Ministerie van Verkeer en Waterstaat

Directoraat – Generaal Rijkswaterstaat

RIZA Rijksinstituut voor Integraal Zoetwaterbeheer en Afvalwaterbehandeling

Bivariate correlatiemodellen met exponentiële en asymptotisch exponentiële marginale verdelingen

RIZA – werkdocument 2002.104x
Auteur: C.P.M. Geerse

RIZA, afdeling WSH
Lelystad, 26 mei 2004
(Definitieve versie van concept uit augustus 2002)

Inhoudsopgave

Samenvatting

Bij Rijkswaterstaat worden voor de berekening van toetspeilen en kruinhoogten probabilistische modellen gebruikt. Daarin is soms sprake van een tweetal gecorreleerde stochasten, die beschreven worden door een *bivariate* kansverdeling. Dit rapport beschrijft een correlatiemodel waarmee deze verdeling bepaald kan worden, op zo'n manier dat enerzijds de correlatie volgens de meetgegevens goed wordt weergegeven en anderzijds de vooraf bekend veronderstelde marginale verdelingen exact door het model worden gerepresenteerd.

In de afleiding van de bivariate verdeling wordt de data op de horizontale as getransformeerd naar een standaardexponentieel verdeelde stochast X en die op de verticale as naar een *asymptotisch* standaardexponentieel verdeelde stochast Y. Het correlatiemodel vormt een verbetering en veralgemenisering van een in onder meer [Vrouwenvelder *et al*, 1999] gebruikt model. Het model uit [Volker, 1987] voor de correlatie tussen hoogwaterstand en windsnelheid te Hoek van Holland vormt daarnaast een speciaal geval van het model. Het nieuwe model zal worden gebruikt in een probabilistisch model voor de Vecht- en de IJsseldelta: het gaat dan om de correlatie tussen Vechtafvoer en IJsselmeerpeil en die tussen IJsselafvoer en IJsselmeerpeil.

De hoofdstukken 3 t/m 6 bevatten de kern van dit rapport. Ze zijn nogal wiskundig van aard en zullen voor de meesten lastig leesbaar zijn. Met het oog op de praktische toepassing van het model zijn daarom de hoofdstukken 2 en 7 geschreven. Die kunnen vrijwel los van de rest worden gelezen: ze geven een overzicht van het model (hoofdstuk 2) en een uitgebreid voorbeeld (hoofdstuk 7) met daarbij veel figuren, en ook tips voor de selectie van data en voor de beoordeling van de modelfit aan de data.

Ter informatie: recent is het rapport [Beijk en Geerse, 2004] verschenen, dat een MATLAB-applicatie beschrijft waarin het correlatiemodel is geïmplementeerd. Dat rapport is bedoeld voor de lezer die primair interesse heeft voor de toepassing van het model. Naast een beknopte modelbeschrijving licht het rapport de werking van de applicatie toe met een concreet voorbeeld.

1 Inleiding

1.1 Algemeen

Bij Rijkswaterstaat worden voor de berekening van toetspeilen en kruinhoogten probabilistische modellen gebruikt. Daarin komen stochasten voor die al of niet gecorreleerd kunnen zijn. De situatie van een tweetal gecorreleerde stochasten komt nogal eens voor. Dan is sprake van een *bivariate* kansverdeling. Denk als voorbeeld aan de tweetallen: IJsselafvoer en IJsselmeerpeil, Rijnafvoer en Maasafvoer, hoogwaterstand te Hoek van Holland en de windsnelheid tijdens dit hoogwater.

Dit rapport beschrijft een mogelijke modellering voor zo'n correlatie. Uitgangspunt is dat de marginale verdelingen van de stochasten reeds bekend zijn en dat tevens wordt beschikt over 'gepaarde' waarnemingen. Dit kunnen dagwaarnemingen zijn (ieder waarnemingspaar bestaat dan uit dagwaarnemingen van beide stochasten) maar ook jaarmaxima of nog een ander soort waarnemingen. Op grond van deze gegevens levert het correlatiemodel dan een bivariate kansverdeling met als marginalen de vooraf gegeven verdelingen, waarbij de spreiding van de bivariate verdeling (mits tenminste een variant van het model de data goed beschrijft) overeenstemt met die in de data. In de afleiding van de bivariate verdeling wordt de data op de horizontale as getransformeerd naar een standaardexponentieel verdeelde stochast X en die op de verticale as naar een *asymptotisch* standaardexponentieel verdeelde stochast Y.

Het correlatiemodel uit dit rapport vormt een verbetering en uitbreiding van een model dat in onder andere [Vrouwenvelder *et al*, 1999] is gebruikt: de verbetering is dat de vooraf bekende marginale verdelingen exáct door het model worden weergegeven en niet alleen in nogal ruwe benadering, de uitbreiding dat (na transformatie) de conditionele verdeling van Y gegeven X een willekeurige vorm mag hebben. Die hoeft niet langer zoals in [Vrouwenvelder *et al,* 1999] de normale verdeling te zijn. Aldus is het nieuwe correlatiemodel nauwkeuriger en algemener dan dat uit laatstgenoemde referentie.

Het nieuwe model zal worden gebruikt in een probabilistisch model voor de Vecht- en de IJsseldelta: het gaat dan om de correlatie tussen Vechtafvoer en IJsselmeerpeil en om de correlatie tussen IJsselafvoer en IJsselmeerpeil.

In een probabilistisch model voor het Benedenrivierengebied, Hydra-B genaamd, is een belangrijke correlatie die tussen de hoogwaterstand te Hoek van Holland en de tijdens dit hoogwater optredende windsnelheid. De bivariate verdeling die deze stochasten beschrijft wordt gewoonlijk aangeduid als de *wind-waterstandstatistiek.* In 1987 is door Volker, zie [Volker, 1987], hiervoor een correlatiemodel opgesteld (in 2002 aangepast voor nieuwe gegevens). Dit rapport demonstreert (zie hoofdstuk 8) dat Volker's model een speciaal geval is van het nieuwe correlatiemodel.

1.2 Achtergronden en leeswijzer

Het grootste deel van dit rapport is reeds begin 2002 geschreven. Directe aanleiding voor het opstellen van een nieuw correlatiemodel was dat de zojuist genoemde wind-waterstandstatistiek vanwege nieuwe gegevens herzien moest worden. Die aanpassing stuitte op problemen. Pas na gedegen analyse van Volker's model was duidelijk hoe die aanpassing te verrichten. Die analyse leerde dat een algemener model mogelijk was dan dat van Volker, dat bovendien aansloot bij het model uit [Vrouwenvelder *et al,* 1999]. Begin 2002 is in eerste instantie een rapport geschreven met de precieze wiskundige formulering van het nieuwe model, tezamen met allerlei verbanden tussen de onderdelen van het model. Wat later bestond de behoefte aan handvatten voor praktisch gebruik. Toen is voor één variant van het model een 'recept' voor de praktische toepassing uitgewerkt. Dat resulteerde in een hoofdstuk met een globale uitleg van het model (hoofdstuk 2), samen met een uitgewerkt voorbeeld voor de praktische toepassing van het recept (hoofdstuk 7). Dat laatste hoofdstuk biedt ook aanwijzingen voor de selectie van data (paragraaf 7.1) en voor de beoordeling van de modelfit aan de data (paragraaf 7.5).

Als service aan de lezer die vooral in de praktische toepassing geïnteresseerd is, zijn de hoofdstukken 2 en 7 zó geschreven dat ze nagenoeg los van de rest kunnen worden gelezen. Die lezer kan ook terecht bij een ander rapport: in [Beijk en Geerse, 2004] wordt een MATLAB-applicatie beschreven waarin het correlatiemodel – of tenminste een variant daarvan – is geïmplementeerd. Dat rapport geeft eveneens een beknopte beschrijving van

het model, tezamen met de meest relevante formules. De werking van de applicatie wordt bovendien toegelicht met een concreet voorbeeld, namelijk het opstellen van de wind-waterstandstatistiek voor de richting Noordwest (uitgaande van een normale verdeling als conditionele verdeling van Y gegeven X).

Hoofdstuk 2 geeft een korte omschrijving van de modelaspecten: de gegevens, transformaties, het eigenlijke model en de toepassing van een variant daarvan. Hoofdstuk 3 geeft een gedetailleerde beschrijving van het model zoals dat er ná transformatie uitziet. Hoofdstuk 4 geeft voor de liefhebber extra informatie. In hoofdstuk 5 wordt het model toegepast op de uniforme verdeling als conditionele verdeling van Y gegeven X – daarvoor kunnen diverse zaken analytisch worden uitgerekend. Hoofdstuk 6 beschrijft het model zoals dat er zónder transformatie uitziet, waarbij paragraaf 6.3 voor de echte liefhebber bestemd is. Hoofdstuk 7 geeft, voorzien van veel plaatjes, de uitwerking van een voorbeeld. Hoofdstuk 8 laat zien dat Volker's model een speciaal geval vormt van het correlatiemodel.

Hoofdstuk 9 staat enigszins los van de rest van het rapport. Daarin wordt, gebaseerd op [Ditlevsen en Madsen, 1996], een algemener correlatiemodel beschreven. In het bijzonder maakt dat hoofdstuk duidelijk dat in zekere zin 'oneindig veel' correlatiemodellen bestaan. De klasse van modellen uit de hoofdstukken 2 t/m 8 vormt daarvan slechts één voorbeeld – zij het dan wel een praktisch erg goed toepasbaar voorbeeld. Paragraaf 9.2 geeft de formules voor een model gebaseerd op de bivariate normale verdeling. Dat model is toegepast in een (inmiddels niet meer gebruikt) probabilistisch model voor de IJsseldelta, waarvan het correlatiemodel nog niet eerder was gerapporteerd.

De appendix geeft bewijzen van beweringen uit hoofdstuk 3 t/m 7.

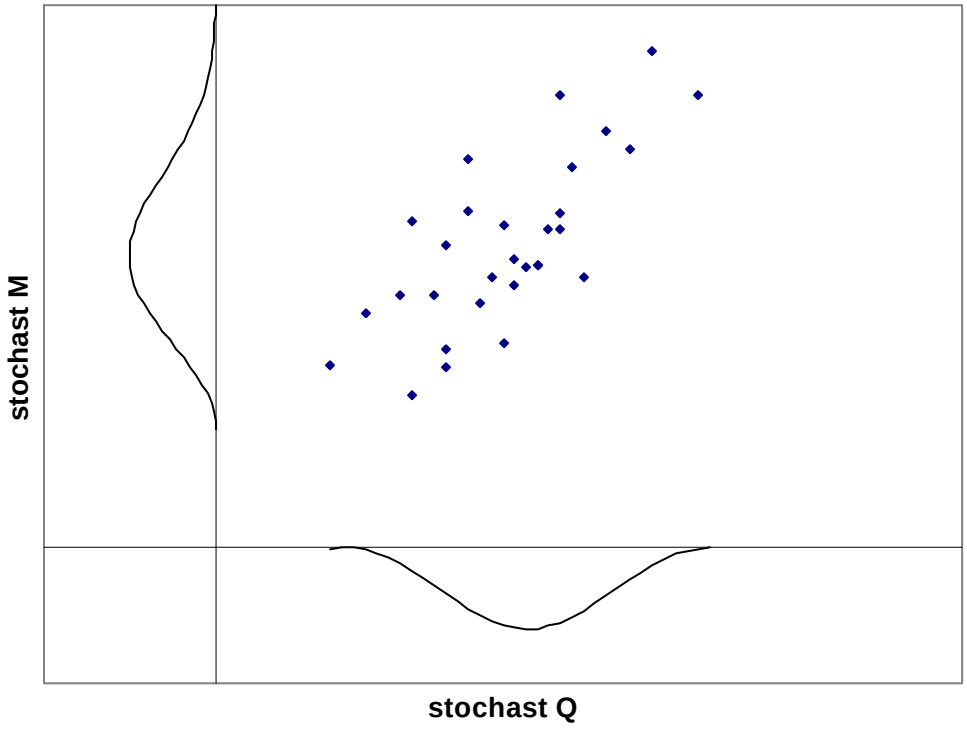
2 Globale beschrijving van correlatiemodel en zijn toepassing

Dit hoofdstuk beschrijft globaal het correlatiemodel uit dit rapport (zie paragraaf 2.1 t/m 2.3) en de toepassing daarvan op praktijksituaties (paragraaf 2.4). Het model wordt veel uitgebreider behandeld in hoofdstuk 3, 4 en 6 en de toepassing daarvan in hoofdstuk 7.

2.1 De gegevens

Het uitgangspunt is dat twee stochasten Q en M gegeven zijn, met reeds bekende kansverdelingen. Denk om het concreet te maken aan Q als de IJsselafvoer en aan M als het IJsselmeerpeil. De kansdichtheden van de stochasten Q en M zullen worden aangeduid als $g_Q(q)$ en $g_M(m)$ of iets korter als g(q) en g(m). Er wordt hier niet verder gespecificeerd op welk soort waarnemingen Q en M betrekking hebben. Dat kunnen dagwaarnemingen zijn maar ook jaarmaxima of nog een ander soort waarnemingen. In het kader van dit stuk is slechts van belang dat de kansdichtheden waaruit de kansen op de waarnemingen van Q en M volgen vooraf gegeven zijn.

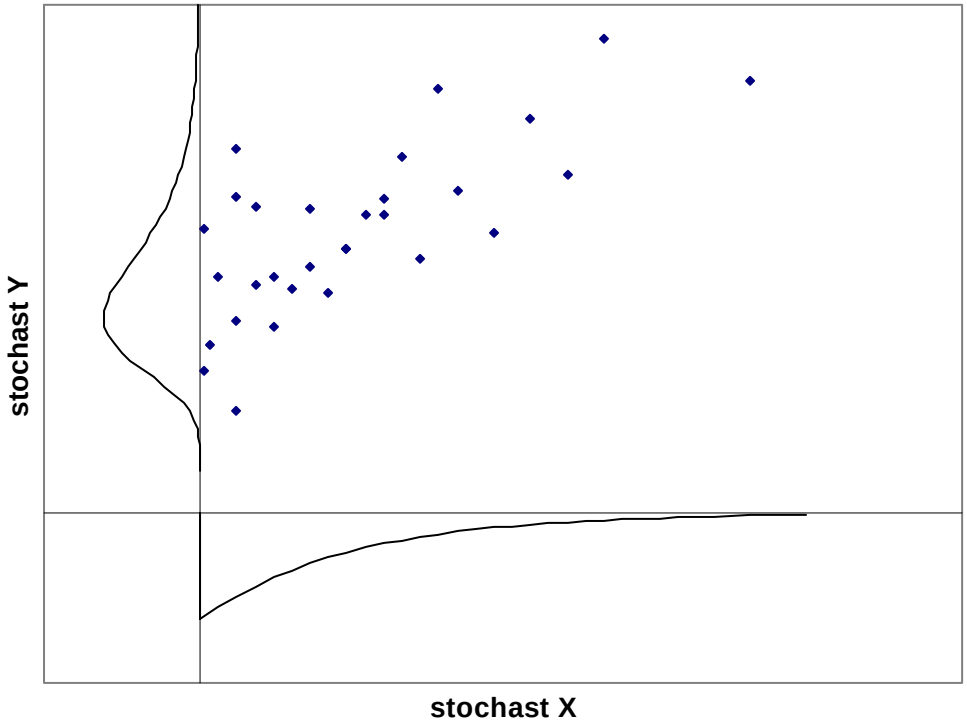
Wanneer de correlatie tussen Q en M onderzocht moet worden, zullen gecombineerde waarnemingen ter beschikking staan van de vorm (q_i, m_i), i = 1, 2, 3, …, N. Voor bijvoorbeeld dagwaarnemingen wordt dan beschikt over N dagen met op dag i een waarneming q_i en een waarneming m_i. Figuur 2.1 geeft een illustratie van de waarnemingen. Op de assen zijn de kansdichtheden schetsmatig aangegeven. Het doel van dit stuk is een gezamenlijke kansdichtheid g(q,m) te vinden die de puntenwolk in figuur 2.1 adequaat beschrijft. Meer specifiek: deze g(q,m) dient als marginale verdelingen de vooraf gegeven g(q) en g(m) te hebben en dient de correlatie tussen Q en M voldoende nauwkeurig te beschrijven.



Figuur 2.1 Illustratie van de puntenwolk bestaande uit N datapunten met op de assen de kansdichtheden g(q) en g(m) aangegeven.

2.2 De transformaties

De methode die in dit stuk gevolgd wordt om g(q,m) te vinden bestaat er onder meer uit eerst Q en M te transformeren naar nieuwe stochasten X en Y. Die transformaties worden voor Q en M ieder apart uitgevoerd. Hierbij wordt Q getransformeerd naar X, waarbij X een standaardexponentiële verdeling volgt. De g(q) gaat dan over in de kansdichtheid $g(x) = \exp(-x)$ voor $x \geq 0$. De M wordt getransformeerd naar Y, waarbij g(m) overgaat in de kansdichtheid g(y). Hieronder zal worden uitgelegd hoe g(y) kan worden berekend. Deze transformaties kunnen ruwweg worden opgevat als het uitrekken en/of indrukken van de assen, op zo'n manier dat g(q) en g(m) worden 'vervormd' tot de nieuwe kansdichtheden g(x) en g(y). Daarbij gaan de N waarnemingen (q_i, m_i) over in N getransformeerde waarnemingen (x_i, y_i), zie figuur 2.2 ter illustratie[1]. In de figuur is te zien hoe (in dit voorbeeld) de transformatie van g(q) naar g(x) de waarnemingen enigszins 'doet ophopen' in de buurt van x = 0, waar g(x) zijn grootste kansbijdragen heeft, terwijl de meest naar rechts gelegen punten in de transformatie nog iets verder naar rechts komen te liggen omdat g(x) een langere staart heeft dan g(q). Omdat g(m) en g(y) wat meer op elkaar lijken dan g(q) en g(x) is het effect van de transformatie van M naar Y iets minder duidelijk te zien in figuur 2.2. Samenvattend kan gesteld worden dat onder de genoemde transformaties het (q,m)-vlak wordt getransformeerd naar het (x,y)-vlak. Daarbij worden niet alleen de kansdichtheden getransformeerd maar ook ieder datapunt. Het (q,m)-vlak zal in dit stuk worden aangeduid als het *originele vlak* en het (x,y)-vlak als het *getransformeerde vlak*.



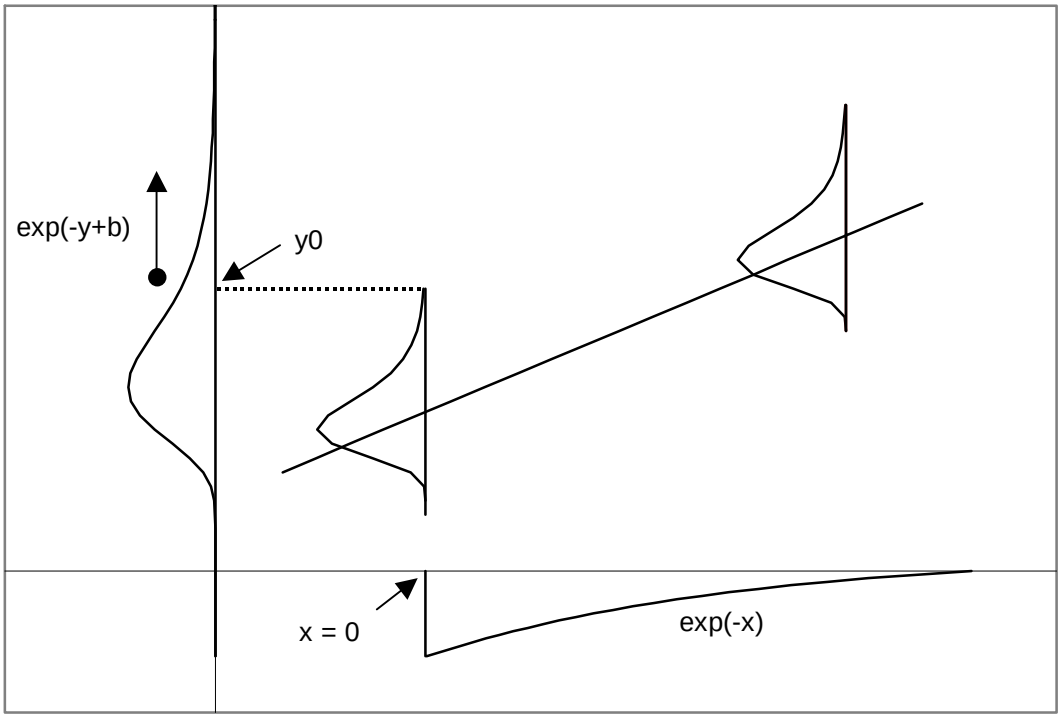
Figuur 2.2 De getransformeerde puntenwolk bestaande uit de N datapunten met op de X-as de standaardexponentiële kansdichtheid en op de Y-as de kansdichtheid g(y) zoals beschreven in de tekst.

2.3 Het correlatiemodel

Om aan g(q,m) te komen wordt een kansdichtheid g(x,y) in het getransformeerde vlak aangenomen die als marginale verdelingen g(x) en g(y) heeft. Door de inverse transformaties toe te passen op g(x,y) volgt dan een formule voor g(q,m) die als marginale verdelingen de gewenste g(q) en g(m) heeft.

[1] Figuur 2.2 geeft slechts schetsmatig aan hoe de puntenwolk er na transformatie uitziet. Indien de coördinaten van de getransformeerde punten exact zouden worden uitgerekend zou de puntenwolk er iets anders uitzien.

We geven nu aan hoe g(x,y) er uit ziet. Zie daartoe figuur 2.3 (de interpretatie van y_0 wordt hieronder nog uitgelegd). Algemeen geldt dat g(x,y) = g(x)g(y|x) waarbij g(y|x) de conditionele verdeling van Y gegeven X = x aangeeft. Omdat g(x) = exp(-x) bekend is, volgt dus dat g(x,y) volledig vastligt wanneer g(y|x) gegeven is. We zullen g(y|x) op een verschuiving langs een rechte lijn na gelijk kiezen aan een vaste kansdichtheid λ(y). In figuur 2.3 zijn twee van deze conditionele verdelingen aangegeven. Tevens is de rechte lijn aangegeven waarlangs λ(y) verschoven moet worden om aan g(y|x) te komen. Meer expliciet kunnen we schrijven g(y|x) = λ(y-x-δ) waarin δ een willekeurig maar vast gekozen reëel getal is. Omdat nu g(x,y) vastligt, ligt daarmee tevens de marginale verdeling g(y) vast, welke in figuur 2.3 op de verticale as is aangegeven. De g(y) is afhankelijk van de beschouwde λ(y) en van de waarde van δ. De transformatie van M naar Y hangt dus eveneens af van λ(y) en δ. Hiermee is die transformatie iets gecompliceerder dan die van Q naar X, omdat de laatstgenoemde transformatie direct kan worden uitgevoerd, onafhankelijk van de beschouwde λ(y) en δ.



