen data, en daarom trekkingen oplevert die beter bij de data passen. Zie de rode curves in het onderste plaatje. Hun gemiddelde zal ook de ware curve ongeveer moeten geven, en de verschillen tussen de trekkingen een idee over hoe precies deze reconstructie is.

Wiskundig is een kansverdeling op een verzameling van curven een stochastisch proces. Het gegeven voorbeeld gebruikt een *Gaussisch proces*, waarvan veel voorbeelden bestaan. De vraag is: werkt de procedure, en welke invloed heeft de keuze van het specifieke stochastische proces? Deze vraag heeft ons in de afgelopen 15 jaar bezig gehouden. Zowel de vraag als het antwoord zijn te ingewikkeld om hier in detail te behandelen. Ruwweg komt het erop neer dat de keuze van de a-priori veel uitmaakt, maar dat verkeerde keuzen kunnen worden vermeden door een bandbreedte parameter in te bouwen, die dan uit de data moet worden geschat. Er zijn echter ook uitzonderlijke gevallen ('inconvenient truths'), waarin geen enkele methode voor curve schattingen goed werkt.

Curve schatting is maar één voorbeeld van een hoog-dimensionale statistische techniek. Het idee is eenvoudig uitbreidbaar tot het schatten van oppervlakken, zoals in figuur 8, die drie trekkingen uit een a-posteriori verdeling laat zien van een frictie-coëfficiënt van de bodem van Antarctica [1]. (blauw is laag, rood hoog). Een ander voorbeeld zijn netwerken, bijvoorbeeld van interacties tussen genen. Kennis over de werking van genen is in toenemende mate beschikbaar in databanken. Deze kennis kan worden verwerkt in een a-priori verdeling op de topologie van het netwerk: sommige verbindingen bestaan waarschijnlijk wel en andere niet. Met de Bayesiaanse methode kan deze kennis worden gecombineerd met nieuwe data, wellicht aangaande een andere groep patiënten of een andere configuratie.

Figuur 9 toont een schatting van het apoptosis netwerk. Ieder van de cirkels aan de omtrek representeert een gen, en een verbindingslijn een interactie. In dit geval gaat het om slechts 84 genen; goed zichtbaar is dat van de 3486 mogelijke interacties slechts weinig significant zijn. Gedacht wordt dat veel fenomenen *sparse* zijn in de zin dat



Figuur 8: drie realisaties van een frictie-coëfficiënt van de bodem van Antarctica [1] (blauw is laag, rood hoog).

van het grote aantal mogelijke vrijheidsgraden er slechts weinig worden benut. Dat is een gelukkige situatie, want zonder deze eigenschap zou de hoeveelheid beschikbare data veel te klein zijn om de waarheid te achterhalen. Moderne statistische methoden proberen deze beperking te overwinnen door de resultaten van de ene analyse te combineren met die van een andere, zogenaamde *large scale inference*. Ook hier spelen Bayesiaanse ideeën een belangrijke rol.

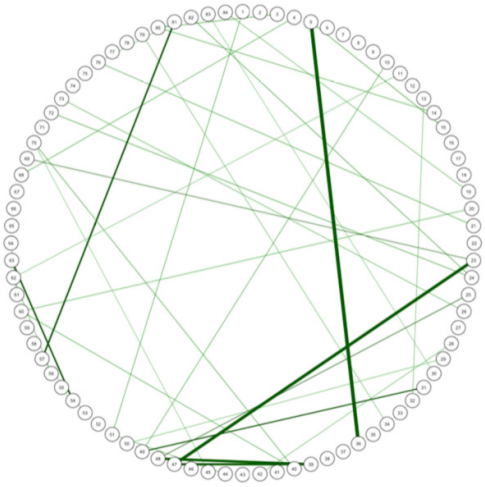
Slotwoord

Slimme data algoritmen kunnen eenvoudige problemen opmerkelijk goed oplossen. Andere problemen over 15 jaar.

Referenties en verder lezen

1. Isaac, Petra, Stadler, Ghattas, 2015, *Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet*, J. Computational Physics, 348-368.
2. Efron, Bradley, 2010, *Large-scale inference. Empirical Bayes methods for estimation, testing, and prediction*, Institute of Mathematical Statistics (IMS) Monographs, 1. Cambridge University Press, Cambridge, xii+263 pp. ISBN: 978-0-521-19249-1.
3. Ghosal, Subhashis ; van der Vaart, Aad, 2017, *Fundamentals of nonparametric Bayesian inference*, Cambridge Series in Statistical and Probabilistic Mathematics, 44. Cambridge University Press, Cambridge, xxiv+646 pp. ISBN: 978-0-521-87826-5.
4. van der Vaart, Aad; van Wieringen, Wessel, 2016, *Statistics in high dimensions*, Mathematics and society, 51--70, Eur. Math. Soc., Zürich.

- - -



Figuur 9: apoptosis netwerk voor 84 genen. De groene lijnen geven de (weinige) significante interacties tussen de genen aan.