Figuur 2.3 Het correlatiemodel in de getransformeerde ruimte. De conditionele verdelingen g(y|x) volgen door een verschuiving langs de rechte lijn uit de kansdichtheid λ(y) volgens de formule g(y|x) = λ(y-x-δ). De λ(y) heeft in dit voorbeeld een eindpunt y_e (eindige staart): vanaf $y_0 = y_e + \delta$ geldt dan g(y) = exp(-y+b).

In dit stuk zal worden aangetoond dat voor kansdichtheden λ(y) waarvan de rechterstaart voldoende snel afneemt, de rechterstaart van de verdeling g(y) in benadering een exponentiële verdeling volgt met schaalparameter gelijk aan 1. Voor grote y geldt dan in benadering g(y) = exp(-y + b). Een voldoende voorwaarde hiervoor is dat de rechterstaart van λ(y) zo snel afneemt dat de verwachtingswaarde van de functie $f(t) = e^t$ eindig is. Bijvoorbeeld de normale verdeling voldoet hieraan. Verder voldoet uiteraard iedere λ(y) die 0 wordt vanaf een zekere waarde y_e aan de hier gestelde eis. Het getal y_e geeft dan het *eindpunt* aan van de kansdichtheid λ(y). Voorbeelden van verdelingen met zo'n eindige staart zijn de uniforme en de β-verdeling. Wanneer λ(y) gelijk aan 0 wordt vanaf een waarde y_e, geldt zelfs dat g(y) niet alleen in benadering maar zelfs exáct gelijk wordt aan een exponentiële verdeling, en wel vanaf de waarde $y_0 = y_e + \delta$. Dat is de waarde van y waar de kansdichtheid g(y|x=0) gelijk wordt aan 0, zie figuur 2.3 ter illustratie. Voor waarden van y die in figuur 2.3 boven de stippellijn liggen, geldt dus exact dat g(y) = exp(-y + b).

Door δ gelijk te kiezen aan een geschikte waarde δ_0 is het altijd mogelijk ervoor te zorgen dat de locatieparameter b in bovenstaande gelijk wordt aan 0, in welk geval voor grote y exact dan wel in benadering geldt g(y) = exp(-y). In dat geval heeft het correlatiemodel in de getransformeerde ruimte de eigenschap dat g(x)

standaardexponentieel is en dat g(y) asymptotisch standaardexponentieel is. Door de inverse transformaties toe te passen op g(x,y) volgt zoals eerder vermeld de kansdichtheid g(q,m). Het blijkt dat de keuze van δ geen invloed heeft op de uitkomst voor g(q,m)! Daarom wordt ten behoeve van toepassingen steeds de waarde $\delta = \delta_0$ gekozen, zodat de locatieparameter b altijd gelijk aan 0 is.

2.4 Toepassing van het model

Wanneer in een toepassing twee stochasten met hun marginale verdelingen ter beschikking staan, moet eerst een keuze worden gemaakt welke naar X wordt getransformeerd en welke naar Y. Anders gezegd, eerst moet worden uitgemaakt (bij de hier gebruikte terminologie) welke als Q en welke als M wordt aangeduid. Beide keuzes lijken even goed werkbaar, zie desgewenst paragraaf 7.1 voor nader commentaar. Die paragraaf geeft ook advies over hoe de data te selecteren. Losjes gezegd dienen bij conditionering op q álle waarnemingen tot de selectie te behoren, en niet slechts een deel – anders resulteert een g(q,m) met een foutieve spreiding.

Het model is bedoeld voor de modellering van een *positieve* correlatie tussen Q en M. Indien sprake is van een negatieve correlatie, dient Q of M te worden voorzien van een minteken. In die situatie is sprake van een positieve correlatie, waarop het model dan weer van toepassing is (zie ook paragraaf 7.1).

In een concrete toepassing dient een keuze voor $\lambda(y)$ te worden gemaakt. Door de inverse transformaties toe te passen op g(x,y) volgt dan de kansdichtheid g(q,m) die de data op de juiste manier dient te beschrijven. In feite gaat het om een iteratief proces! De transformatie die M overvoert in Y kan namelijk pas uitgevoerd worden nadat $\lambda(y)$ gekozen is. Echter: *om te beoordelen of $\lambda(y)$ juist gekozen is dient de getransformeerde puntenwolk te worden beschouwd*. Het getransformeerde plaatje laat immers zien of alle g(y|x) inderdaad op een verschuiving langs een rechte lijn na gelijk zijn en of inderdaad de (verticale) spreiding overal gelijk is. In een toepassing zullen dus in principe meerdere keuzes van $\lambda(y)$ beschouwd moeten worden, met daarnaast steeds de puntenwolk in het getransformeerde vlak.

Ter informatie van de lezer: met het beschouwde correlatiemodel kan lang niet iedere g(q,m) worden verkregen. Anders gezegd, er zijn heel veel formules voor g(q,m) te geven met de voorgeschreven marginale verdelingen g(q) en g(m), terwijl g(q,m) *niet* door een transformatie samenhangt met de beschouwde g(x,y). Mogelijk stemmen de data dus voor geen enkele keuze van $\lambda(y)$ overeen met het correlatiemodel. Wanneer bijvoorbeeld voor *elke* keuze van $\lambda(y)$ de puntenwolk in het getransformeerde vlak een spreiding vertoont die toeneemt met x, is het model niet van toepassing. Eveneens is het niet van toepassing als voor *elke* keuze van $\lambda(y)$ de g(y|x) in de getransformeerde ruimte niet door verschuiving langs een *rechte* lijn samenhangen, maar slechts door verschuiving langs een *niet-rechte* lijn.

In de praktijk is het vaak een kwestie van oordeelkundig inzicht of het hier beschouwde correlatiemodel de data voldoende nauwkeurig beschrijft. Ook als dat wel het geval is zal een eenduidige keuze van $\lambda(y)$ niet altijd te maken zijn: verschillende keuzes kunnen min of meer even goed de data beschrijven. De normale verdeling zal als keuze voor $\lambda(y)$ vaak de voorkeur verdienen. Eén reden is de grote bekendheid van deze verdeling. Daarnaast blijkt voor deze verdeling te gelden dat niet alleen g(y|x) maar ook de 'omgekeerde' verdeling g(x|y) dan voor grote y een normale verdeling volgt, waarbij dat laatste exact dan wel in benadering geldt.

De wiskundige aspecten van de bovengenoemde resultaten komen uitgebreid aan de orde in hoofdstuk 3, 4 en 6. In hoofdstuk 7 wordt een manier aangegeven om het model toe te passen in de praktijk. Daarbij wordt aangenomen dat de beschouwde $\lambda(y)$ behoren tot één type verdeling, bijvoorbeeld de normale of de uniforme verdeling. Binnen de beschouwde klasse van verdelingen verschillen de $\lambda(y)$ slechts van elkaar door hun standaarddeviatie s, terwijl elke $\lambda(y)$ gemiddelde nul heeft; ruwweg gezegd hebben alle verdelingen dezelfde vorm maar verschillen ze door hun spreiding. Aldus beschouwen we een klasse van verdelingen die we aangeven met $\lambda_s(y)$, voor $s > 0$, waarbij iedere $\lambda_s(y)$ van hetzelfde type is en gemiddelde nul en standaarddeviatie s heeft. De verdeling $\lambda_s(y)$ volgt eenvoudig uit die voor $s = 1$ met de formule $\lambda_s(y) = (1/s)*\lambda_1(y/s)$. Als we bijvoorbeeld de klasse van de normale verdelingen beschouwen, stelt $\lambda_s(y)$ de normale verdeling voor met gemiddelde nul en standaarddeviatie s. De parameter s is dan – bij beschouwen van één type verdeling voor $\lambda(y)$ – de enige vrijheidsgraad in het model. Iedere waarde van s correspondeert met een gezamelijk kansdichtheid $g_s(q,m)$ die de voorgeschreven marginale verdelingen g(q) en g(m) oplevert. Voor hele kleine waarden van s is de correlatie tussen Q en M zeer sterk, voor hele grote zeer zwak. In de limiet s nadert tot oneindig blijkt $g_s(q,m)$ gelijk te worden aan g(q)g(m), wat correspondeert met onafhankelijkheid van Q en M.

Wanneer de data door het correlatiemodel – voor het beschouwde type verdeling – voldoende nauwkeurig beschreven kunnen worden, zal een 'juiste waarde' van s bestaan, namelijk degene waarvoor de spreiding in $g_s(q,m)$ overeenstemt met die in de data. Hoofdstuk 7 geeft in dat geval twee manieren om deze waarde te vinden. De eerste betreft een *iteratieproces* waarbij successievelijk standaarddeviaties s_1, s_2, s_3, … worden bepaald: de limietwaarde is de gezochte s. De tweede manier is de *Maximum Likelyhood* methode (ML-methode), waarbij s zo wordt gekozen dat de kans op de opgetreden waarnemingen maximaal wordt.

Het gebruik van beide methodes wordt becommentarieerd in paragraaf 7.5. Eigenlijk is het wat misleidend te spreken over de 'juiste waarde' van s. De werkelijkheid, dat wil zeggen de werkelijke g(q,m), wordt nooit exact door het beschouwde correlatiemodel beschreven (behalve dan in simulaties). Dus is er ook geen waarde van s die de data exact beschrijft. Hooguit is sprake van een waarde die de data voldoende nauwkeurig beschrijft. Hierdoor hebben de methodes om s te bepalen een beperkt nut. Het blijft altijd een kwestie van oordeelkundig inzicht of het type kansverdeling, in combinatie met de gevonden waarde van s, de data voldoende nauwkeurig beschrijft. Een advies aan de gebruiker is de met het iteratieproces of met de ML-methode gevonden waarde 'handmatig' aan te passen als dat nodig lijkt.

3 Het correlatiemodel in de getransformeerde ruimte

In dit hoofdstuk wordt een correlatiemodel voor stochasten X en Y beschouwd van de volgende vorm. De verdeling van X is standaardexponentieel. De conditionele verdeling van Y gegeven X is op een verschuiving langs een rechte lijn na gelijk aan een vaste kansdichtheid $\lambda(t)$. Deze twee verdelingen leggen de gezamenlijke kansdichtheid g(x,y) van X en Y vast. De kansdichtheid $\lambda(t)$ kan een eindige zowel als een onbegrensde drager hebben. In toepassingen zal het correlatiemodel gebruikt worden nadat de 'oorspronkelijke' stochasten, bijvoorbeeld de rivierafvoer Q en het IJsselmeerpeil M, eerst naar X en Y getransformeerd zijn. De formules in dit hoofdstuk betreffen dus een gezamenlijke kansdichtheid in een 'getransformeerde ruimte'. De genoemde transformatie komt in hoofdstuk 6 aan de orde.

3.1 Beschrijving van het correlatiemodel

In deze paragraaf worden wat definities gegeven en wordt het correlatiemodel gedetailleerd beschreven. Beschouw de stochast X met de standaard exponentiële verdeling

$$(3.1) \qquad F_X(x) = \begin{cases} 1 - e^{-x} & , x > 0 \\ 0 & , x \le 0 \end{cases}$$

en kansdichtheid

$$(3.2) \qquad g_X(x) = \begin{cases} e^{-x} & , x > 0 \\ 0 & , x \le 0 \end{cases}$$

Soms zal gemakshalve indien geen verwarring kan ontstaan de index X worden weggelaten en wordt bijvoorbeeld geschreven F(x) in plaats van $F_X(x)$. Voor de overschrijdingskans wordt geschreven $\overline{F}(x) = 1 - F(x) = P(X>x)$. We beschouwen een kansdichtheid $\lambda(t)$ met gemiddelde $\mu = 0$ en standaarddeviatie $s > 0$. De cumulatieve verdelingsfunctie van $\lambda(t)$ wordt gegeven door

$$(3.3) \qquad \Lambda(y) = \int_{-\infty}^{y} \lambda(t)dt$$

In dit stuk wordt aangenomen dat $\lambda(t)$ een continue functie is, behalve eventueel in eindig veel punten. In dat geval is $\Lambda(y)$ een continue functie, zie bijvoorbeeld [Billingsley, 1995; pagina 400, 401]. (In het bijzonder heeft een 'geïsoleerde uitkomst' altijd kans 0). De afgeleide van $\Lambda(y)$ bestaat dan overal, behalve in een eventueel discontinuïteitspunt van $\lambda(y)$, en is gelijk aan $\Lambda'(y) = \lambda(y)$. De reden dat discontinue $\lambda(t)$ worden toegelaten is dat we voor $\lambda(t)$ onder meer de uniforme verdeling alsmede een 'afgeknotte Gumbelverdeling' (zie hoofdstuk 8) willen kunnen beschouwen; deze verdelingen hebben discontinuïteitspunten. Verder wordt aangenomen dat $\lambda(t)$ begrensd is, dus dat geldt voor zekere C dat $\lambda(t) < C$ voor alle t. In nagenoeg alle praktische toepassingen zal aan deze begrensdheid voldaan zijn. Daarnaast wordt aangenomen dat de linker- en rechterstaart van $\lambda(t)$ geen 'grillige hobbels' vertonen, in de zin dat voor grote $|t|$ moet gelden dat $\lambda(t) < K/t^c$ voor zekere $c > 1$ en zekere K. Dit is een uitermate milde voorwaarde waar in praktische toepassingen (bijna) altijd aan voldaan zal zijn; ter illustratie merken we op dat een kansdichtheid $\lambda(t)$ waarvoor $\lambda(t) = K/t^2$ voor $|t| > t_1$ zulke zware staarten heeft dat het gemiddelde en de standaarddeviatie van de verdeling niet bestaan maar oneindig zijn. De voorgaande voorwaarden kunnen worden gekenschetst als zeer mild, in de zin dat in enigermate reguliere toepassingen daar altijd aan voldaan zal zijn. We stellen ook een strengere eis aan $\lambda(t)$, namelijk dat de rechterstaart van $\lambda(t)$ zo snel afneemt dat de verwachtingswaarde van de functie e^t eindig is. Wanneer deze verwachtingswaarde wordt aangegeven met $E(e^T)$ stellen we dus de eis dat $E(e^T) < \infty$.

De drager van D van $\lambda(t)$ wordt hier gedefinieerd als

(3.4) $$D = \{t \mid \lambda(t) > 0\}$$

We zullen alleen kansdichtheden $\lambda(t)$ beschouwen waarvoor de drager een aaneengesloten open interval vormt dat al of niet begrensd kan zijn. Dat interval zal worden aangegeven met (y_b, y_e), waarbij de indices b en e slaan op het **b**egin en **e**ind van de drager. Losjes gezegd 'leeft' de kansverdeling op het interval (y_b, y_e), omdat buiten dit interval $\lambda(t) = 0$. Bijvoorbeeld voor de uniforme verdeling geldt $(y_b, y_e) = (-s\sqrt{3}, s\sqrt{3})$ en voor de normale verdeling $(y_b, y_e) = (-\infty, \infty)$. De restrictie aan $\lambda(t)$ dat de drager een *open* verzameling vormt dient slechts om de wiskundige notatie van een en ander eenvoudig te houden, maar vormt geen wezenlijke restrictie. Indien een kansdichtheid een niet-open interval als drager heeft hoeft deze slechts op het linker- en/of rechterpunt van het interval te worden aangepast om een open interval als drager te krijgen. Een dergelijke aanpassing laat $\Lambda(y)$ onveranderd en is daarmee niet van (wezenlijke) invloed op de beweringen in dit stuk. Samenvattend dient $\lambda(t)$ in dit stuk altijd (minstens) te voldoen aan de volgende voorwaarden[4].

Voorwaarden aan $\lambda(t)$

(3.5) $\lambda(t)$ is begrensd en continu behalve eventueel in eindig veel discontinuiteitspunten

(3.6) De drager D van $\lambda(t)$ vormt een aaneengesloten open interval

(3.7) Het gemiddelde μ van $\lambda(t)$ is gelijk aan 0 en de standaarddeviatie $s > 0$ bestaat als eindig getal

(3.8) De staarten van $\lambda(t)$ mogen niet te veel onregelmatigheden vertonen in de zin dat voor zekere $t_1 < \infty$ en $c > 1$ geldt

$$\lambda(t) < \frac{K}{t^c} \qquad \text{voor } |t| > t_1$$

(3.9) $$E(e^T) = \int e^t \lambda(t)\, dt < \infty$$

De voorwaarden (3.5) t/m (3.8) gelden voor dit hele stuk en worden altijd (meestal stilzwijgend) verondersteld. Voorwaarde (3.9) zal in hoofdstuk 4, 6 en 7 iets worden aangescherpt. Voor een naar boven begrensde drager ($y_e < \infty$) is duidelijk dat aan (3.9) is voldaan. Voor de normale verdeling is eveneens aan deze voorwaarde voldaan (overigens ook aan de andere voorwaarden indien de verdeling gemiddelde 0 heeft). Er geldt namelijk

(3.10)
$$\begin{aligned} E(e^T) &= \int e^t \frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{t^2}{2s^2}\right) dt \\ &= \exp\left(\frac{s^2}{2}\right) \frac{1}{s\sqrt{2\pi}} \int \exp\left(-\frac{(t-s^2)^2}{2s^2}\right) dt \\ &= \exp\left(\frac{s^2}{2}\right) \end{aligned}$$

Een voorbeeld van een verdeling waarvoor niet is voldaan aan (3.9) is de exponentiële verdeling met $s \geq 1$. De staart van de verdeling neemt dan te langzaam af, waardoor de verwachtingswaarde van e^t gelijk aan oneindig wordt. Voor $0 < s < 1$ is $E(e^T) = e^{-s}/(1-s)$ wel eindig. Voor dit type verdeling bepaalt de waarde van s dus of $E(e^T)$ al dan niet eindig is.

[4] De $E(e^T)$ kan in verband worden gebracht met de zogenaamde *moment generating function*, zie bijvoorbeeld [Cassella, 1990], indien één type verdeling voor $\lambda(t)$ wordt beschouwd met verschillende standaarddeviaties. Geef met $\lambda_1(t)$ de kansdichtheid met standaarddeviatie 1 aan en met $\lambda_s(t) = \lambda_1(t/s)/s$ de kansdichtheid met standaarddeviatie s. Dan is de verwachtingswaarde $E_s(e^T)$ met betrekking tot $\lambda_s(t)$ gelijk aan de moment generating function M(s) van de verdeling van $\lambda_1(t)$, dus $M(s) = E_s(e^T)$.

Beschouw voor een vast reëel getal δ de gezamenlijke kansdichtheid g(x,y), met marginale verdeling g(x), waarvan de conditionele verdeling g(y|x) gegeven wordt door

(3.11) $g(y \mid x) = \lambda(y - x - \delta)$

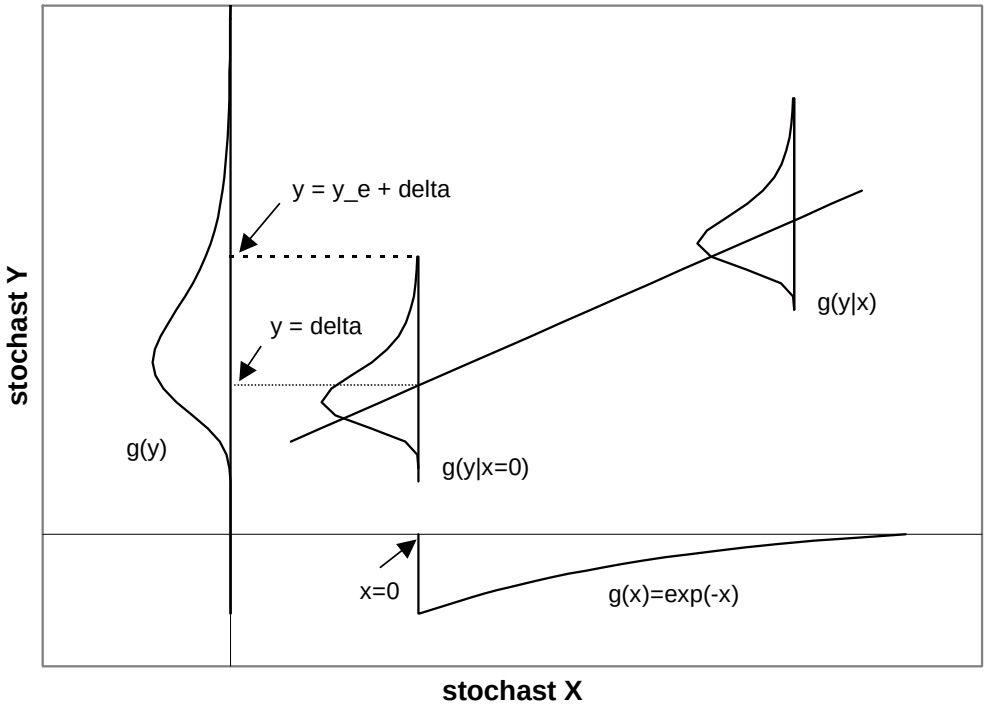
Zie figuur 3.1 ter illustratie. De gezamenlijke kansdichtheid wordt dan gegeven door, voor x > 0,

(3.12) $g(x, y) = e^{-x}\lambda(y - x - \delta)$

terwijl g(x,y) = 0 voor x ≤ 0. Formule (3.11) zegt dat de kansdichtheid van Y bij gegeven X = x wordt verkregen uit de kansdichtheid λ(y) door deze te verschuiven over een afstand x + δ. Omdat λ(y) per definitie gemiddelde 0 heeft, volgt dat het gemiddelde van Y bij gegeven X = x wordt gegeven door

(3.13) $E(Y \mid X = x) = x + \delta$

De gemiddelden van de conditionele kansdichtheden g(y|x) liggen dus op de lijn y(x) = x + δ.



Figuur 3.1 Illustratie van het correlatiemodel met daarin aangegeven de marginale verdelingen g(x) =e^{-x} en g(y). De conditionele verdelingen g(y|x) zijn op een verschuiving na gelijk aan de kansdichtheid λ(t). De gemiddelden E(Y|X=x) liggen op de aangegeven rechte y(x) = x + δ. In de figuur is aangenomen dat $y_e < \infty$. De horizontale stippellijn geeft de waarde $y = y_e + \delta$ aan. In paragraaf 3.2 wordt uitgelegd dat voor $y \geq y_e + \delta$ de verdeling van Y een exponentiële verdeling volgt.

We zullen losjesweg zeggen dat het *correlatiemodel* voor de stochasten X en Y wordt vastgelegd door de functies g(x) en g(y|x). In feite is dit niets anders dan zeggen dat het correlatiemodel wordt vastgelegd door g(x,y). Ofwel, iedere gezamenlijke kansdichtheid definieert een correlatiemodel. Toch is het handig om te denken over een correlatiemodel als vastgelegd door een marginale verdeling enerzijds en een conditionele verdeling anderzijds.

3.2 De marginale verdeling van Y

Het is een bekend feit, zie bijvoorbeeld [Vrouwenvelder et al, 1999], dat ingeval $\lambda(y)$ een normale verdeling volgt de verdeling van Y net als die van X voor grote y in *benadering* een exponentiële verdeling volgt. In het vervolg zullen we laten zien dat deze eigenschap niet alleen voor de normale verdeling geldt maar vrij algemeen geldig is, mits tenminste de rechterstaart van de kansdichtheid $\lambda(t)$ voldoende snel afneemt. Ingeval $\lambda(t)$ een naar boven begrensde drager heeft geldt zelfs dat vanaf een zekere vaste waarde van y de verdeling van Y *exact* exponentieel wordt.

De verdeling $F_Y(y)$ van Y wordt gegeven door g(x,y) eerst te integreren over x van 0 tot ∞ en de resulterende marginale verdeling vervolgens te integreren van $-\infty$ tot y. Na enige herschrijving volgt dan

$$F_Y(y) = \int_0^\infty e^{-x}\Lambda(y-x-\delta)dx \tag{3.14}$$

Het is eenvoudig te verifiëren dat de drager van de verdeling van Y gelijk is aan $(y_b + \delta, \infty)$. Voor $y \le y_b + \delta$ geldt dus $F_Y(y) = 0$. Tevens valt te verifiëren dat $F_Y(y)$ strikt stijgend is op $(y_b + \delta, \infty)$. Voor de overschrijdingskans kan, met behulp van de substitutie $t = y - x - \delta$, worden geschreven

$$\begin{aligned}\overline{F}_Y(y) &= \int_0^\infty e^{-x}\overline{\Lambda}(y-x-\delta)dx \\ &= e^{\delta-y}\int_{-\infty}^{y-\delta} e^t\overline{\Lambda}(t)dt\end{aligned} \tag{3.15}$$

Omdat de integrand in het laatste lid continu is volgt dat (3.15) differentieerbaar is naar y. Voor de kansdichtheid g(y) kan dan worden geverifieerd

$$\begin{aligned}g(y) &= e^{\delta-y}\int_{-\infty}^{y-\delta} e^t\overline{\Lambda}(t)dt \; - \overline{\Lambda}(y-\delta) \\ &= \Lambda(y-\delta) - F_Y(y)\end{aligned} \tag{3.16}$$

Formule (3.15) en (3.16) gelden voor alle $y \in \mathbb{R}$. Zowel $\overline{F}_Y(y)$ als g(y) zijn continu op heel $\mathbb{R}$, waarbij voor $y \le y_b + \delta$ geldt $\overline{F}_Y(y) = 1$ en $g(y) = 0$.

In hoofdstuk A.3 van de appendix wordt aangetoond dat voorwaarde (3.9) equivalent is aan de voorwaarde

$$\int e^t\overline{\Lambda}(t)\,dt < \infty \tag{3.17}$$

Als aan (3.9) (of aan (3.17)) is voldaan volgt door partieel te integreren dat

$$\int e^t\overline{\Lambda}(t)\,dt = E(e^T) \tag{3.18}$$

Omdat $f(t) = e^t$ strikt convex is en omdat $\lambda(t)$ gemiddelde 0 heeft volgt vanwege Jensens ongelijkheid, zie Stelling A.2.7 uit de appendix, dat

$$E(e^T) > e^{E(T)} = 1 \tag{3.19}$$

We kunnen daarom definiëren

$$\delta_0 = -\ln\left(E(e^T)\right) < 0 \tag{3.20}$$

Voor de normale verdeling met standaarddeviatie s volgt uit (3.10) dat

(3.21) $$\delta_0 = -\frac{s^2}{2}$$

Situatie $y_e < \infty$

We zullen nu laten zien dat ingeval $\lambda(t)$ een naar boven begrensde drager heeft de verdeling van Y vanaf zekere y exact exponentieel wordt. Beschouw $y_e < \infty$ en $y \geq y_e + \delta$. Dan geldt voor de bovengrens in de laatste integraal van (3.15) dat deze groter is dan y_e en hoeft, omdat de integrand 0 wordt voor $t > y_e$, slechts geïntegreerd te worden tot aan de waarde y_e. Voor $y \geq y_e + \delta$ geldt dus vanwege (3.18)

$$\begin{aligned} \overline{F}_Y(y) &= e^{\delta - y} \int_{-\infty}^{y_e} e^t \overline{\Lambda}(t) dt \\ &= e^{\delta - y} E(e^T) \\ &= e^{-y+\delta-\delta_0} \end{aligned} \tag{3.22}$$

wat inhoudt dat voor grote y de verdeling van Y exponentieel wordt met schaalparameter 1 en locatieparameter $\delta - \delta_0$. Voor de keuze $\delta = \delta_0$ wordt de locatieparameter gelijk aan 0. In dat geval geldt dus dat de verdeling van Y voor grote y, namelijk voor $y \geq y_e + \delta_0$, *standaard*exponentieel wordt. Merk wel op dat y_e impliciet van de beschouwde standaarddeviatie s van $\lambda(y)$ afhangt, waarbij een grotere s ook een grotere y_e betekent. Bijvoorbeeld in het geval dat $\lambda(y)$ de uniforme verdeling is, geldt $y_e = s\sqrt{3}$.

Situatie $y_e = \infty$

We beschouwen nu de situatie $y_e = \infty$. In dat geval volgt uit (3.15) dat de verdeling van Y nooit exact exponentieel kan zijn, omdat de integraal in het laatste lid van (3.15) dan altijd van y blijft afhangen. Omdat uit (3.18) volgt dat

(3.23) $$\lim_{y \to \infty} \int_{-\infty}^{y-\delta} e^t \overline{\Lambda}(t) dt = E(e^T)$$

kan die integraal voor grote y wel worden benaderd als

(3.24) $$\int_{-\infty}^{y-\delta} e^t \overline{\Lambda}(t) dt \cong E(e^T) \quad \text{, voor } y \text{ groot}$$

in welk geval uit (3.15) en (3.20) volgt

(3.25) $$\overline{F}_Y(y) \cong e^{\delta - y} E(e^T) = e^{-y+\delta-\delta_0} \quad \text{, voor } y \text{ groot}$$

De kwaliteit van de benadering wordt bepaald door de snelheid waarmee de limiet in (3.23) tot $E(e^T)$ nadert. Dat zal des te sneller gebeuren naarmate de staart van $\lambda(t)$ sneller afneemt. Indien dat slechts zeer langzaam gebeurt zal de benadering vrij slecht zijn en zal (3.25) voor de praktijk van weinig nut zijn. De kwaliteit van de benadering hangt dus samen met het type kansverdeling dat beschouwd wordt. Daarnaast zal een grotere waarde van de standaarddeviatie s van $\lambda(t)$ een langzamere afname van de staart van de verdeling leveren en daarmee een minder goede benadering vormen dan een kleinere waarde van s.

Door differentiatie van (3.25) volgt voor de kansdichtheid

(3.26) $$g(y) \cong e^{-y+\delta-\delta_0} \quad \text{, voor } y \text{ groot}$$

We merken nog op dat uit (3.16), (3.18) en (3.20) volgt dat voor alle y geldt (bedenk dat $y_e = \infty$)

(3.27) $g(y) < e^{-y+\delta-\delta_0}$, voor alle y

Naarmate y groter wordt nadert g(y) dus ‘vanaf de onderkant’ naar het rechterlid in (3.26). Voor alle duidelijkheid melden we dat voor $y_e < \infty$ vanaf $y_e + \delta$ het gelijkteken in (3.27) zal gelden.

Exponentiële verdeling voor λ(t) met standaarddeviatie ≥ 1

Zoals eerder opgemerkt is voor λ(t) gelijk aan de exponentiële verdeling met s ≥ 1 niet aan voorwaarde (3.9) voldaan. De verdeling van Y kan (in ieder geval) voor s = 1 wel analytisch worden berekend, voor alle waarden van y. Indien

(3.28) $\lambda(t) = e^{-(t+1)}$, $t > -1$

wordt g(y) gegeven door

(3.29) $g(y) = (y - \delta + 1)e^{-y+\delta-1}$, $y > \delta - 1$

Het bewijs vergt enig schrijfwerk en wordt aan de lezer overgelaten. De g(y) is in dit geval geen exponentiële verdeling maar een zogeheten gammaverdeling (met vormparameter 2 en schaalparameter 1). Deze verdeling heeft een ‘dikkere staart’ dan de exponentiële verdeling, wat komt omdat de hier beschouwde λ(t) een dikkere staart heeft dan de hiervoor beschouwde kansdichtheden λ(t) die aan (3.9) voldoen.

4 Conditionele verdeling van X gegeven Y en samenvatting model

In hoofdstuk 3 zijn de verdelingen g(y) en g(y|x) behandeld. In paragraaf 4.1 wordt g(x|y) behandeld. Deze nagal wiskundig-technische en wat lastige paragraaf is bedoeld voor de geïnteresseerde lezer. In de rest van dit rapport wordt slechts sporadisch gebruik gemaakt van de inhoud van deze paragraaf. In paragraaf 4.2 wordt een samenvatting van het correlatiemodel gegeven.

4.1 De conditionele verdeling van X gegeven Y

Om g(x|y) te behandelen is het handig de karakteristieke functie van het interval [0, ∞) te gebruiken, die wordt gegeven door

$$\chi(x) = \begin{cases} 1 & , x > 0 \\ 0 & , x \le 0 \end{cases} \qquad (4.1)$$

Voor g(x|y) kan dan worden geschreven, zie (3.12),

$$g(x \mid y) = \frac{g(x,y)}{g(y)} = \frac{e^{-x}\chi(x)\lambda(y-x-\delta)}{g(y)} \qquad , y > y_b + \delta \qquad (4.2)$$

Het correlatiemodel is zo opgesteld dat g(y|x) = λ(y-x-δ) op een verschuiving over x + δ na gelijk is aan de kansdichtheid λ(t). Formule (4.2) suggereert dat er in zijn algemeenheid geen kansdichtheid bestaat waaraan de g(x|y) op een verschuiving na gelijk is. *Voor grote y* blijkt dat echter wel *in benadering* het geval te zijn; de betreffende kansdichtheid zal hieronder worden aangeduid als γ(t). In het geval $y_e < \infty$ blijkt dat de verdelingen g(x|y) voor grote waarden van y zelfs *exact* gelijk worden, op een verschuiving na, aan γ(t). We zullen nu echter een iets strengere eis stellen aan de snelheid waarmee de staart van λ(t) afneemt. In plaats van de eerdere aanname uit (3.9) dat $E(e^T) < \infty$ eisen we nu[5], naast de voorwaarden (3.5) t/m (3.8),

Voorwaarde aan λ(t)

$$\text{(ii)} \quad E(Te^T) = \int te^t \lambda(t)\,dt < \infty \qquad (4.3)$$

Voor een normale verdeling is aan (4.3) voldaan, evenals voor de exponentiële verdeling met standaarddeviatie s < 1. Voor praktische toepassingen zullen de eisen (3.9) en (4.3) ongeveer op hetzelfde neer komen. Dat de laatste eis wel degelijk strenger is blijkt door het beschouwen van een λ(t) die vanaf zekere t_0 gelijk is aan $\lambda(t) = e^{-t}/t^2$. Daarvoor is wel voldaan aan (3.9) maar niet aan (4.3).

De zojuist genoemde functie γ(t) wordt gedefinieerd als

$$\gamma(t) = e^{\delta_0 - t}\lambda(-t) \qquad (4.4)$$

Met behulp van (3.20) kan worden geverifieerd dat γ(t) inderdaad zoals hierboven gesteld een kansdichtheid vormt (want ≥ 0 en genormeerd op 1). Merk op dat een snel afnemende *linker*staart van γ(t) correspondeert met een (zeer) snel afnemende *rechter*staart van λ(t). Het gemiddelde van γ(t) wordt aangegeven met μ_γ en is gelijk aan

$$\mu_\gamma = \int \gamma(t)t\,dt = -e^{\delta_0}\int \lambda(t)te^t\,dt \qquad (4.5)$$

Het laatste lid hiervan maakt duidelijk waarom de eis (4.3) is gesteld. Indien daaraan niet voldaan is geldt $\mu_\gamma = \infty$, hetgeen in de context van dit stuk geen interessant geval vormt om te behandelen. In appendix A.4 wordt aangetoond dat onder de voorwaarde (4.3) altijd geldt

[5] Merk op dat $E(Te^T)$ nooit $-\infty$ kan worden omdat voor $t < 0$ de functie $f(t) = te^t$ begrensd is.

(4.6) $\mu_\gamma < 0$

De gezamenlijke kansdichtheid g(x,y) uit (3.12) kan in termen van λ(t) of γ(t) worden geschreven als

(4.7) $g(x,y) = e^{-x}\chi(x)\lambda(y-x-\delta) = e^{-y+\delta-\delta_0}\chi(x)\gamma(x-y+\delta)$

terwijl voor g(x|y) volgt

(4.8) $g(x \mid y) = \dfrac{e^{-y+\delta-\delta_0}}{g(y)}\chi(x)\gamma(x-y+\delta) \quad , y > y_b + \delta$

Situatie $y_e < \infty$

Beschouw nu $y_e < \infty$ en $y \geq y_e + \delta$. Omdat vanwege (3.22) nu geldt $g(y) = \exp(-y+\delta-\delta_0)$ volgt dan

(4.9) $g(x \mid y) = \gamma(x-y+\delta)\chi(x) \quad , y \geq y_e + \delta$

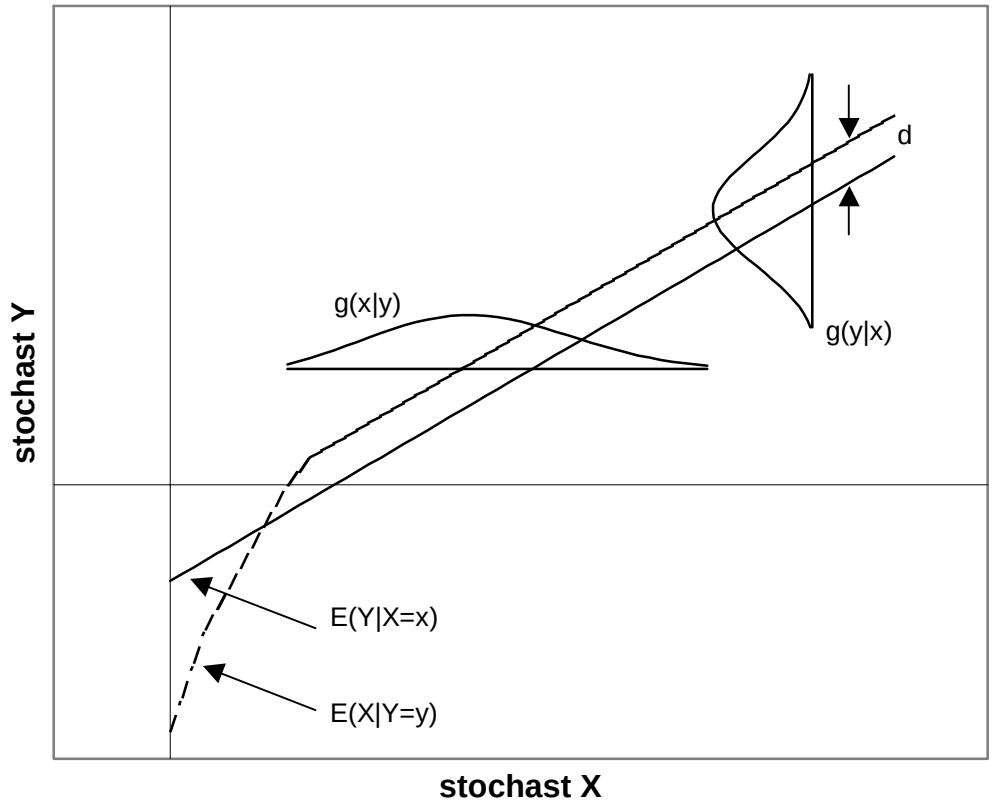
Voor de hier beschouwde waarden van y geldt indien $x \leq 0$ dat $x-y+\delta \leq x-y_e \leq -y_e$. Uit (4.4) volgt dan dat voor deze waarden van x en y moet gelden $\gamma(x-y+\delta) = 0$, zodat behalve (4.9) tevens geldt

(4.10) $g(x \mid y) = \gamma(x-y+\delta) \quad , y \geq y_e + \delta$

Hieruit blijkt dat voor $y \geq y_e + \delta$ de verdelingen g(x|y) op een verschuiving over $y - \delta$ na gelijk zijn aan de functie γ(x). Daaruit volgt dat voor deze waarden van y de gemiddelden van de verdelingen g(x|y) op een rechte lijn moeten liggen; voor lagere waarden van y zullen deze gemiddelden in het algemeen niet op een rechte lijn liggen. Geef de beschouwde gemiddelden aan met de functie x(y) = E(X|Y=y). Dan geldt, voor $y \geq y_e + \delta$,

$$\begin{aligned} x(y) &= E(X \mid Y = y) = \int g(x \mid y)x\,dx \\ &= \int \gamma(x-y+\delta)x\,dx = \int \gamma(t)t\,dt + y - \delta \\ &= \mu_\gamma + y - \delta \end{aligned} \tag{4.11}$$

Een logische vraag is hoe x(y) en de eerder, zie (3.13), behandelde y(x) = E(Y|X=x) = x + δ zich tot elkaar verhouden. Het is eenvoudig na te gaan dat voor de hier beschouwde grote waarden van y de lijnen y(x) en x(y) evenwijdig aan elkaar lopen, waarbij x(y) altijd de (verticaal gemeten) afstand $|\mu_\gamma|$ bóven de lijn y(x) ligt. Zie figuur 4.1 ter illustratie.



Figuur 4.1 Het verband tussen de lijnen y(x) = E(Y|X=x) = x + δ en x(y) = E(X|Y=y). Voor y ≥ y_e + δ loopt de lijn x(y) evenwijdig aan y(x) op een afstand d = $|\mu_\gamma|$ in welk geval x(y) altijd boven y(x) ligt.

Situatie $y_e = \infty$

Beschouw nu $y_e = \infty$. Het voorgaande blijkt voor grote y dan niet meer exact maar slechts in benadering te gelden. Zoals in (3.26) reeds werd gesteld geldt nu

(4.12) $g(y) \cong e^{-y+\delta-\delta_0}$, voor y groot

Uit (4.8) volgt dan als analogon van (4.9) dat

(4.13) $g(x \mid y) \cong \gamma(x-y+\delta)\,\chi(x)$, voor y groot

Bedenk nu dat voor x ≤ 0 en voor grote waarden van y men 'ver in de linkerstaart van γ(t) zit'. Voor deze waarden van x en y zal γ(x-y+δ) dus klein zijn. Het rechterlid van (4.13) zal dus niet veel veranderen wanneer de term χ(x) daaruit wordt weggelaten. Naast (4.13) geldt dus tevens

(4.14) $g(x \mid y) \cong \gamma(x-y+\delta)$, voor y groot

In feite is de laatste benadering veel praktischer om mee te werken dan de voorgaande, omdat het rechterlid van (4.13) geen kansdichtheid vormt terwijl dat wel het geval is voor het rechterlid van (4.14). De integratie van het rechterlid van (4.13) over alle x levert namelijk geen 1 op maar, zoals volgt door (4.8) te integreren over alle x,

(4.15) $\int \chi(x)\gamma(x-y+\delta)\,dx = \frac{g(y)}{e^{-y+\delta-\delta_0}} < 1$

waarbij de laatste ongelijkheid volgt uit (3.27) of uit de overweging dat $\gamma(t)$ een oneindig lange linkerstaart heeft. Het verband tussen de exacte formule (4.8) voor g(x|y) en de benadering (4.14) kan als volgt worden beschouwd. De benadering wordt gevormd door de kansdichtheid $f(x) = \gamma(x-y+\delta)$. Wanneer van deze kansdichtheid het gedeelte links van 0 wordt afgekapt en de resulterende functie opnieuw wordt genormeerd ontstaat de exacte uitdrukking voor g(x|y). De kansmassa die zich links van 0 bevindt in de betreffende kansdichtheid blijkt met behulp van (4.15) gelijk te zijn aan

$$(4.16) \qquad \int_{-\infty}^{0} \gamma(x-y+\delta)\,dx = 1 - \int_{0}^{\infty} \gamma(x-y+\delta)\,dx = 1 - \frac{g(y)}{e^{-y+\delta-\delta_0}}$$

Hieruit blijkt dat hoe beter de benadering (4.12) is, hoe beter ook de benadering (4.14) zal zijn, omdat bij een betere benadering de 'afgekapte' kansmassa kleiner zal zijn.

Als analogon van (4.11) geldt nu vanwege (4.14)

$$(4.17) \qquad x(y) \cong E(X \mid Y = y) = \mu_\gamma + y - \delta \qquad \text{, voor } y \text{ groot}$$

Voor grote y liggen de gemiddelden E(X|Y=y) dus in benadering op een rechte lijn, waarmee de in figuur 4.1 geschetste situatie dus voor $y_e = \infty$ in benadering juist blijft voor grote y. We merken wel op dat in een concrete toepassing beoordeeld dient te worden hoe goed de benadering is en of deze bruikbaar is; zie ook de eerdere opmerkingen volgend op (3.25).

Voorbeeld: de normale verdeling

Vanwege het grote belang van de normale verdeling in toepassingen wordt het bovenstaande toegelicht voor de uit de literatuur reeds bekende situatie waarin $\lambda(t)$ de normale verdeling vormt, zie bijvoorbeeld [Vrouwenvelder et al, 1999]. In dat geval geldt volgens (3.21) $\delta_0 = -s^2/2$. Enig schrijfwerk leert dat $\gamma(t)$ dan gegeven wordt door

$$(4.18) \qquad \gamma(t) = \frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{(t+s^2)^2}{2s^2} \right)$$

Dus $\gamma(t)$ volgt net als $\lambda(t)$ weer een normale verdeling met standaarddeviatie s, echter nu niet met gemiddelde 0 maar met gemiddelde $\mu_\gamma = -s^2$. De kansdichtheden $\lambda(t)$ en $\gamma(t)$ behoren in dit geval dus tot hetzelfde type kansverdeling. In het algemeen zal $\gamma(t)$ niet tot hetzelfde type kansverdeling als $\lambda(t)$ behoren. Volgens (4.14) geldt nu

$$(4.19) \qquad g(x \mid y) \cong \frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{\left[x - (y - \delta - s^2) \right]^2}{2s^2} \right) \qquad \text{, voor } y \text{ groot}$$

Dus g(x|y) volgt voor grote y in benadering een normale verdeling met standaarddeviatie s en gemiddelde $y-\delta-s^2$. In toepassingen, zie bijvoorbeeld [Vrouwenvelder et al, 1999] wordt vaak de keuze $\delta = \delta_0$ gemaakt, omdat g(y) dan asymptotisch standaard exponentieel verdeeld is. Dan is $y-\delta-s^2 = y - s^2/2$. De lijn met de gemiddelden E(Y|X=x) ligt dan op een (verticaal gemeten) afstand $s^2/2$ onder de lijn y = x. De lijn met de gemiddelden E(X|Y=y) vormt voor grote y dan in benadering een rechte lijn die op een (verticaal gemeten) afstand $s^2/2$ boven de lijn y = x ligt.

4.2 Samenvatting van het correlatiemodel

De belangrijkste punten betreffende het correlatiemodel uit hoofdstuk 3 en paragraaf 4.1 worden nu samengevat. De kansdichtheid $\lambda(t)$ heeft gemiddelde 0 en willekeurige standaarddeviatie $s > 0$.

Voor $y_e < \infty$ (naar boven begrensde drager van $\lambda(t)$) geldt:

1. De marginale verdeling van X wordt, voor $x > 0$, gegeven door $g(x) = e^{-x}$ en de conditionele verdeling van Y gegeven X door $g(y|x) = \lambda(y-x-\delta)$.

2. De marginale verdeling van Y wordt, voor alle $y \geq y_e + \delta$, gegeven door $g(y) = \exp(-y+\delta-\delta_0)$ en de conditionele verdeling van X gegeven Y door $g(x|y) = \gamma(x-y+\delta)$. Hierin is $\gamma(t) = \exp(\delta_0-t)\lambda(-t)$. De kansdichtheid $\gamma(t)$ heeft gemiddelde $\mu_\gamma < 0$.
   Voor $y < y_e + \delta$ volgt $g(y)$ uit (3.16) en volgt $g(x|y)$ uit (4.2) of (4.8).

3. De conditionele gemiddelden $E(Y|X=x)$ liggen op de rechte lijn met vergelijking $y = x + \delta$, voor $x > 0$.

4. De conditionele gemiddelden $E(X|Y=y)$ liggen, voor alle $y \geq y_e + \delta$, op de rechte lijn met vergelijking $y = x + \delta - \mu_\gamma$. Omdat $\mu_\gamma < 0$ ligt deze lijn bóven die in punt (3).
   Voor $y < y_e + \delta$ liggen de gemiddelden in het algemeen niet op een rechte lijn.

Voor $y_e = \infty$ (oneindig lange staart) gelden de beweringen in punt (2) en (4) slechts in benadering voor grote y; de frase 'voor alle $y \geq y_e + \delta_0$' moet dan dus worden gelezen als 'geldt in benadering voor grote y'. Voor punt (2) dient de staart van $\lambda(t)$ zo snel af te nemen dat is voldaan aan $E(e^T) < \infty$ en voor punt (4) dat is voldaan aan $E(Te^T) < \infty$. De kwaliteit van de benaderingen hangt af van de snelheid waarmee de staart van de kansdichtheid $\lambda(t)$ naar nul gaat. Hoe sneller dat gebeurt, hoe beter de benaderingen zullen zijn. De kwaliteit van de benaderingen hangt dus samen met het type kansverdeling dat beschouwd wordt. Daarnaast zal een grotere waarde van de standaarddeviatie s van $\lambda(t)$ een langzamere afname van de staart van de verdeling leveren, waardoor minder goede benaderingen resulteren dan voor een kleinere waarde van s. In toepassingen dient beoordeeld te worden in welke mate de genoemde benaderingen bruikbaar zijn.

In toepassingen wordt vaak de keuze $\delta = \delta_0 = -s^2/2$ gemaakt, omdat $g(y)$ dan asymptotisch standaard exponentieel verdeeld is. In dat geval volgt $\gamma(t)$ eveneens een normale verdeling met standaarddeviatie s en gemiddelde $\mu_\gamma = -s^2$. De lijn met de gemiddelden $E(Y|X=x)$ ligt dan op een (verticaal gemeten) afstand $s^2/2$ onder de lijn $y = x$. De lijn met de gemiddelden $E(X|Y=y)$ vormt voor grote y dan in benadering een rechte lijn die op een (verticaal gemeten) afstand $s^2/2$ boven de lijn $y = x$ ligt.

5 Toepassing correlatiemodel met λ(t) de uniforme verdeling

De formules uit de voorgaande hoofdstukken worden hieronder toegepast voor de situatie dat λ(t) de uniforme verdeling vormt. De conditionele verdelingen g(y|x) zijn dan dus op een verschuiving na gelijk aan de uniforme verdeling. In deze situatie kunnen diverse grootheden analytisch worden bepaald. De kansdichtheid λ(t) is van de vorm, voor a > 0,

$$\lambda(t) = \begin{cases} \frac{1}{2a} & , -a < t < a \\ 0 & , \text{anders} \end{cases} \tag{5.1}$$

De drager wordt dan gegeven door $(y_b, y_e) = (-a,a)$. De standaarddeviatie wordt gegeven door

$$s = \frac{a}{\sqrt{3}} \tag{5.2}$$

In appendix A.5 wordt uitgerekend dat de grootheden $E(e^T)$, δ_0 en μ_γ uit (3.9), (3.20) en (4.6) worden gegeven door

$$\begin{aligned} E(e^T) &= \frac{e^a - e^{-a}}{2a} \\ \delta_0 &= -\ln\left(\frac{e^a - e^{-a}}{2a}\right) \\ \mu_\gamma &= \frac{e^{\delta_0}}{2a}\left(e^a(1-a) - e^{-a}(1+a)\right) \end{aligned} \tag{5.3}$$

Omdat λ(t) een even functie is in t volgt voor γ(t) uit (4.4)

$$\gamma(t) = e^{\delta_0 - t}\lambda(t) \tag{5.4}$$

Figuur 5.1 geeft een illustratie van λ(t) en γ(t) voor het geval a = 2. Ter illustratie melden we dat in dat geval

$$\begin{aligned} E(e^T) &= 1.813 \\ \delta_0 &= -0.595 \\ \mu_\gamma &= -1.075 \end{aligned} \tag{5.5}$$

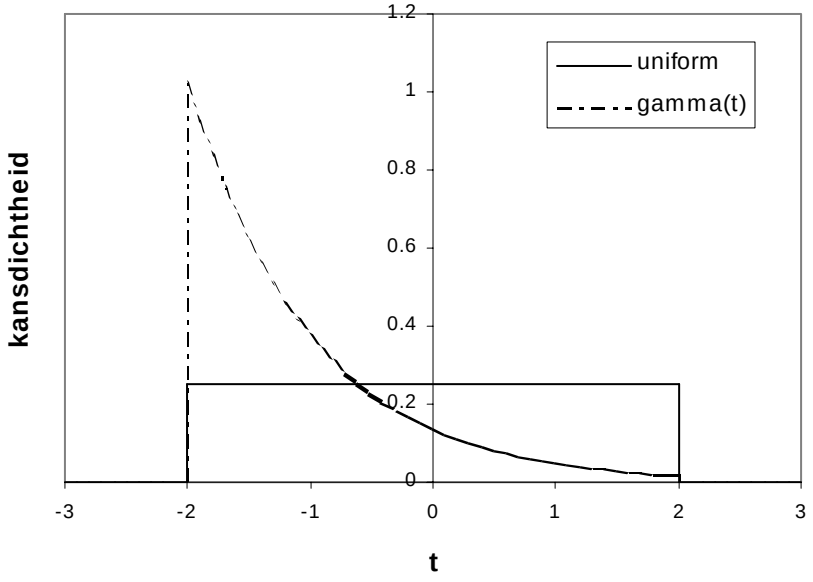
In de appendix wordt ook de formule voor de overschrijdingskans van Y afgeleid. Deze wordt gegeven door

$$\overline{F}_Y(y) = \begin{cases} \frac{e^a - e^{-a}}{2a} e^{-y+\delta} & , y \geq a + \delta \\ \frac{1}{2} + \frac{1}{2a}\left(1 - y + \delta - e^{-y+\delta-a}\right) & , -a + \delta < y < a + \delta \\ 1 & , y \leq -a + \delta \end{cases} \tag{5.6}$$

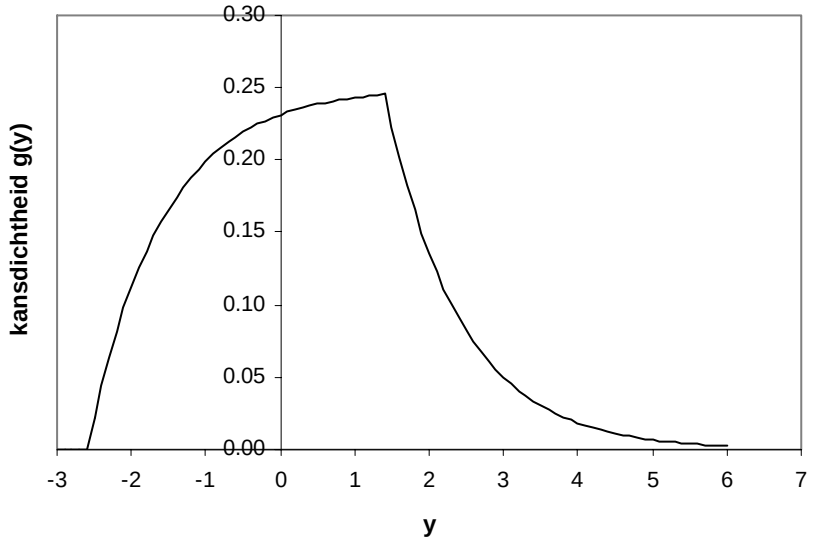
Door differentiatie volgt hieruit de kansdichtheid van Y, die gegeven wordt door

$$(5.7) \qquad g(y) = \begin{cases} \dfrac{e^{a} - e^{-a}}{2a} e^{-y+\delta} & , y \geq a + \delta \\ \dfrac{1}{2a}\left(1 - e^{-y+\delta-a}\right) & , -a + \delta < y < a + \delta \\ 0 & , y \leq -a + \delta \end{cases}$$

Deze kansdichtheid is in figuur 5.2 weergegeven voor het geval a = 2 en $\delta = \delta_0 = -0.595$. Voor deze keuze van δ wordt g(y) voor $y \geq a + \delta_0 = 1.405$ gelijk aan de standaardexponentiële verdeling. Deze overgang is duidelijk te zien in figuur 5.2.



Figuur 5.1 De kansdichtheden $\lambda(t)$ en $\gamma(t)$ voor het geval a = 2.



Figuur 5.2 De kansdichtheid g(y) voor het geval a = 2 en $\delta = \delta_0 = -0.595$.

6 Correlatiemodel in de originele ruimte

6.1 De kansdichtheid in de orginele ruimte

Beschouw twee stochasten Q en M. Deze zouden in een concrete toepassing bijvoorbeeld respectievelijk de IJsselafvoer en het IJsselmeerpeil kunnen voorstellen. We gaan uit van gegeven verdelingsfuncties $F_Q(q)$ en $F_M(m)$. Daarmee liggen de kansdichtheden $g_Q(q)$ en $g_M(m)$, of meestentijds kortweg g(q) en g(m), volledig vast. Het vlak van punten (q, m) zal worden aangeduid als de 'originele ruimte', met als motivatie dat de $F_Q(q)$ en $F_M(m)$ vooraf gegeven zijn en dus kunnen worden opgevat als de originele gegevens. In het voorgaande hoofdstuk werden $g_X(x)$ en $g_Y(y)$ beschouwd, welke meestentijds kortweg werden aangeduid als g(x) en g(y). In het vervolg zullen transformaties worden beschouwd waarmee q overgaat in x en m overgaat in y. Daarbij gaat g(q) over in g(x) en g(m) in g(y). De transformaties worden ieder apart voor q en m uitgevoerd (de eventuele correlatie tussen q en m speelt in het toepassen van de transformaties dus geen rol) en luiden

$$\begin{aligned} x &= J(q) \\ y &= K(m) \end{aligned} \tag{6.1}$$

met J(q) en K(m) vastgelegd door

$$\begin{aligned} F_X(J(q)) &= F_Q(q) \\ F_Y(K(m)) &= F_M(m) \end{aligned} \tag{6.2}$$

Om te garanderen dat deze transformaties eenduidig gedefinieerd zijn zullen we enige eisen stellen aan g(q) en g(m). Ten eerste zullen we aannemen dat de dragers D_Q en D_M van de stochasten Q en M open intervallen vormen, waarbij we in herinnering brengen dat de drager van een stochast volgens paragraaf 3.1 bestaat uit die waarden waarvoor de kansdichtheid van de betreffende stochast positief is. De aanname is dus dat de dragers zijn van de vorm

$$\begin{aligned} D_Q &= (q_b, q_e) \\ D_M &= (m_b, m_e) \end{aligned} \tag{6.3}$$

Deze intervallen mogen naar boven zowel als beneden begrensd dan wel onbegrensd zijn. Omdat buiten deze dragers de kansdichtheden nul zijn hoeven de transformaties (6.1) slechts beschouwd te worden voor $q \in D_Q$ en $m \in D_M$. De kern van de aanname dat de dragers een interval vormen is dat de waarden waarvoor de kansdichtheid positief is in een *aaneengesloten* gebied liggen. In toepassingen zal hier veelal aan voldaan zijn. De reden dat de dragers een *open* interval vormen heeft het wiskundige voordeel dat de transformatie op de hele drager eenduidig is gedefinieerd. Indien bijvoorbeeld, voor een situatie waarin D_Q begrensd is, $D_Q = (q_b, q_e]$ zou worden genomen zou de transformatie niet gedefinieerd zijn voor $q = q_e$. Een tweede eis die wordt gesteld aan de stochasten Q en M is

$$\begin{aligned} &g(q) \text{ is continu op } D_Q = (q_b, q_e) \\ &g(m) \text{ is continu op } D_M = (m_b, m_e) \end{aligned} \tag{6.4}$$

Merk op dat deze voorwaarden niet uitsluiten dat bijvoorbeeld g(q) in een randpunt q_b of q_e wel een discontinuïteit heeft. De g(q) zou bijvoorbeeld een exponentiële verdeling kunnen volgen op het interval (q_b, ∞) waarbij de kansdichtheid links van q_b gelijk aan nul is. Onder de voorwaarden (6.3) en (6.4) zijn de verdelingsfuncties $F_Q(q)$ en $F_M(m)$ strikt stijgend en differentieerbaar op hun dragers. Voor meer details betreffende de wiskundige eigenschappen van de transformaties wordt de lezer verwezen naar paragraaf 6.1 van de Appendix.

Onder de voorwaarden (6.3) en (6.4) geldt de theorie zoals uiteengezet in het vervolg van dit hoofdstuk en in de volgende hoofdstukken. Wel is het goed op te merken dat in praktische toepassingen waarin niet aan deze voorwaarden is voldaan onderstaande formules in essentie geldig zullen blijven. Wel zullen dan wat aanpassingen moeten worden verricht, die afhankelijk zijn van de beschouwde toepassing. Er is hier dus gekozen

voor een wiskundig zeer exacte formulering waarbij het aan de lezer wordt overgelaten in een praktische situatie indien nodig de formules iets aan te passen.

Omdat X een standaardexponentiële verdeling volgt is de transformatie van q naar x zeer eenvoudig uit te voeren. Er geldt vanwege (3.1)

$$J(q) = -\ln\left(\overline{F}_Q(q)\right) \tag{6.5}$$

De functie K(m) dient numeriek bepaald te worden, door gebruik te maken van de gegeven $F_M(m)$ en formule (3.14). Merk op dat K(m) afhangt van de in hoofdstuk 3 beschouwde $\lambda(t)$ en δ.

Het in hoofdstuk 3 beschouwde correlatiemodel luidde, voor x > 0,

$$\begin{aligned} g(x) &= e^{-x} \\ g(y \mid x) &= \lambda(y - x - \delta) \end{aligned} \tag{6.6}$$

Dit vormt het correlatiemodel in de getransformeerde ruimte. We zullen in paragraaf A.6.2 van de appendix laten zien dat in de originele ruimte het correlatiemodel de volgende vorm heeft

$$\begin{aligned} g(q) &= F_Q'(q) = e^{-J(q)} J'(q) \\ g(m \mid q) &= \lambda\left(K(m) - J(q) - \delta\right) K'(m) \end{aligned} \tag{6.7}$$

Hierin slaat het accent op het nemen van de afgeleide, dus bijvoorbeeld J’(q) = dJ(q)/dq. De gezamenlijke kansdichtheid wordt in termen van $\lambda(t)$ of $\gamma(t)$ gegeven door

$$\begin{aligned} g(q,m) &= \lambda\left(K(m) - J(q) - \delta\right) e^{-J(q)} J'(q) K'(m) \\ &= \gamma\left(J(q) - K(m) + \delta\right) e^{-K(m)+\delta-\delta_0} J'(q) K'(m) \end{aligned} \tag{6.8}$$

Deze gezamenlijke kansdichtheid blijkt (zoals gewenst) de g(q) en g(m) als marginale verdelingen te hebben. Formules (6.7) en (6.8) worden in de appendix bewezen. Verder geldt

$$P(M < m \mid q) = \Lambda\left(K(m) - J(q) - \delta\right) \tag{6.9}$$

Dat kan worden ingezien als volgt. Met behulp van (6.7) volgt dat zowel het linker- als rechterlid van (6.9) g(m|q) als afgeleide hebben. Dan volgt dat het rechterlid van (6.9) op een constante na gelijk moet zijn aan P(M<m|q). Deze constante moet echter nul zijn, omdat beide leden van (6.9) naar 1 gaan in de limiet $m \to \infty$.

De formules voor het correlatiemodel in de originele ruimte kunnen dus betrekkelijk eenvoudig worden berekend uit de kansdichtheden $\lambda(t)$ en $\gamma(t)$ en de transformaties J(q) en K(m). We brengen in herinnering dat in de getransformeerde ruimte de gemiddelden van de kansdichtheden g(y|x) liggen op een rechte lijn met vergelijking, zie (3.13), y(x) = E(Y|X=x) = x + δ. In de originele ruimte liggen de gemiddelden m(q) = E(M|Q=q) in het algemeen niet meer op een rechte lijn. Overigens volgt uit (6.7) eenvoudig dat dat wel het geval is indien K(m) en J(q) beide lineair zijn. Die laatste situatie zal zich voor K(m) in toepassingen (behalve misschien in benadering) in de regel niet voordoen. J(q) zal lineair zijn indien g(q) een exponentiële verdeling volgt, zoals onmiddelijk blijkt uit (6.5). De gemiddelden E(Q|M=m) liggen in het algemeen eveneens niet op een rechte lijn.

6.2 De rol van δ

In de kansdichtheden g(x,y) en g(q,m) komt de parameter δ voor. Uit (3.12) en eveneens uit figuur 3.1 blijkt dat een andere keuze van δ wat g(x,y) betreft neerkomt op een verschuiving van de kansdichtheid langs de Y-as. We zullen nu aantonen dat hoewel δ voorkomt in formule (6.8) voor g(q,m) een andere keuze van δ toch dezelfde g(q,m) oplevert. De reden blijkt te zijn dat de functie K(m) die eveneens in formule (6.8) voorkomt op een

bepaalde manier van δ afhangt. Het blijkt dat een verandering in δ als het ware 'geneutraliseerd wordt' door de manier waarop de K(m) verandert bij deze verandering in δ.

Beschouw om het voorgaande aan te tonen twee verschillende δ_1 en δ_2 en geef de bijbehorende functies K(m) aan met $K(m,\delta_1)$ en $K(m,\delta_2)$. Uit (3.14) en (6.2) volgt dan

$$\begin{aligned} F_M(m) &= \int_0^\infty e^{-x}\Lambda\left(K(m,\delta_1)-x-\delta_1\right)dx \\ &= \int_0^\infty e^{-x}\Lambda\left(K(m,\delta_2)-x-\delta_2\right)dx \end{aligned} \tag{6.10}$$

Omdat K(m) door (6.2) uniek wordt vastgelegd, moet dan gelden

$$K(m,\delta_1)-\delta_1 = K(m,\delta_2)-\delta_2 \tag{6.11}$$

Uit (6.8) volgt dan dat g(q,m) hetzelfde is voor δ_1 en δ_2 zodat g(q,m) niet afhangt van de beschouwde δ. Met het oog op toepassingen is het dan handig δ gelijk aan δ_0 uit (3.20) te kiezen. Zoals in hoofdstuk 3 werd behandeld volgt g(y) dan asymptotisch de standaardexponentiële verdeling.

Het correlatiemodel uit hoofdstuk 3 is geformuleerd voor een kansdichtheid λ(t) met een bepaalde standaarddeviatie s. In toepassingen zal λ(t) echter dienen te worden bepaald uit de beschikbare data. Veelal zal men voor λ(t) dan één type kansverdeling beschouwen, bijvoorbeeld de normale verdeling of de uniforme verdeling. Naast eventuele andere parameters van λ(t) zal de standaarddeviatie van de verdeling bepaald moeten worden die het beste de data beschrijft. In een toepassing zal men λ(t) dus voor verschillende waarden van s beschouwen; zie het volgende hoofdstuk voor een voorbeeld waarin λ(t) de uniforme verdeling volgt. In het vervolg van deze paragraaf en in hoofdstuk 7 zullen we kansdichtheden $\lambda_s(t)$ met gemiddelde 0 beschouwen die zijn van de vorm

$$\lambda_s(t) = \frac{1}{s}\lambda_1\left(\frac{t}{s}\right) \quad ,s>0 \tag{6.12}$$

waarbij de standaarddeviatie van $\lambda_1(t)$ gelijk is aan 1. De kansdichtheden die hier worden beschouwd bevatten dus als enige parameter de standaarddeviatie s van de verdeling. Om gebruik te kunnen maken van de theorie uit hoofdstuk 3 moeten we aannemen dat de $\lambda_s(t)$ voor alle s voldoen aan (3.5) t/m (3.9). Het is eenvoudig na te gaan dat indien $\lambda_1(t)$ voldoet aan (3.5) t/m (3.8) hetzelfde geldt voor iedere $\lambda_s(t)$ met s > 0. Voorwaarde (3.9) zegt dat de verwachtingswaarde van de functie $f(t) = e^t$ met betrekking tot $\lambda_s(t)$ eindig moet zijn. Wanneer deze verwachtingswaarde wordt aangegeven met $E_s(e^T)$ volgt eenvoudig dat deze in termen van $\lambda_1(t)$ kan worden berekend als de integraal over de functie $e^{st}\lambda_1(t)$. De $\lambda_s(t)$ uit (6.12) voldoen dus aan de gestelde voorwaarden indien door $\lambda_1(t)$ is voldaan, naast de voorwaarden (3.5) t/m (3.8), aan:

Voorwaarde aan $\lambda_s(t)$

$$E_s(e^T) = \int e^t \lambda_s(t)\,dt = \int e^{st}\lambda_1(t)\,dt < \infty \quad ,s>0 \tag{6.13}$$

Indien $\lambda_1(t)$ de standaardnormale verdeling volgt is aan (6.13) voldaan. De voor toepassingen zeer belangrijke normale verdeling voldoet dus aan de gestelde voorwaarden. Daarnaast mag $\lambda_1(t)$ iedere verdeling zijn die een eindige rechterstaart ($y_e < \infty$) heeft, mits er voor wordt gezorgd dat de verdeling gemiddelde nul heeft. Voor $\lambda_1(t)$ mag bijvoorbeeld de uniforme verdeling met standaarddeviatie 1 worden genomen (gelijk aan 1 op $(-\sqrt{3},\sqrt{3})$ en 0 daarbuiten).

We merken nog op dat $E_s(e^T) < \infty$ voor alle s > 0 impliceert dat $E_s(Te^T) < \infty$ voor alle s > 0. Als s'> s geldt namelijk dat voor t voldoende groot $te^{st} < \exp(s't)$. Wanneer het rechterlid een eindige verwachtingswaarde heeft moet dat tevens gelden voor het linkerlid[6]. Voorwaarde (6.13) impliceert dus dat ook is voldaan aan

[6] Merk op dat op dezelfde manier volgt dat (6.13) impliceert dat $E_s(T^n e^T) < \infty$ voor alle n = 0, 1, 2,… en alle s>0.

voorwaarde (4.3) voor alle $\lambda_s(t)$. Samenvattend: de theorie van hoofdstuk 3 en 4 is van toepassing op de hier beschouwde $\lambda_s(t)$ indien door $\lambda_1(t)$ is voldaan aan (3.5) t/m (3.8) en indien daarnaast is voldaan aan (6.13).

Omdat de beschouwde kansdichtheden alleen van de parameter s afhangen volgt dat δ_0 uit (3.20) een functie is van s. Om deze afhankelijkheid expliciet te maken zullen we vanaf nu $\delta(s)$ schrijven in plaats van δ_0. De index 0 is in het vervolg overbodig en wordt daarom weggelaten. Volgens (3.20) en (6.13) volgt dan

(6.14) $$\delta(s) = -\ln\left(\int e^{st}\lambda_1(t)\,dt\right) < 0 \qquad ,s > 0$$

In appendix A.6 wordt aangetoond dat $\delta(s)$ en zijn afgeleide $\delta'(s)$ voldoen aan

(6.15) $$\lim_{s\downarrow 0}\delta(s) = \lim_{s\downarrow 0}\frac{\delta(s)}{s} = \lim_{s\downarrow 0}\delta'(s) = 0$$

(6.16) $$\delta'(s) < 0 \quad ,s > 0$$

Blijkbaar heeft volgens (6.16) een grotere s in absolute zin altijd een grotere $\delta(s)$ tot gevolg waarbij volgens (6.15) $\delta(s)$ tot nul nadert als s tot nul nadert. De afgeleide van $\delta(s)$ wordt in de buurt van nul gelijk aan nul, wat meetkundig inhoudt dat de raaklijn aan de grafiek daar horizontaal gaat lopen. Anders gezegd, voor kleine s (sterke correlatie) is $\delta(s)$ een orde kleiner dan s zelf. Omdat de afgeleide van $\delta(s)$ negatief is, volgt dat $\delta(s)$ een inverse heeft. Iedere $\delta < 0$ correspondeert dus met een unieke waarde van s. Wellicht is nog interessant dat (6.15) ons in staat stelt de functie $\delta(s)$ als continue functie op het domein $[0, \infty)$ te beschouwen in plaats van op $(0, \infty)$, waarbij dan $\delta(0) = 0$ terwijl voor de rechterafgeleide dan volgt $\delta'(0+) = 0$.

6.3 De conditionele verdeling van Q gegeven M

Voor de liefhebber gaan we nu in op formules voor g(m) en g(q|m) die voor grote m exact (ingeval $y_e < \infty$) of in benadering (ingeval $y_e = \infty$) gelden. Beschouw eerst $y_e < \infty$. Definieer m_e als de unieke[7] waarde m waarvoor voldaan is aan

(6.17) $$K(m_e) = y_e + \delta$$

Beschouw nu $m \geq m_e$. Omdat K(m) stijgend is volgt dan $K(m) \geq K(m_e) = y_e + \delta(s)$. Uit (6.2) en (3.22) volgt dan

(6.18) $$\overline{F}_M(m) = e^{-K(m)}, \qquad m \geq m_e$$

waaruit door differentiatie volgt

(6.19) $$g(m) = K'(m)e^{-K(m)}, \qquad m \geq m_e$$

Voor g(q|m) volgt dan uit (6.8)

(6.20) $$g(q \mid m) = \gamma\left(J(q) - K(m) + \delta(s)\right)J'(q), \qquad m \geq m_e$$

Merk nog op dat K(m) voor de hier beschouwde waarden van m volgens (6.18) kan worden berekend uit $F_M(m)$ volgens

(6.21) $$K(m) = -\ln\left(\overline{F}_M(m)\right), \qquad m \geq m_e$$

Deze relatie komt overeen met (6.5), zij het dat de huidige relatie slechts geldt voor grote waarden van m.

[7] Merk op dat de aanname dat K(m) strikt stijgend is impliceert dat m_e uniek bepaald is.

Beschouw nu $y_e = \infty$. De relaties (6.18) tot en met (6.21) gelden dan niet meer exact maar in benadering voor grote m. We zullen dat voor (6.18) en (6.20) toelichten. Merk daartoe eerst op dat grote waarden van y in de transformatie y = K(m) uit (6.1) samengaan met grote waarden van m en omgekeerd, hetgeen volgt omdat K(m) strikt stijgend is in m. Op grond van (6.2), (6.8) en (3.25) volgt dan als analogon van (6.18) en (6.20)

(6.22) $\overline{F}_M(m) \cong e^{-K(m)}$, voor m groot

(6.23) $g(q \mid m) \cong \gamma\left(J(q) - K(m) + \delta(s)\right) J'(q)$, voor m groot

De kwaliteit van de benaderingen valt niet in zijn algemeenheid aan te geven. In een concrete toepassing dient beoordeeld te worden in welke mate de genoemde benaderingen bruikbaar zijn.

7 Het bepalen van de bivariate kansdichtheid uit de data

Dit hoofdstuk gaat net als het voorgaande uit van stochasten Q en M met bekende kansdichtheden $g_Q(q)$ en $g_M(m)$. Hoewel strikt genomen niet helemaal correct zullen deze kansdichtheden indien geen verwarring kan ontstaan veelal korter worden geschreven als g(q) en g(m). Verder wordt aangenomen dat de data bestaan uit N paren (q_i, m_i), i = 1, 2, ..., N. Het doel van dit hoofdstuk is een gezamenlijke kansdichtheid g(q,m) af te leiden die de correlatie in de data adequaat beschrijft en die g(q) en g(m) als marginale verdelingen heeft. De manier om g(q,m) af te leiden bestaat eruit eerst Q en M te transformeren naar stochasten X en Y op de manier die reeds globaal werd beschreven in hoofdstuk 2 en op gedetailleerde wijze in hoofdstuk 6. In de getransformeerde ruimte is de conditionele kansdichtheid g(y|x) van Y gegeven X = x dan op een verschuiving langs een rechte lijn na gelijk aan een kansdichtheid $\lambda(t)$. Paragraaf 7.2 en 7.3 hebben enige overlap met de uiteenzetting in de hoofdstukken 3 t/m 6. Dat is gedaan als service aan de lezer – die kan in principe die hoofdstukken dan overslaan. Voor meer details en achtergronden zijn met name de hoofdstukken 3 en 5 t/m 6 echter wel nodig. De volgende paragrafen veronderstellen dat hoofdstuk 2 bekend is bij de lezer.

Paragraaf 7.1 gaat in op de vraag wanneer het correlatiemodel wel of niet mag worden toegepast. Paragraaf 7.2 geeft een recept om met een iteratieve methode de parameter s te bepalen (zie paragraaf 2.2); deze stelt de standaarddeviatie van $\lambda(t)$ voor in de getransformeerde ruimte. Dat is de enige parameter die een rol speelt in het model, zij het dat een andere 'vrijheidsgraad' nog de keuze van het type verdeling voor $\lambda(t)$ is. In paragraaf 7.3 wordt het recept toegepast op een voorbeeld, met de uniforme verdeling als keuze voor het type verdeling. Ook wordt een bepaald soort figuren toegelicht, die als 'diagnostische tool' dienen in het bepalen van de juiste keuze voor s. Paragraaf 7.4 beschrijft de Maximum Likelyhood methode als alternatief voor het iteratieve recept. Paragraaf 7.5 geeft een discussie van de methoden en adviezen voor concrete toepassingen.

7.1 Wanneer is het correlatiemodel toepasbaar?

Deze paragraaf gaat over de vraag wanneer het correlatie model wel of niet toepasbaar is, en daarnaast over keuzes in de analyse. We noemen vier punten, die hieronder nader worden toegelicht:

1. Niet iedere g(q,m), en dus niet iedere vorm van correlatie, kan door het correlatiemodel worden beschreven.

2. Data met een *negatieve* correlatie kunnen zeker niet door het correlatiemodel worden beschreven.

3. Alle relevante waarnemingen dienen bij conditionering op een waarde q tot de geselecteerde data te behoren.

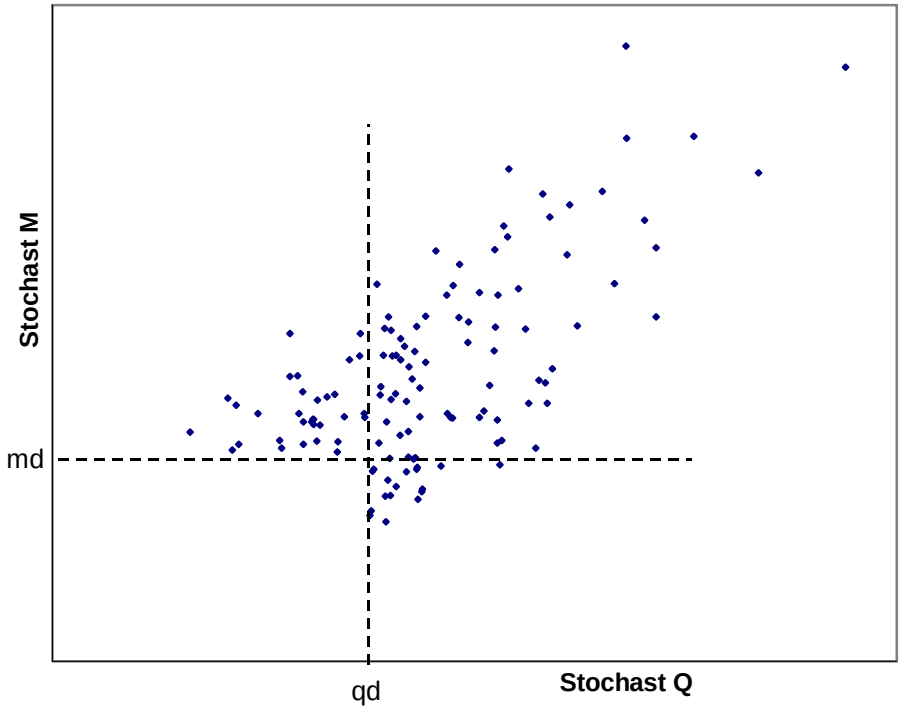
4. Voor de gecorreleerde stochasten Q en M dient een keus te worden gemaakt: welke van de twee variabelen wordt naar de standaardexponentiële variabele X getransformeerd en welke naar Y.

Als uitleg over punt 1: er zijn 'oneindig veel' kansdichtheden g(q,m) met dezelfde marginale kansdichtheden. Anders gezegd, er zijn oneindig veel formules voor g(q,m) te geven die alle dezelfde marginale kansdichtheden hebben (dat blijkt onder meer uit paragraaf 9.1). In zekere zin kan slechts 'een klein deel' van deze kansdichtheden g(q,m) exact met het beschouwde correlatiemodel worden beschreven: de meeste van deze kansdichtheden kunnen voor geen enkele keuze van s en van het type verdeling voor $\lambda(t)$ worden verkregen. (Of iets explicieter: formule (7.12) kan niet alle g(q,m) opleveren die g(q) en g(m) als marginale verdelingen hebben.) Eigenlijk is de vraag niet zozeer of de werkelijke correlatie exact door het model kan worden beschreven. De 'werkelijke correlatie' is ook nooit bekend (behalve eventueel in simulaties). De vraag is meer of het correlatiemodel een voldoend nauwkeurige beschrijving geeft van de uit de data blijkende correlatie. Wat hier voldoende nauwkeurig is, is dan een kwestie van oordeelkundig inzicht. Paragraaf 7.5 geeft een aantal criteria om de kwaliteit van de fit aan de data te beoordelen, maar het uiteindelijke oordeel zal altijd een subjectief element bevatten.

Wat punt 2 betreft: een negatieve correlatie kan nooit goed door het model worden weergegeven. Zonder daar uitgebreid op in te gaan melden we als reden dat de helling van de lijn $y = x + \delta$ in figuur 2.3 *positief* is (namelijk gelijk aan 1). Voor het beschouwen van een negatieve correlatie zouden de formules allemaal herschreven moeten worden voor een situatie met een negatieve helling, namelijk $y = - x + \delta$. Veel simpeler is het om bij een negatieve correlatie in de data één van de stochasten Q of M te voorzien van een minteken. Neem bijvoorbeeld

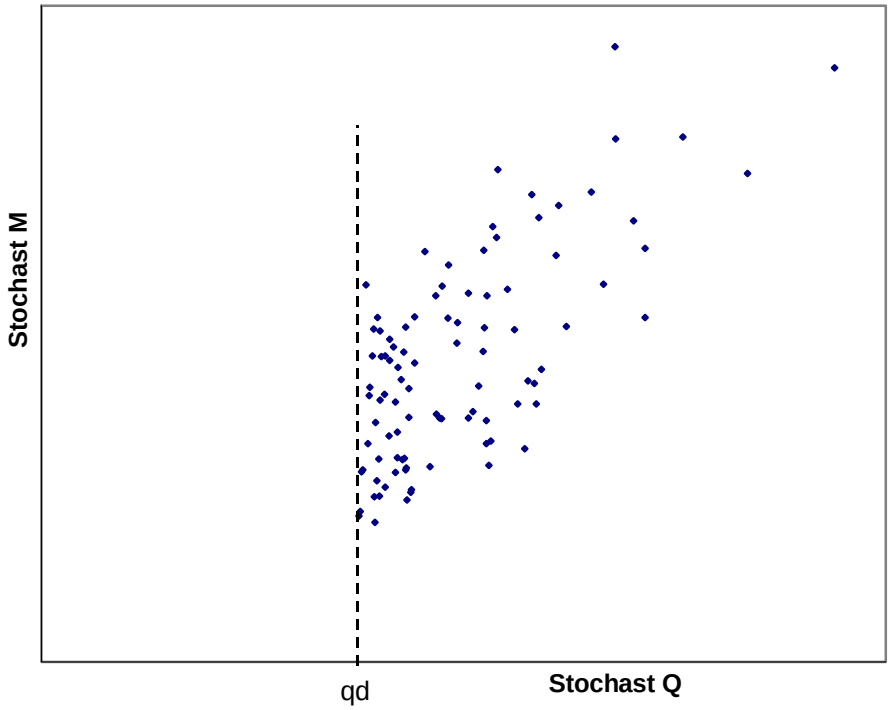
M' = - M, met uiteraard m_i' = - m_i voor i = 1 t/m N: dan vertonen de data behorend bij M' en Q een positieve correlatie. Vervolgens kan dan voor deze stochasten worden onderzocht of het model de correlatie voldoende nauwkeurig beschrijft. Als dat het geval is, met $g_{Q,M'}(q,m')$ als resultaat, volgt de gewenste kansdichtheid in termen van de oorspronkelijke stochasten simpelweg als $g_{Q,M}(q,m) = g_{Q,M'}(q, m'= -m)$.

Punt 3 kan het best met een voorbeeld worden uitgelegd, zie daartoe figuur 7.1. De data in deze figuur zijn tot stand gekomen door het opleggen van twee drempels: een drempel q_d voor Q en een drempel m_d voor M. Van de aanvankelijk ter beschikking staande data (q_i, m_i) zijn alleen díe paren geselecteerd waarvoor $q_i > q_d$ en/of $m_i > m_d$: indien zowel $q_i < q_d$ en $m_i < m_d$ is het punt (q_i, m_i) niet meegenomen in de selectie. Een dergelijke selectie van data komt in toepassingen nogal eens voor. Stel nu dat we ons correlatiemodel willen gebruiken, met Q en M getransformeerd naar X en Y. In de getransformeerde ruimte ziet de puntenwolk er iets anders uit dan in figuur 7.1, maar ook in die ruimte zal het 'uitgeknipte stuk' van figuur 7.1 zichtbaar zijn. De drempels q_d en m_d gaan dan over in drempels x_d en y_d. Zoals in paragraaf 2.2 kort werd besproken en zoals uitgebreider in paragraaf 7.2 aan de orde komt, moet dan de parameter s worden bepaald die gelijk is aan de standaarddeviatie van de conditionele verdeling $g(y|x) = \lambda_s(y-x-\delta(s))$ in de getransformeerde ruimte, vergelijk ook figuur 2.3. Het is duidelijk dat dat misgaat voor de getransformeerde versie van de data in figuur 7.1: voor $x < x_d$ is de verticale spreiding in de data (de spreiding gemeten langs een verticale lijn) kleiner dan voor $x > x_d$. Het meenemen van alle data levert dus een te kleine waarde voor s – de oplossing is alleen waarnemingen met $x > x_d$ mee te nemen, of equivalent hiermee, alleen waarnemingen met $q > q_d$. De analyse dient dus uitgevoerd te worden voor de in figuur 7.2 getoonde deelset van de data.

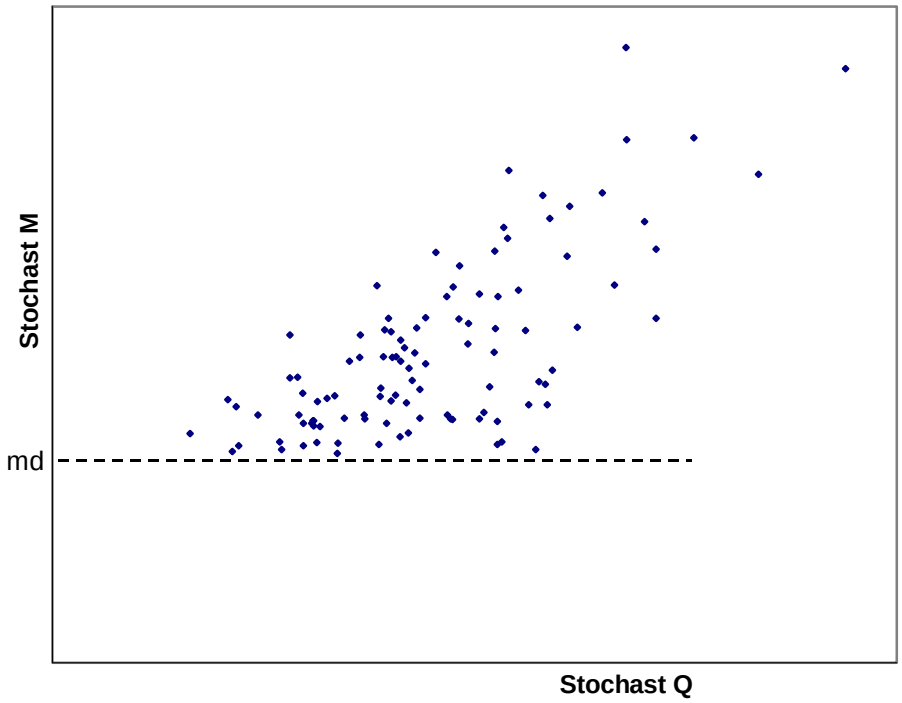


Figuur 7.1 Geselecteerde data: gepaarde data waarvoor of $q_i > q_d$ en/of $m_i > m_d$.

Tot slot de uitleg van punt 4. De stochast Q is in het voorgaande steeds getransformeerd naar de standaardexponentieel verdeelde variabele X en M naar de *asymptotisch* standaardexponentieel verdeelde variabele Y. De rol van de stochasten Q en M is dus niet symmetrisch. De transformatie van Q naar X blijkt namelijk veel eenvoudiger uit te voeren dan die van M naar Y. (Volgens paragraaf 7.2 hangt de laatste transformatie af van s en van het type verdeling voor $\lambda(t)$, terwijl dat voor de eerste transformatie niet geldt.) In toepassingen dient een keuze te worden gemaakt: welke stochast transformeren we naar X en welke naar Y? Of iets anders gezegd: wanneer we de notatie van dit rapport aanhouden, waarbij Q altijd naar X wordt getransformeerd en M naar Y: welke van de twee gegeven stochasten noemen we Q en welke M? Beide mogelijkheden lijken even werkbaar – de keus is aan degene die de analyse uitvoert. Let er wel op dat bij verwisseling van de rol van Q en M een andere deelset van de data wordt gebruikt: namelijk de deelset uit figuur 7.3.



Figuur 7.2 Deelset van de data uit figuur 7.1 waarvoor of $q_i > q_d$.



Figuur 7.3 Deelset van de data uit figuur 7.1 waarvoor of $m_i > m_d$.

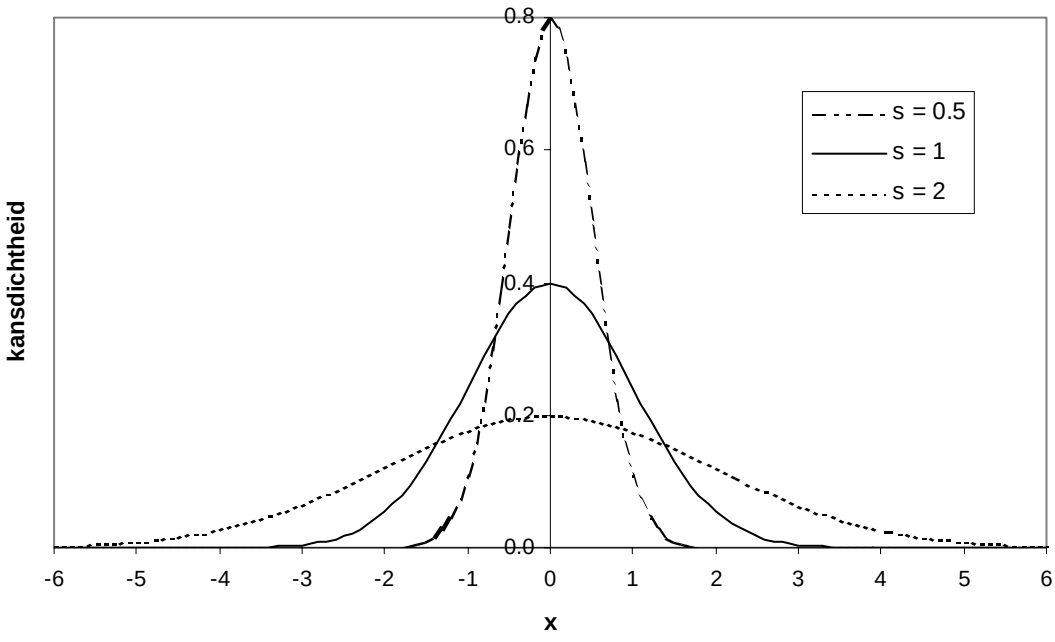
7.2 Recept voor het bepalen van de kansdichtheid uit de data

In deze paragraaf wordt een recept gegeven om de kansdichtheid g(q,m) op grond van de data te bepalen met het de hiervoor genoemde transformatieprocedure. De aanname is vanzelfsprekend, zoals in de voorgaande paragraaf werd behandeld, dat de dataset dusdanig is dat deze op voldoend nauwkeurige wijze door het beschouwde correlatiemodel kan worden beschreven. Het recept bestaat er vooral uit de juiste vorm voor $\lambda(t)$ te vinden. In de praktijk zijn de beschikbare data vaak zo spaarzaam dat het niet goed mogelijk is $\lambda(t)$ al te nauwkeurig te bepalen. Veelal zal men aannemen dat de verschillende kansdichtheden $\lambda(t)$ die in aanmerking komen behoren tot één type verdeling, bijvoorbeeld de normale verdeling. Men beschouwt dan normale verdelingen met verschillende standaarddeviaties en probeert díe standaarddeviatie te vinden die de data het best beschrijft. Het recept uit deze paragraaf beperkt zich dan ook tot één type verdeling.

Tevens zullen we slechts verdelingen met slechts twee parameters beschouwen, namelijk een locatie en een schaalparameter. De locatieparameter speelt feitelijk in het recept geen rol, omdat deze altijd zo wordt gekozen dat $\lambda(t)$ gemiddelde 0 heeft. De enige parameter die dan een rol speelt is de standaardeviatie s van $\lambda(t)$. Het is handig de parameter s expliciet in de notatie tot uitdrukking te brengen. We schrijven daarom $\lambda_s(t)$ om de kansdichtheid met standaardeviatie s aan te geven. Het komt er dan op neer dat een collectie kansdichtheden wordt beschouwd van de vorm

$$\lambda_s(t) = \frac{1}{s}\lambda_1\left(\frac{t}{s}\right) \quad , s > 0 \tag{7.1}$$

Deze kansdichtheden volgen uit $\lambda_1(t)$ door de t-as te schalen met een factor s en op de juiste manier te normeren. Figuur 7.4 geeft een voorbeeld voor de normale verdeling voor de waarden s = 1 (standaarnormale verdeling), s = 0.5 en s = 2. De collectie kansdichtheden ligt dus volledig vast door het geven van de kansdichtheid die standaarddeviatie 1 heeft. De kansdichtheden $\lambda_s(t)$ werden reeds in detail behandeld in paragraaf 6.2, waarnaar we ook verwijzen voor de preciese wiskundige voorwaarden waaraan $\lambda_s(t)$ moet voldoen. We herhalen hier slechts dat voor de normale verdeling aan de gestelde voorwaarden is voldaan en dat tevens is voldaan aan deze voorwaarden indien $\lambda_s(t)$ een eindige rechterstaart heeft.



Figuur 7.4 De normale kansdichtheden $\lambda_s(t)$ voor de waarden s = 0.5, s = 1 en s = 2.

In het vervolg wordt een iteratief recept gegeven om uit de beschikbare data g(q,m) te bepalen. De relevante formules zijn al uitgebreid aan de orde geweest in hoofdstuk 3 en paragraaf 4.2, 6.1 en 6.2 waarnaar we ook verwijzen voor verdere details. Als service aan de lezer worden de formules die van belang zijn voor het recept hier nog eens gegeven, dan bovendien, vanwege de belangrijke rol die s hier heeft, in termen van $\lambda_s(t)$ in plaats

van $\lambda(t)$. Ook andere grootheden krijgen hier de parameter s als index. (Voor δ zal de specifieke keuze $\delta(s)$ uit (6.14) worden genomen; in hoofdstuk 3 t/m 6 werd in plaats van $\delta(s)$ het symbool δ_0 gebruikt.)

De marginale kansdichtheid van X wordt gegeven door

(7.2) $$g(x) = e^{-x} \qquad , x \geq 0$$

De conditionele verdeling van Y gegeven X = x wordt volgens (3.11) gegeven door

(7.3) $$g_s(y \mid x) = \lambda_s(y - x - \delta(s)) \qquad , x \geq 0$$

met volgens (6.14)

(7.4) $$\delta(s) = -\ln\left(\int e^{st}\lambda_1(t)\,dt\right) < 0 \qquad , s > 0$$

De bivariate kansdichtheid in de getransformeerde ruimte wordt volgens (3.12) gegeven door

(7.5) $$g_s(x,y) = e^{-x}\lambda_s(y - x - \delta(s)) \qquad , x \geq 0\ , y \in \mathbb{R}$$

De cumulatieve verdeling geassocieerd met $\lambda_s(t)$ wordt gegeven door

(7.6) $$\Lambda_s(y) = \int_{-\infty}^{y} \lambda_s(t)dt = \Lambda_1(\frac{y}{s}) \qquad , s > 0$$

en de cumulatieve verdeling van Y door, zie (3.14)

(7.7) $$F_{Y,s}(y) = \int_{0}^{\infty} e^{-x}\Lambda_s(y - x - \delta(s))dx$$

Voor grote y wordt volgens (3.25) de verdeling van Y in benadering standaardexponentieel:

(7.8) $$F_{Y,s}(y) \cong 1 - e^{-y} \qquad , s > 0 \text{ en } y \text{ groot}$$

Hierbij geldt dat de benadering beter wordt naarmate s kleiner wordt. De marginale verdeling $g_s(y)$ wordt verkregen door differentiatie van (7.7) naar y; zie desgewenst (3.16) voor een alternatieve manier om $g_s(y)$ uit te rekenen.

De formules (7.1) t/m (7.8) zijn alle geassocieerd met de getransformeerde ruimte. Het verband tussen de originele en de getransformeerde ruimte wordt volgens (6.1) en (6.2) gegeven door de transformaties x = J(q) en y(s) = K_s(m). Ieder punt (x,y) in de getransformeerde ruimte hangt (behoudens in irreguliere situaties) samen met precies één punt (q,m) in de originele ruimte en vice versa. De transformaties van q en m worden ieder apart uitgevoerd. Om bijvoorbeeld de transformatie van q naar x uit te voeren zijn de waarden van m en y niet relevant. Merk op dat we in plaats van J(q) ook simpelweg x(q) zouden kunnen schrijven omdat de transformatie van x naar q simpelweg inhoudt dat x een functie is van q; analoog zouden we ook y(m,s) in plaats van y(s) = K_s(m) kunnen schrijven. Omdat x en y soms als 'zelfstandige' variabelen voorkomen, zonder dat ze van q en m afhangen, is er voor de duidelijkheid voor gekozen om de functies J(q) en K_s(m) te gebruiken indien x en y wél van q en m afhangen.

Volgens (6.1) en (6.5) kan de transformatie van q naar x eenvoudig worden berekend met

(7.9) $$x = J(q) = -\ln\left(1 - F_Q(q)\right) \geq 0$$

waarbij F_Q(q) de (bekend veronderstelde) cumulatieve verdeling geeft van de stochast Q in de originele ruimte. De transformatie van m naar y is volgens (6.1) en (6.2) van de vorm

(7.10) $y = K_s(m) \qquad , s > 0$

waarbij $K_s(m)$, bij een vaste beschouwde waarde van s, volgt door het (numeriek) oplossen van de vergelijking

(7.11) $F_{Y,s}(K_s(m)) = F_M(m) \qquad , s > 0$

Hierbij wordt $F_{Y,s}$ gegeven door (7.7) en geeft $F_M(m)$ de (bekend veronderstelde) cumulatieve verdeling geeft van de stochast M in de originele ruimte. De bivariate kansdichtheid in de originele ruimte wordt volgens (6.7) en (6.8) gegeven door

(7.12) $g_s(q,m) = g(q)\,\lambda_s\left(K_s(m) - J(q) - \delta(s)\right)\dfrac{dK_s(m)}{dm}$

Deze kansdichtheid heeft als marginale verdelingen g(q) en g(m). We merken nog op dat de (dimensieloze grootheid) s in de getransformeerde ruimte kan worden geïnterpreteerd als de spreiding van de puntenwolk bij gegeven x. In de originele ruimte heeft s geen directe interpretatie, zij het dat deze indirect wel samenhangt met de spreiding van de puntenwolk in de originele ruimte bij gegeven q. De cumulatieve verdeling van M, gegeven Q = q, wordt volgens (6.9) gegeven door

(7.13) $F_s(m \mid q) = \Lambda_s\left(K_s(m) - J(q) - \delta(s)\right)$

We zijn nu gereed om het recept aan te geven om op grond van de data de bivariate kansdichtheid $g_s(q,m)$ te bepalen. Uit formules (7.1) t/m (7.12) blijkt dat de enige vrijheidsgraad in het hier gebruikte correlatiemodel de parameter s is. Het recept bestaat eruit op iteratieve wijze de waarde van s te bepalen die het beste bij de data past. In iedere iteratiestap (behalve de eerste) wordt daarbij de transformatie van m naar y(s) uitgevoerd voor de dan beschouwde waarde van s. De transformatie van q naar x wordt slechts eenmaal uitgevoerd, in de eerste stap van het recept. Nadat het recept gegeven is volgt een uitgebreide toelichting aan de hand van een voorbeeld waarin $\lambda_1(t)$ uit (7.1) de standaard uniforme verdeling volgt.

Recept voor het bepalen van s door iteratie

Neem aan dat g(q) en g(m) vooraf gegeven zijn en ga uit van de formules (7.1) t/m (7.12). Neem tevens aan dat N gecombineerde waarnemingen (q_i, m_i), i = 1, 2, ..., N beschikbaar zijn die beschouwd mogen worden als onafhankelijke trekkingen uit een (vooralsnog) onbekende kansdichtheid g(q,m).

Stap 1
Transformeer q_i naar de standaardexponentieel verdeelde x_i volgens (7.9). Dus

(7.14) $x_i = -\ln\left(1 - F_Q(q_i)\right) \geq 0$

Transformeer m_i op dezelfde wijze. Dus

(7.15) $y_i = -\ln\left(1 - F_M(m_i)\right) \geq 0$

Dit levert N datapunten (x_i, y_i), i = 1, 2, ..., N in de getransformeerde ruimte. Een schatting van de 'standaarddeviatie ten opzichte van de lijn y = x' wordt gegeven door

(7.16) $s_1 = \sqrt{\dfrac{\sum_{i=1}^{N}(y_i - x_i)^2}{N-1}}$

Stap 2
Ga nu uit van de waarde van de standaarddeviatie uit de voorgaande stap en transformeer dan m_i naar $y_i(s_1)$ volgens (7.10). Dus

(7.17) $y_i(s_1) = K_{s_1}(m_i)$

Dat levert N datapunten (x_i, $y_i(s_1)$) in de getransformeerde ruimte. Een schatting van de 'standaarddeviatie ten opzichte van de lijn $y = x + \delta(s_1)$' wordt gegeven door

(7.18) $$s_2 = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i(s_1) - x_i - \delta(s_1)\right)^2}{N-1}}$$

Stap 3
De voorgaande stap wordt nu herhaald met de standaarddeviatie s_2. Voor de duidelijkheid wordt deze stap hier uitgeschreven. Er geldt dus

(7.19) $y_i(s_2) = K_{s_2}(m_i)$

Dat levert N datapunten (x_i, $y_i(s_2)$) in de getransformeerde ruimte. Een schatting van de 'standaarddeviatie ten opzichte van de lijn $y = x + \delta(s_2)$' wordt gegeven door

(7.20) $$s_3 = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i(s_2) - x_i - \delta(s_2)\right)^2}{N-1}}$$

Zo verdergaande worden successievelijk s_1, s_2, s_3,..... bepaald. Het iteratieproces kan worden gestopt wanneer de standaarddeviaties in opeenvolgende stappen (volgens een of ander criterium) nauwelijks meer verschillen. De limietwaarde vormt dan de gezochte s uit $\lambda_s(t)$. Het wordt ten strengste aanbevolen de puntenwolk, die in de n-de stap bestaat uit de N punten (x_i, $y_i(s_{n-1})$), in een plaatje uit te zetten. Aldus kan grafisch beoordeeld worden of met het gebruikte correlatiemodel de data goed kunnen worden beschreven. Tevens verdient het aanbeveling de puntenwolk in de n-de stap te vergelijken met de percentielenlijnen (zie het voorbeeld hieronder voor nadere uitleg) van $g_{s_{n-1}}(y \mid x)$.

7.3 Toepassing van het recept op een voorbeeld

Het recept uit de voorgaande paragraaf zal worden toegelicht met een voorbeeld waarin Q de IJsselafvoer en M het IJsselmeerpeil voorstelt. Met nadruk wordt gesteld dat de gegevens in het voorbeeld *louter fictief zijn!* Dat hier over de IJsselafvoer en het IJsselmeerpeil wordt gesproken dient om zoveel mogelijk een concrete situatie als voorbeeld te nemen. We stellen ons voor dat maandmaxima beschikbaar zijn over een periode van 30 winterhalfjaren, waarbij het winterhalfjaar bestaat uit de maanden oktober t/m maart. Aldus zijn N = 6*30 = 180 waarnemingen (q_i, m_i), i = 1, 2, 3,..., N beschikbaar die onafhankelijk worden verondersteld. Zie figuur 7.5 voor de in dit voorbeeld gebruikte data. Voor de cumulatieve verdelingsfunctie van Q en M zijn in dit voorbeeld Gumbelverdelingen genomen van de volgende vorm. Hoewel het fictieve gegevens betreft zijn deze verdelingen wel zo gekozen dat ze voor hoge q en m wel enigszins realistische uitkomsten tot gevolg hebben.

$$F_Q(q) = \exp\left(-\exp\left[-\frac{q - b_Q}{a_Q}\right]\right)$$

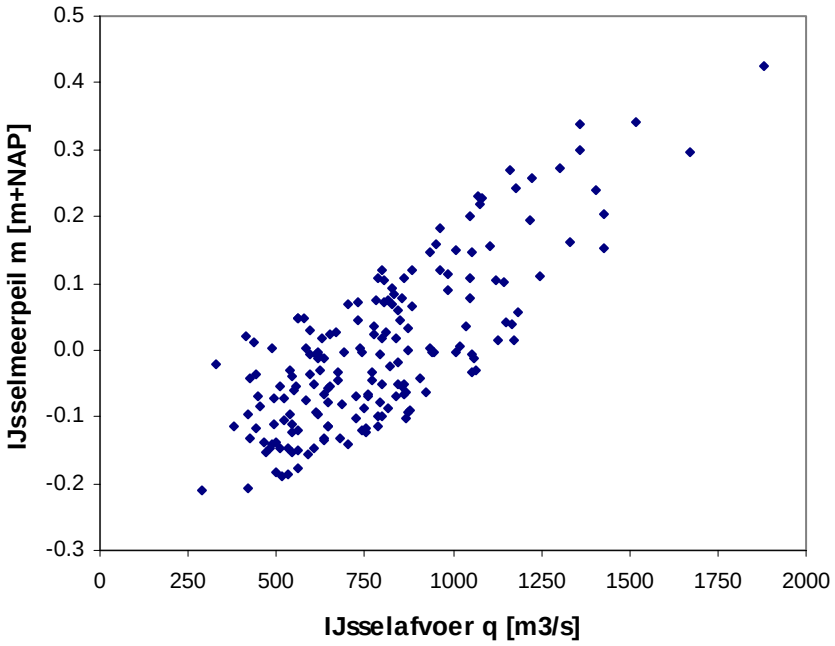
(7.21) $a_Q = 220\ m^3/\text{s}$

$b_Q = 675\ m^3/\text{s}$

$$F_M(m) = \exp\left(-\exp\left[-\frac{q-b_M}{a_M}\right]\right)$$

(7.22) $a_M = 0.10$ m

$b_M = -0.05$ m + NAP



Figuur 7.5 Gecombineerde maandmaxima voor het IJsselmeerpeil en de IJsselafvoer, N = 180 (fictieve gegevens).

De data in figuur 7.5 zijn afkomstig uit een simulatie, op zo'n manier dat in de getransformeerde ruimte de kansdichtheid $\lambda_s(t)$ een uniforme verdeling vormt met standaarddeviatie s = 0.693; deze waarde correspondeert met de uniforme verdeling op het interval (-1.2, 1.2). Vanwege 'de herkomst' van de data is het evident dat de puntenwolk na transformatie kan worden beschreven door het hiervoor beschouwde correlatiemodel. De correlatie die uit de data in figuur 7.5 blijkt is 'bijzonder netjes'. Werkelijke data laten in de regel een veel minder fraaie puntenwolk zien.

We geven nu enkele formules die betrekking hebben op de getransformeerde ruimte waarin g(y|x) de uniforme verdeling volgt. Deze formules werden reeds in hoofdstuk 5 afgeleid. De formules worden hier enigszins herschreven om de afhankelijkheid van s expliciet te maken. Daarbij wordt voor δ uit hoofdstuk 5 de keuze $\delta = \delta(s)$ gemaakt (waarbij δ(s) ook weer gelijk is aan δ_0 uit hoofdstuk 5). De uniforme kansdichtheid $\lambda_s(t)$ met standaarddeviatie s wordt gegeven door, zie (5.1) en (5.2),

$$\lambda_s(t) = \begin{cases} \dfrac{1}{2\sqrt{3}\,s} & , -\sqrt{3}\,s < t < \sqrt{3}\,s \\ 0 & , \text{anders} \end{cases} \tag{7.23}$$

De parameter δ(s) wordt volgens (5.3) gegeven door

$$\delta(s) = -\ln\left(\frac{e^{\sqrt{3}s} - e^{-\sqrt{3}s}}{2\sqrt{3}\,s}\right) \tag{7.24}$$

terwijl de cumulatieve verdeling van Y wordt gegeven door, zie (5.6),

$$(7.25)\quad F_{Y,s}(y)=\begin{cases}1-e^{-y} & ,\ y\ge\sqrt{3}\,s+\delta(s)\\ \dfrac{1}{2}-\dfrac{1}{2\sqrt{3}\,s}\left(1-y+\delta(s)-e^{-y+\delta(s)-\sqrt{3}\,s}\right) & ,\ -\sqrt{3}\,s+\delta(s)<y<\sqrt{3}\,s+\delta(s)\\ 0 & ,\ y\le-\sqrt{3}\,s+\delta(s)\end{cases}$$

en de kansdichtheid $g_s(y)$ door, zie (5.7),

$$(7.26)\quad g_s(y)=\begin{cases}e^{-y} & ,\ y\ge\sqrt{3}\,s+\delta(s)\\ \dfrac{1}{2\sqrt{3}\,s}\left(1-e^{-y+\delta(s)-\sqrt{3}\,s}\right) & ,\ -\sqrt{3}\,s+\delta(s)<y<\sqrt{3}\,s+\delta(s)\\ 0 & ,\ y\le-\sqrt{3}\,s+\delta(s)\end{cases}$$

Nu zullen achtereenvolgens de stappen uit het recept in paragraaf 7.2 worden doorlopen, waarbij de standaarddeviatie door middel van een iteratieproces wordt bepaald. Als dit een goed recept vormt zal de uit het simulatieproces reeds bekende standaarddeviatie in redelijke benadering moeten volgen uit het iteratieproces.

Stap 1
De transformatie van q_i en m_i naar standaardexponentieel verdeelde x_i en y_i kan vanwege (7.14), (7.15), (7.21) en (7.22) worden geschreven als, voor i = 1, 2,..., N,

$$(7.27)\quad x_i=-\ln\left\{1-\exp\left(-\exp\left[-\frac{q_i-b_Q}{a_Q}\right]\right)\right\}$$

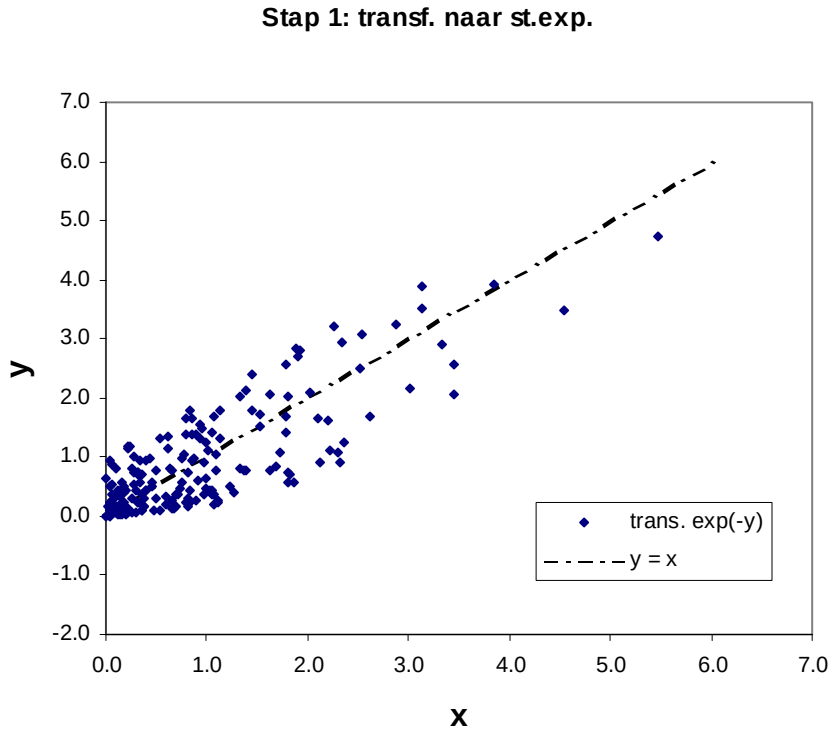
$$(7.28)\quad y_i=-\ln\left\{1-\exp\left(-\exp\left[-\frac{m_i-b_M}{a_M}\right]\right)\right\}$$

Onder deze transformatie gaat figuur 7.5 over in figuur 7.6. In de figuur is te zien dat de y_i zich als het ware 'ophopen' in de buurt van y = 0. De puntenwolk voldoet niet aan de eis van het correlatiemodel dat g(y|x) op een verschuiving langs de lijn y = x na gelijk is aan een vaste kansdichtheid; de spreiding voor lage x-waarden is duidelijk kleiner dan die voor hoge x-waarden. In deze eerste stap moet dat echter ook zeker niet verwacht worden. De hier toegepaste transformatie van de m_i naar de y_i is namelijk niet de juiste transformatie. De juiste transformatie zou van de vorm $y_i(s) = K_s(m_i)$ zijn, met dan ook nog eens de juiste waarde van s. Van deze transformatie is slechts bekend volgens (7.8) en (7.11) dat deze neerkomt op een transformatie naar de standaardexponentiële verdeling voor *grote* waarden van y. In de figuur lijkt inderdaad voor grote waarden van y wel voldaan te zijn aan de eis dat g(y|x) op een verschuiving langs een rechte lijn na gelijk is aan een vaste kansdichtheid.

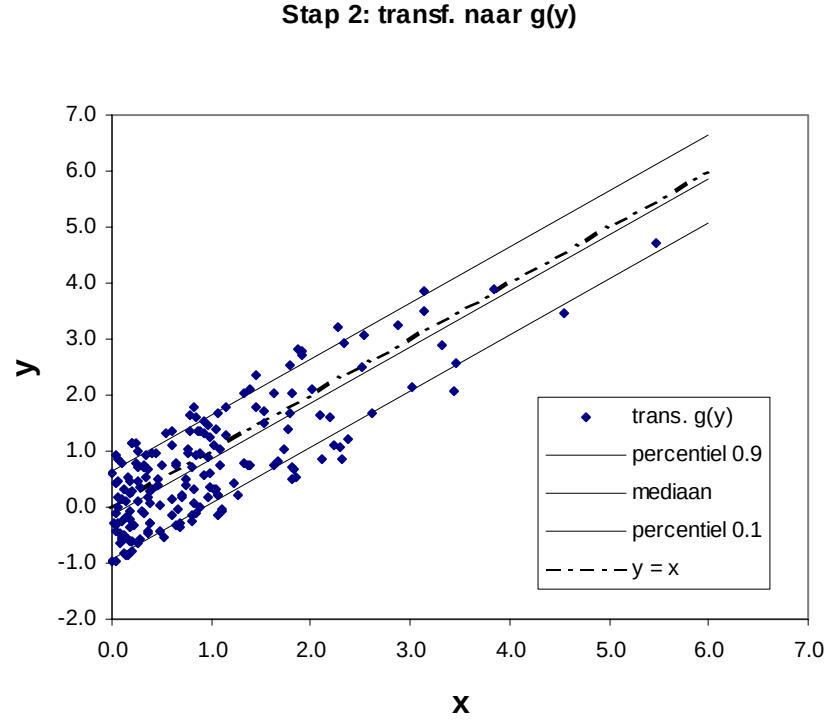
De waarde s_1 volgens (7.16) blijkt gelijk te zijn aan

$$(7.29)\quad s_1=0.564$$

Het verschil $y_i - x_i$ dat in de sommatie onder de wortel in (7.16) voorkomt is gelijk aan de verticaal gemeten afstand van (y_i, x_i) tot de lijn y = x. Vandaar dat (7.16) herkend kan worden als een gebruikelijke formule om de standaarddeviatie te schatten van een kansdichtheid die gecentreerd is rond de lijn y = x. Feitelijk zijn de data in figuur 7.6 niet in overeenstemming met een dergelijke kansdichtheid omdat de data duidelijk niet symmetrisch met eenzelfde spreiding rond de lijn y = x liggen. Dat s_1 niet werkelijk als een standaarddeviatie is op te vatten is geen enkel probleem, omdat dit getal slechts een beginwaarde vormt in een iteratieproces dat uiteindelijk wel tot een helder te interpreteren standaarddeviatie moet leiden. Soortgelijke opmerkingen zijn ook te maken bij de volgende iteratiestappen zolang nog geen convergentie naar een 'limietwaarde' van s is opgetreden.



Figuur 7.6 De waarnemingen (x_i, y_i) die resulteren na de transformatie naar de standaardexponentiële verdeling.



Figuur 7.7 De waarnemingen (x_i, $y_i(s_1)$) die resulteren na de transformatie in stap 2. De mediane lijn valt in dit voorbeeld samen met de lijn $y = x + \delta(s_1)$.

Stap 2
Met de waarde s_1 als standaarddeviatie uit de voorgaande stap wordt m_i getransformeerd naar $y_i(s_1)$ volgens (7.17); de functie $K_{s1}(m)$ is daarbij numeriek bepaald door gebruik te maken van (7.11), (7.22) en (7.25). De resulterende N datapunten $(x_i, y_i(s_1))$ in de getransformeerde ruimte zijn weergegeven in figuur 7.7. Een schatting van de 'standaarddeviatie ten opzichte van de lijn $y = x + \delta(s_1)$' wordt volgens (7.18) gegeven door

(7.30) $s_2 = 0.653$

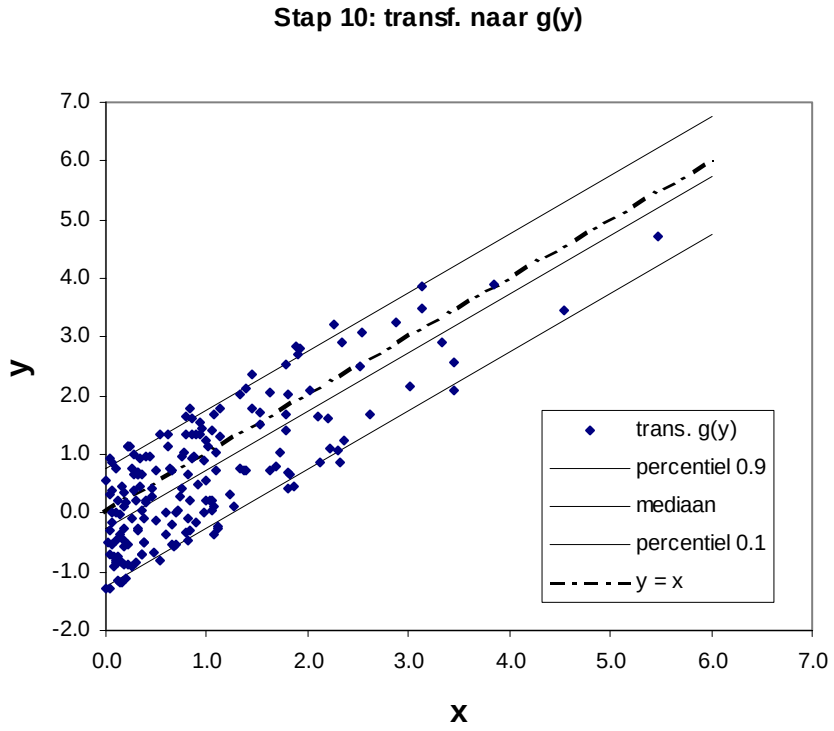
Op de genoemde lijn liggen de gemiddelden $E(Y|X=x)$ van de conditionele kansdichtheden $g(y|x)$; zie eventueel (3.13) voor nadere uitleg. Omdat de mediaan van de uniforme verdeling samenvalt met het gemiddelde van deze verdeling is de mediane lijn (waarop de medianen van de kansdichtheden $g(y|x)$ liggen) in figuur 7.7 gelijk aan de lijn $y = x + \delta(s_1)$. Tevens zijn in de figuur het 10% - en het 90% - percentiel aangegeven.

Zo op het oog het oog lijkt de waarde $s_2 = 0.653$ al vrij goed bij de data te passen. Tevens lijkt het zo te zijn dat bij iedere gegeven x de conditionele kansdichtheid $g(y|x)$ op een verschuiving langs de lijn $y = x + \delta(s_1)$ na gelijk is aan een vaste kansdichtheid, die dan gelijk is aan $\lambda_{0.653}(t)$ volgens formule (7.23). Hooguit lijken er toch nog iets te veel waarnemingen buiten de percentiellijnen (waarbuiten ongeveer 20% van de waarnemingen mag liggen) te vallen.

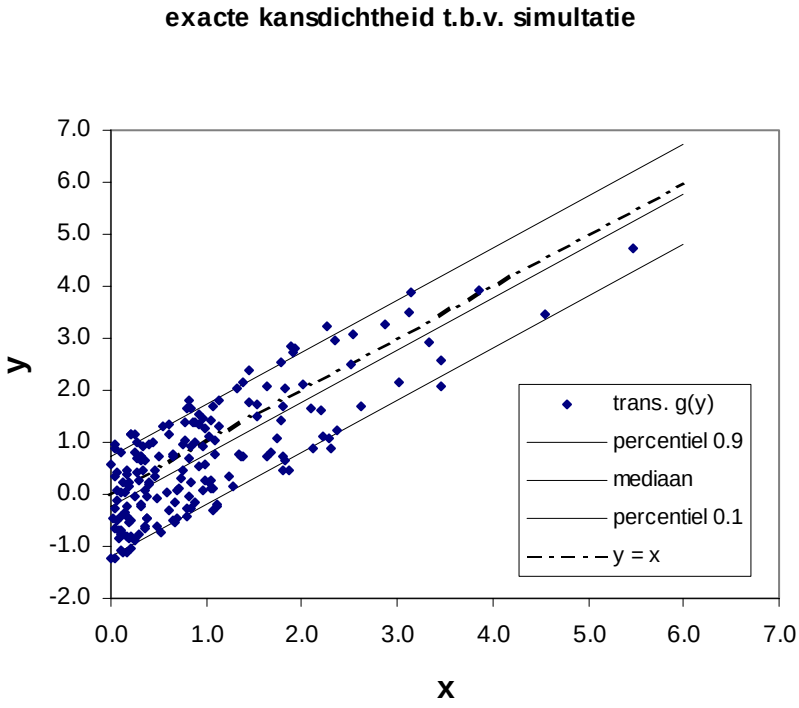
stap i	s_i	delta(s_i)	a(s_i)
1	0.564	-0.154	0.978
2	0.653	-0.205	1.132
3	0.689	-0.227	1.193
4	0.704	-0.237	1.220
5	0.711	-0.241	1.232
6	0.715	-0.244	1.238
7	0.716	-0.244	1.240
8	0.717	-0.245	1.241
9	0.717	-0.245	1.242
10	0.717	-0.245	1.242
exact	0.693	-0.229	1.2

Tabel 7.1 De waarden van s_i, $\delta(s_i)$ en $a(s_i) = \sqrt{3}\, s_i$ voor i = 1, 2,..., 10. Tevens zijn de exacte waarden op basis waarvan de simulatie is uitgevoerd gegeven.

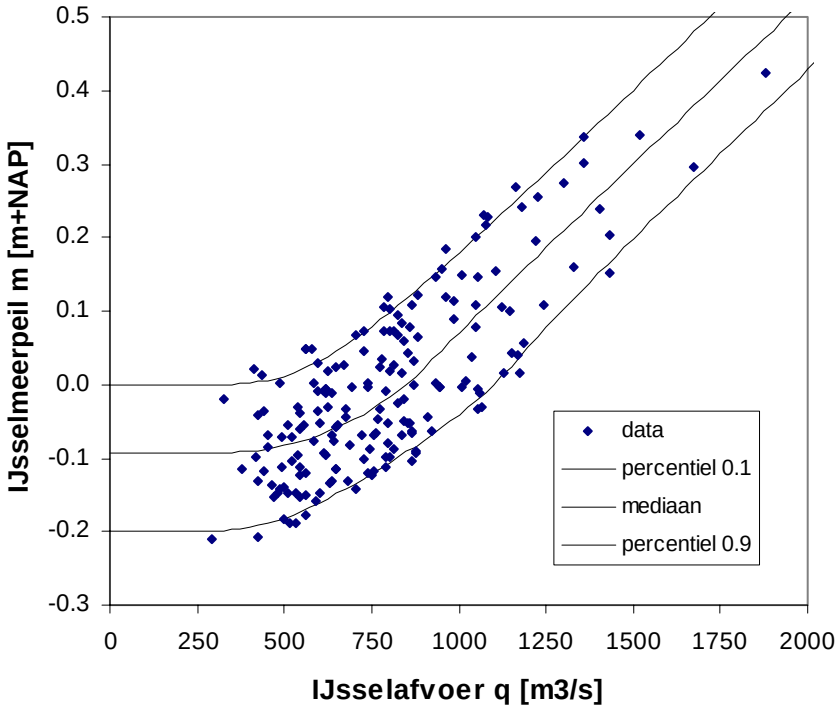
Wanneer tot en met stap 10 de volgende stappen uit het iteratieproces worden doorlopen resulteren de uitkomsten volgens tabel 7.1. De uniforme verdeling $\lambda_s(t)$ heeft als drager het interval (-a(s), a(s)). Naast de waarden s_i en $\delta(s_i)$ worden de $a(s_i)$ eveneens in de tabel aangegeven. Uit de tabel blijkt dat na ongeveer 5 á 6 iteraties convergentie optreedt. Het verschil van de uiteindelijk gevonden $s_{10} = 0.717$ met de exacte waarde van s op basis waarvan de simultatie is uitgevoerd blijkt gelijk te zijn aan $0.717 - 0.693 = 0.024$. Figuur 7.8 geeft het plaatje voor de tiende stap. In figuur 7.9 worden de gegevens weergegeven op basis van de exacte waarden die in het simulatieproces zijn gebruikt. Het is duidelijk dat het gebruikte recept (tenminste in dit voorbeeld) de oorspronkelijke kansdichtheid vrij nauwkeurig weet te reproduceren. In de volgende paragraaf zullen 9 andere runs worden beschouwd. Ter informatie melden we hier dat op een enkele uitzondering na steeds na 5 á 10 stappen convergentie blijkt te zijn opgetreden op zo'n manier dat de derde decimaal dan niet meer verandert. Veelal is reeds na 3 stappen al goeddeels de eindwaarde van s bereikt.



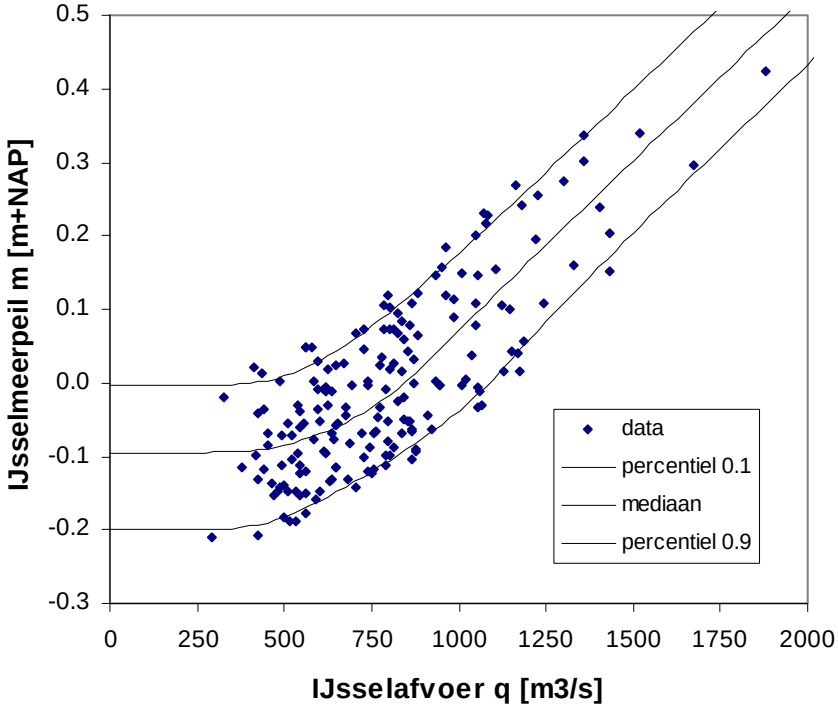
Figuur 7.8 De waarnemingen (x_i, $y_i(s_9)$) die resulteren na de transformatie in stap 10. De mediane lijn valt in dit voorbeeld samen met de lijn $y = x + \delta(s_9)$.



Figuur 7.9 De waarnemingen in de getransformeerde ruimte zoals bepaald met de exacte waarde s = 0.693 De mediane lijn valt hier samen met de lijn $y = x + \delta(0.693)$.



Figuur 7.10 De waarnemingen in de originele ruimte zoals bepaald met de in het iteratieproces gevonden waarde s = 0.717 tezamen met de percentielenlijnen van de kansdichtheid $g_s(m|q)$.



Figuur 7.11 De waarnemingen in de originele ruimte zoals bepaald met de exacte waarde s = 0.693 tezamen met de percentielenlijnen van de kansdichtheid $g_s(m|q)$.

De figuren 7.6 tot en met 7.9 geven de plaatjes in de getransformeerde ruimte. Voor de originele ruimte geeft figuur 7.10 het plaatje, volgens de met het recept gevonden waarde s = 0.717. Ter vergelijking geeft figuur 7.11 het plaatje met de exacte waarde s = 0.693. Beide figuren lijken sterk op elkaar. Het is dus duidelijk dat de met het recept gevonden waarde van s en de daarmee corresponderende kansdichtheid de werkelijke waarde van s en de werkelijke kansdichtheid goed benadert. Uiteraard moest dat laatste in het licht van de figuren 7.8 en 7.9 ook wel verwacht worden.

Misschien acht de lezer het overbodig om zowel de plaatjes in de getransformeerde als de originele ruimte te geven. Omdat het voorbeeld in deze paragraaf een erg nette puntenwolk betreft is dat hier feitelijk ook zo. In toepassingen zullen echter minder nette puntenwolken worden aangetroffen, die nooit helemaal exact door het hier beschouwde correlatiemodel kunnen worden beschreven. Dan is het zeker belangrijk om in beide ruimtes de plaatjes te maken. In paragraaf 7.5 wordt hier op teruggekomen.

Het is belangrijk om in te gaan op de rol van de parameter s in de originele ruimte. Deze parameter geeft in de getransformeerde ruimte de standaarddeviatie van de als conditionele verdeling gebruikte $\lambda_s(t)$. Het ligt voor de hand dat s in de originele ruimte eveneens samenhangt met de spreiding van de conditionele verdeling. Dat blijkt inderdaad zo te zijn, echter niet op de manier dat de spreiding in de originele ruimte recht evenredig is met s. Ter illustratie zijn in figuur 7.12 voor een kleine en een grote waarde van s de percentielenlijnen gegeven. De smalle bundel lijnen correspondeert met s = 0.2 en de brede met s = 2.5. Beide bundels geven zoals uiteraard verwacht moest worden een slechte fit aan de data. De smalle bundel loopt nog wel redelijk door het centrum van de data maar geeft een veel smallere spreiding dan in de data het geval is. De brede bundel geeft een veel grotere spreiding dan de data en loopt bovendien niet door het centrum van de data; zo loopt bijvoorbeeld de mediaan voor de hogere data onder de puntenwolk langs. Voor de duidelijkheid melden we dat $g_s(q,m)$ voor elke waarde s > 0 als marginale verdelingen de voorgeschreven g(q) en g(m) heeft; het feit dat voor te kleine en te grote waarden van s de correlatie niet goed wordt beschreven staat dus los van het feit dat $g_s(q,m)$ de juiste marginale verdelingen heeft.

Het is interessant te bekijken hoe $g_s(q,m)$ er uit gaat zien in ten eerste de situatie waarin s naar nul gaat en ten tweede de situatie waarin s naar oneindig gaat. In de limiet s naar nul blijkt de smalle bundel over te gaan in een enkele lijn; in de limiet s nadert tot oneindig blijken de percentielenlijnen horizontaal te gaan lopen. Zie ter illustratie figuur 7.13. Het gedrag voor kleine s is zoals verwacht. Misschien wekt het in eerste instantie verbazing dat voor grote s de spreiding niet langer meer toeneemt maar (in de originele ruimte) naar een eindige waarde gaat. Dat laatste moet echter wel het geval zijn! De marginale verdeling g(m) is namelijk vooraf gegeven en heeft een eindige standaarddeviatie. Dat betekent dat de afstand tussen de percentielenlijnen niet onbeperkt kan toenemen, omdat dan de marginale verdeling g(m) een oneindig grote spreiding zou krijgen.

Het blijkt mogelijk de vergelijking van de lijn voor s = 0 en de vergelijkingen voor de lijnen voor s = ∞ expliciet aan te geven. Dat is zelfs mogelijk voor willekeurige λ(t) en niet alleen voor de hier beschouwde uniforme verdeling. De precieze wiskundige formuleringen van onderstaande beweringen worden gegeven in de appendix – hier geven we slechts de resultaten. Beschouw eerst s = 0. In het algemene geval blijkt de lijn uit figuur 7.13 geen rechte te zijn maar een ‘kromme’ lijn die we zullen weergeven met m(q). Het meerpeil m(q) is dus een functie van de afvoer q. Deze functie blijkt zodanig te zijn dat de onderschrijdingskans van m(q) gelijk is aan die van q, wat in formule betekent

$$F_M\left(m(q)\right) = F_Q(q) \tag{7.31}$$

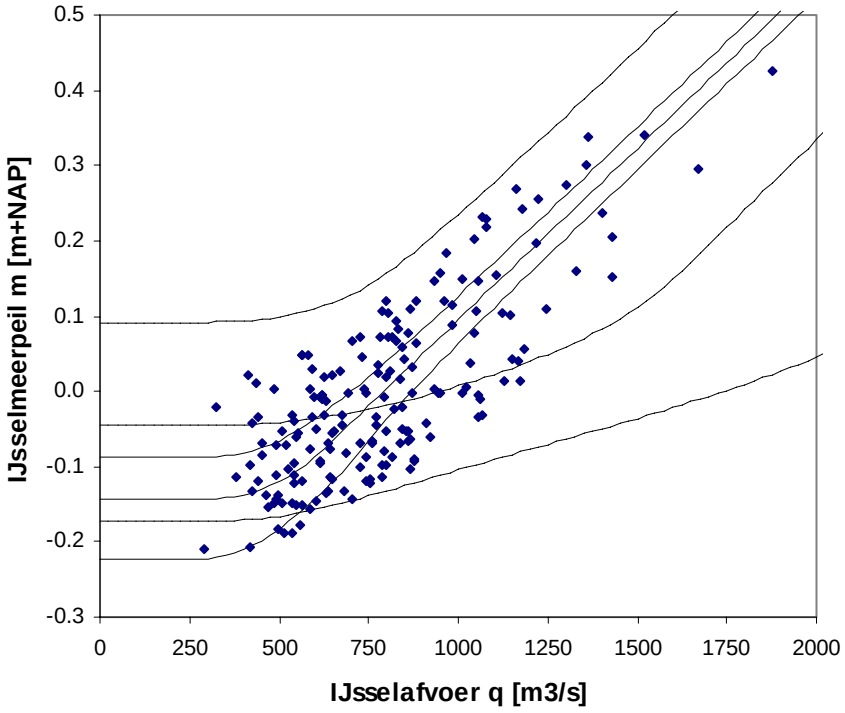
Het is evident dat dit de situatie van maximale correlatie tussen Q en M weergeeft. De lijn m(q) zullen we soms aanduiden als de *lijn met gelijke kansen.* Voor de lezer die bekend is met de *delta functie*, ook wel de *Dirac delta functie* genoemd, merken we op dat de limiet van $g_s(q,m)$ voor s nadert tot nul kan worden weergegeven in termen van deze deltafuntie. Wanneer de deltafunctie met locatieparameter 0 wordt aangegeven als $\delta_D(t)$, geldt

$$\lim_{s\downarrow 0} g_s(q,m) = g(q)\delta_D\left(m - m(q)\right) \tag{7.32}$$

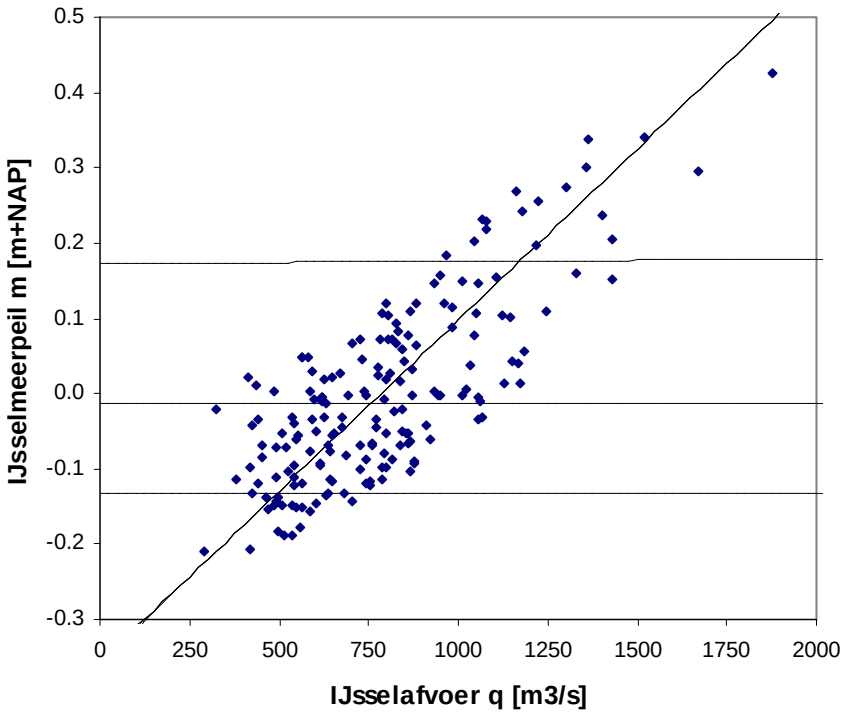
Beschouw nu de situatie waarin s nadert tot oneindig. Hier blijkt simpelweg te gelden dat de limietverdeling correspondeert met onafhankelijkheid van Q en M:

$$\lim_{s\to\infty} g_s(q,m) = g(q)g(m) \tag{7.33}$$

De percentielenlijnen van $g_s(m|q)$ worden in de limiet s nadert tot oneindig dan gelijk aan de percentielenlijnen van g(m).



Figuur 7.12 De percentielenlijnen van de kansdichtheid $g_s(m|q)$ voor s = 0.2 (smalle bundel) en s = 2.5 (brede bundel.



Figuur 7.13 De percentielenlijnen van de kansdichtheid $g_s(m|q)$ in de limiet s gaat naar nul (enkele lijn) en in de limiet s nadert oneindig (horizontale bundel).

7.4 Maximum likelyhood als alternatief recept

Het recept uit de voorgaande paragrafen met iteratie op de standaarddeviatie s vormt slechts één methode om s te bepalen. Een andere voor de hand liggende methode is die van *maximum likelyhood*. Die zal nu worden behandeld. Vervolgens zullen naast de simulatie uit paragraaf 7.3 nog 9 andere simulaties worden bekeken. Deze simulaties worden voortaan *runs* genoemd. De verschillende methoden zullen dan worden vergeleken aan de hand van deze 10 runs.

De maximum likelyhood methode (ML-methode) wordt hieronder toegepast op de originele zowel als op de getransformeerde ruimte, met in het algemeen (behalve in bijzondere gevallen) verschillende resultaten. Zometeen zal worden toegelicht dat bij toepassing op de getransformeerde ruimte feitelijk geen sprake is van een 'nette' standaardtoepassing van de ML-methode. De ML-methode voor de getransformeerde ruimte dient slechts ter vergelijking, maar kan in toepassingen beter niet worden gebruikt.

Nu volgen de formules voor de ML-methode voor de getransformeerde ruimte. We gaan weer uit van de in paragraaf 7.2 beschreven situatie. Na transformatie gaan de N datapunten (q_i, m_i) over in de punten (x_i, $y_i(s)$) met volgens (7.9) en (7.10)

(7.34) $$x_i = J(q_i) = -\ln\left(1 - F_Q(q_i)\right) \geq 0$$

(7.35) $$y_i = K_s(m_i)$$

waarbij $K_s(m)$ wordt vastgelegd door (7.11). Met behulp van (7.5) kan de *likelyhood* in de getransformeerde ruimte dan worden geschreven als

(7.36) $$L_{tr}(s) = \prod_{i=1}^{N} g_s(x_i, y_i(s)) = \prod_{i=1}^{N} e^{-x_i} \lambda_s(y_i(s) - x_i - \delta(s))$$

terwijl de natuurlijke logaritme daarvan kan worden geschreven als

(7.37) $$\ln\left(L_{tr}(s)\right) = -\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} \ln\left(\lambda_s(y_i(s) - x_i - \delta(s))\right)$$

Als schatting voor de standaarddeviatie die de data het best beschrijft kan dan de waarde van s worden genomen waarvoor $L_{tr}(s)$, of equivalent daarmee $\ln(L_{tr}(s))$, maximaal is. Omdat in deze maximalisatie slechts één parameter een rol speelt is de juiste waarde van s eenvoudig op numerieke wijze te bepalen door te beginnen met een zeer kleine startwaarde > 0. Deze wordt dan steeds iets opgehoogd totdat blijkt voor welke waarde van s het maximum optreedt. Waarom is zoals hiervoor gezegd het toepassen van de ML-methode op de getransformeerde ruimte feitelijk incorrect? Dat komt omdat in de maximalisatie geen vaste set van data wordt gebruikt. De $y_i(s)$ hangen immers van s af, zodat bij iedere waarde van s als het ware andere data horen.

We zullen nu de correcte ML-methode toepassen op de originele data. Deze data zijn wél onafhankelijk van s. Voor de originele ruimte wordt de likelyhood gegeven door, zie (7.12),

(7.38) $$L_{orig}(s) = \prod_{i=1}^{N} g_s(q_i, m_i) = \prod_{i=1}^{N} g_Q(q_i) K_s'(m_i) \lambda_s\left(K_s(m_i) - J(q_i) - \delta(s)\right)$$

Door maximalisatie kan weer de juiste waarde van s worden bepaald. Om het verband met (7.36) te onderzoeken herschrijven we (7.38) iets. Merk op dat door differentiatie van (7.11) naar m volgt

(7.39) $$K_s'(m) = \frac{g_M(m)}{g_{Y,s}\left(K_s(m)\right)}$$

Met (7.34) en (7.35) volgt dan dat $L_{orig}(s)$ kan worden geschreven als

$$(7.40) \quad L_{orig}(s) = \prod_{i=1}^{N} g_Q(q_i) g_M(m_i) \frac{\lambda_s\left(y_i(s) - x_i - \delta(s)\right)}{g_{Y,s}\left(y_i(s)\right)}$$

met als natuurlijke logaritme daarvan

$$(7.41) \quad \ln\left(L_{orig}(s)\right) = \sum_{i=1}^{N} \ln\left(g_Q(q_i) g_M(m_i)\right) + \sum_{i=1}^{N} \ln\left(\lambda_s\left(y_i(s) - x_i - \delta(s)\right)\right) - \sum_{i=1}^{N} \ln\left(g_{Y,s}\left(y_i(s)\right)\right)$$

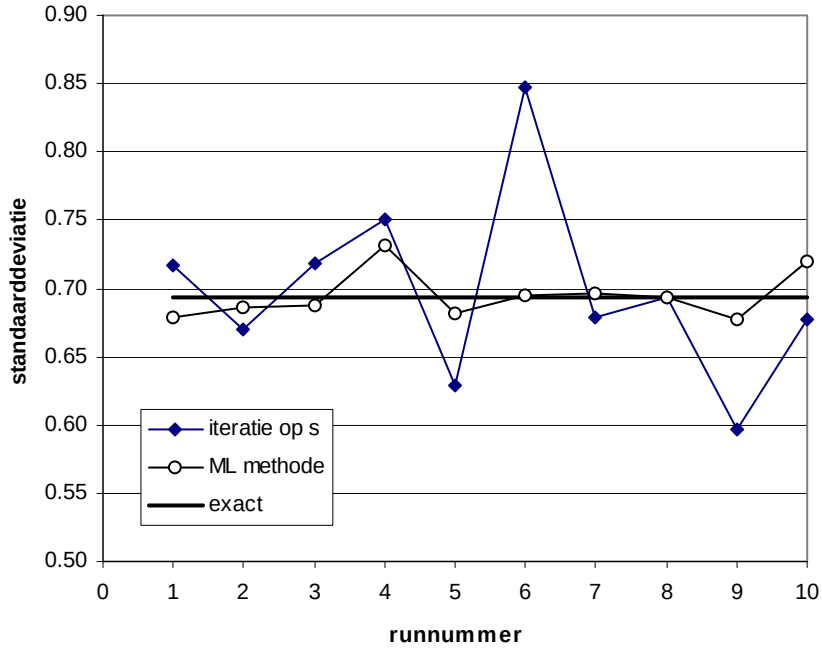
In het maximalisatieproces van (7.37) en (7.41) spelen alleen de termen die van s afhangen een rol. Formule (7.41) heeft als extra s-afhankelijke term ten opzichte van (7.37) de sommatie over $\ln[g_{Y,s}(y_i(s))]$. Dat maakt duidelijk dat het maximalisatieproces in de getransformeerde en de originele ruimte in het algemeen tot verschillende resultaten zal leiden. Overigens blijken de beide ML-methoden voor het voorbeeld uit paragraaf 7.3 (in ieder geval in de beschouwde runs) juist wel dezelfde resultaten op te leveren. Dat komt echter door de wel erg simpele vorm die $\lambda_s(t)$ heeft, namelijk de uniforme verdeling.

runnummer	iteratie op s	ML methode
1	0.717	0.679
2	0.670	0.686
3	0.719	0.687
4	0.751	0.732
5	0.628	0.682
6	0.847	0.695
7	0.679	0.697
8	0.694	0.694
9	0.597	0.678
10	0.677	0.720
gemiddeld	**0.698**	**0.695**
stand. deviatie	0.069	0.018
exacte waarde	**0.693**	**0.693**

Tabel 7.2 Vergelijking van de methode met iteratie op s (waarde na 12 stappen) met de ML-methodes voor 10 runs. In dit voorbeeld geven beide ML-methodes dezelfde resultaten, hetgeen in zijn algemeenheid niet het geval hoeft te zijn.

In tabel 7.2 wordt, voor 10 uitgevoerde runs, de methode met iteratie op s uit paragraaf 7.2 en 7.3 vergeleken met de ML-methode. In figuur 7.14 zijn de resultaten grafisch uitgezet. De eerste run is dezelfde als besproken in paragraaf 7.3. Beide methodes leveren bevredigende resultaten. Het gemiddelde over de 10 runs is in beide gevallen nagenoeg gelijk aan de exacte waarde 0.693 die ten grondslag heeft gelegen aan de simulaties. Wel valt op dat de methode met iteratie op s een grotere spreiding vertoont (circa factor 4 groter) tussen de verschillende runs. Blijkbaar is de ML-methode beter dan de iteratiemethode in staat om de juiste waarde van s uit de data op te sporen. Hierover zometeen meer. In 9 van de 10 runs bleek na 12 stappen (of eerder) convergentie te zijn opgetreden in de zin dat de derde decimaal van s na de 12-stap niet meer wijzigde. Alleen in run 6 bleken 15 stappen nodig voordat convergentie optrad. Niet toevallig levert deze run ook een aanmerkelijk hogere waarde op van s, namelijk 0.847, wat goed zichtbaar is in figuur 7.14. Ter informatie delen we mee dat in 10 extra uitgevoerde runs een dergelijke afwijkende waarde niet meer optrad. Blijkbaar leveren de trekkingen in run 6 puur toevallig een nogal grote spreiding op. Voor run 6 levert de ML-methode wel een goede schatting op van s. In combinatie met de zojuist genoemde kleinere spreiding in de uitkomsten voor de ML-methode lijkt dat laatste er op te wijzen dat de ML-methode in het algemeen de voorkeur verdient boven de methode met iteratie op s. Dat is echter maar zeer de vraag. Ten eerste gaat het hier slechts om één voorbeeld, met gebruik van de nogal simpele uniforme verdeling als conditionele verdeling g(y|x). Ten tweede is het hier beschouwde voorbeeld ‘zeer netjes’ in de zin dat de data na transformatie exact worden beschreven door het gebruikte correlatiemodel, om de simpele reden dat het simulatieproces juist zo in elkaar is gezet dat aan de voorwaarden voor het correlatiemodel is voldaan. Werkelijke data zullen nooit exact door het gebruikte correlatiemodel kunnen worden beschreven

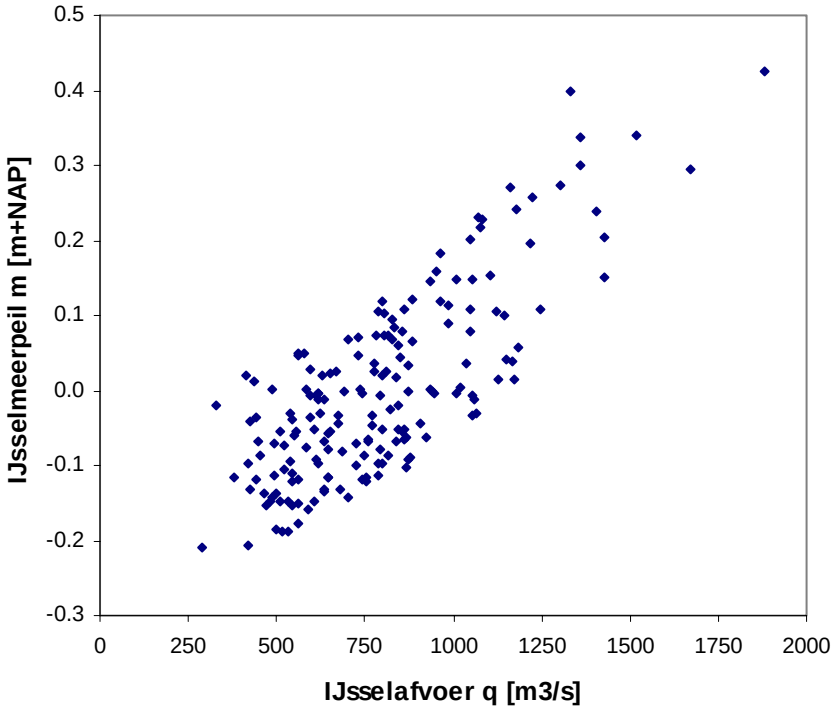
maar slechts in benadering. De eigenlijke vraag is dus hoe de methoden presteren in het geval van werkelijke data.



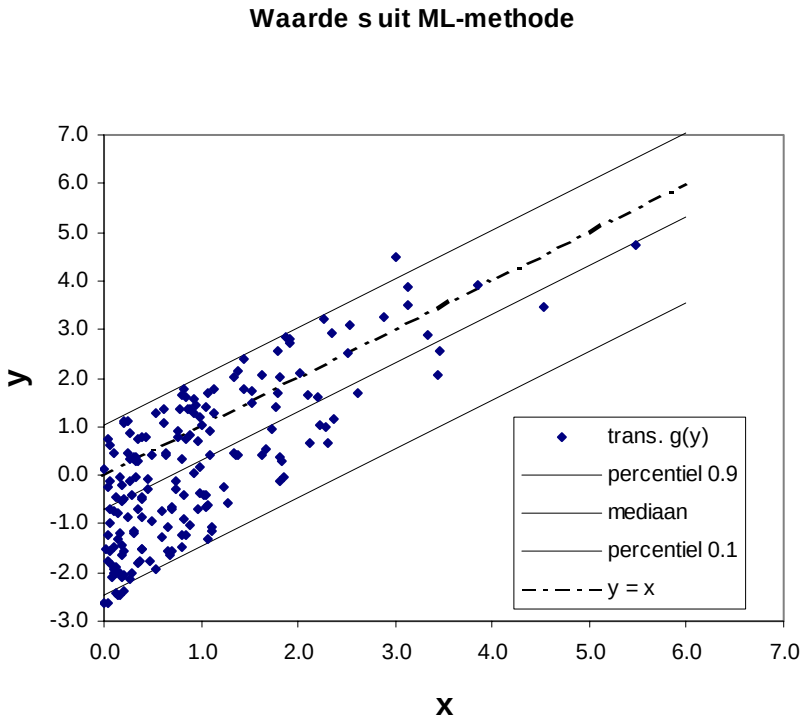
Figuur 7.14 Grafische vergelijking van de methode met iteratie op s met de ML-methodes.

Het is vrij eenvoudig in te zien dat (in ieder geval in dit voorbeeld) de ML-methode niet erg robuust is. Beschouw daartoe figuur 7.15 waarin in run 1 kunstmatig een uitschieter is gecreëerd. Voor de afvoerwaarde 1330 m^3/s is het bijbehorende meerpeil van 0.16 m + NAP verhoogd tot 0.40 m + NAP. Nu hebben de data een uitschieter, die overigens niet al te zeer uit de toon valt bij de rest van de data. Voor de data met uitschieter resulteert nu bij de methode met iteratie op s een standaarddeviatie s = 0.736. Ten opzichte van de oude waarde 0.717 is dit een verhoging van 0.019, wat neerkomt op een verhoging van ongeveer 3 % ten opzichte van de oude waarde. De ML-methodes (die beide in dit voorbeeld weer hetzelfde antwoord geven) komen met uitschieter uit op de waarde s = 1.267, wat ten opzichte van de oude waarde een verhoging van ongeveer 90 % inhoudt. De nieuwe waarde van s leidt in de getransformeerde ruimte tot een onjuiste beschrijving van de data als geheel, hetgeen blijkt uit figuur 7.16. Praktisch alle waarnemingen liggen nu binnen de percentiellijnen, terwijl zich daar ongeveer slechts 80 % van de waarnemingen mag bevinden. Ook is niet langer meer sprake van een constante (verticale) spreiding rond de mediane lijn.

De ML-methode blijkt in dit voorbeeld dus extreem gevoelig voor uitschieters. Gezien de aard van de methode en gezien het feit dat $\lambda_s(t)$, en eveneens $g_s(y|x) = \lambda_s(y\text{-}x\text{-}\delta(s))$, hier de uniforme verdeling volgt is dat zeer verklaarbaar. Iedere waarneming die binnen de drager van $g_s(y|x)$ ligt heeft dezelfde kans van voorkomen, terwijl buiten deze drager de kans abrupt 0 wordt. Zolang voor de beschouwde waarde s > 0 ook maar één van de N waarnemingen $y_i(s)$ buiten de drager van $g_s(y|x_i)$ ligt zal de likelyhood voor deze waarde van s gelijk aan 0 zijn. Pas wanneer s zo groot wordt dat de uitschieter 'binnen het bereik van de uniforme verdeling komt' zal een positieve waarde van de likelyhood resulteren. De ene uitschieter dwingt als het ware een zo grote waarde van s op, dat de uitschieter binnen het bereik van de uniforme verdeling gaat vallen. Het is dus niet verbazend dat de gevonden waarde s = 1.267 zodanig blijkt te zijn dat de uitschieter precies op de 100% - percentiellijn van g(y|x) blijkt te liggen. De hier beschouwde uitschieter ligt bij vrij hoge waarden van q en m. Het is geverifieerd dat iedere uitschieter een soortgelijke uitwerking heeft op de ML-uitkomst, of de uitschieter nou bij hoge of lage waarden van q en m ligt. Tevens heeft elk van deze uitschieters eenzelfde (gering) effect op de uitkomsten voor de methode iteratie op s. Deels zal de gevoeligheid, ofwel het gebrek aan robuustheid, van de ML-methode samenhangen met de eigenschap van de uniforme verdeling dat deze buiten zijn drager abrupt gelijk aan 0 wordt. Ht vermoeden lijkt echter gerechtvaardigd dat de ML-methode in het algemeen erg gevoelig zal zijn voor kleine verstoringen in de data. Naar het zich laat aanzien zal de methode met iteratie op s in praktische toepassingen betere resultaten geven dan de ML-methode. Of dat laatste werkelijk het geval is, is verder niet onderzocht.



Figuur 7.15 De data uit run 1 met een kunstmatige uitschieter (q = 1330, m = 0.40). Vergelijk met figuur 7.5.



Figuur 7.16 De getransformeerde data uit figuur 7.xxx8 met de waarde s = 1.267 uit de ML-methode. Vergelijk met figuur 7.9.

7.5 Discussie van de methoden

Deze paragraaf maakt de balans op van de diverse methoden om in toepassingen een juiste waarde van s te bepalen voor het correlatiemodel. Al eerder werd gezegd dat de ML-methode voor de getransformeerde ruimte in toepassingen niet gebruikt moet worden. Dan zijn er de volgende mogelijkheden:

1. Bepaal s met het iteratierecept uit paragraaf 7.2.

2. Pas de ML-methode toe in de originele ruimte: bepaal s door maximalisatie van (7.38) of (7.41).

3. Beschouw meerdere waarden van s. Beoordeel steeds hoe goed het model de data beschrijft aan de hand van de in paragraaf 7.3 besproken figuren met percentiellijnen. Kies díe waarde van s die op het oog de data het best beschrijft.

Ook voor 1 en 2 dient met figuren te worden beoordeeld of de gevonden waarde van s de data voldoende nauwkeurig beschrijft. Denk aan deze vragen:

- Voor de originele en getransformeerde ruimte: loopt de mediane lijn netjes door het centrum van de data?

- Voor de originele en getransformeerde ruimte: ligt het bij benadering juiste aantal punten buiten de 10%- en 90% percentiellijnen?

- Voor de getransformeerde ruimte: liggen de data als functie van x min of meer gecentreerd rond een rechte lijn met helling 1 en is de spreiding rond deze lijn voor elke x min of meer hetzelfde?

- Voor de originele en getransformeerde ruimte: geeft het model voor met name het meest extreme deel van de data een overtuigende beschrijving?

Wat het laatste punt betreft: het belangrijkste in toepassingen is meestal dat de correlatie voor de meest extreme waarnemingen goed wordt beschreven. Indien dat het geval is kan soms wat water bij de wijn worden gedaan voor de meer frequentere waarnemingen: daarvoor is een wat minder goede beschrijving dan geoorloofd. (Bedenk dat het correlatiemodel voor elke waarde van s in ieder geval de voorgeschreven marginale verdelingen oplevert.) De persoonlijke voorkeur van schrijver dezes is een combinatie van de manieren 1 en 3. Dus eerst s bepalen met het iteratierecept, en vervolgens desgewenst die waarde aanpassen om op het oog een optimale fit te verkrijgen, voor met name het meest extreme deel van de data.

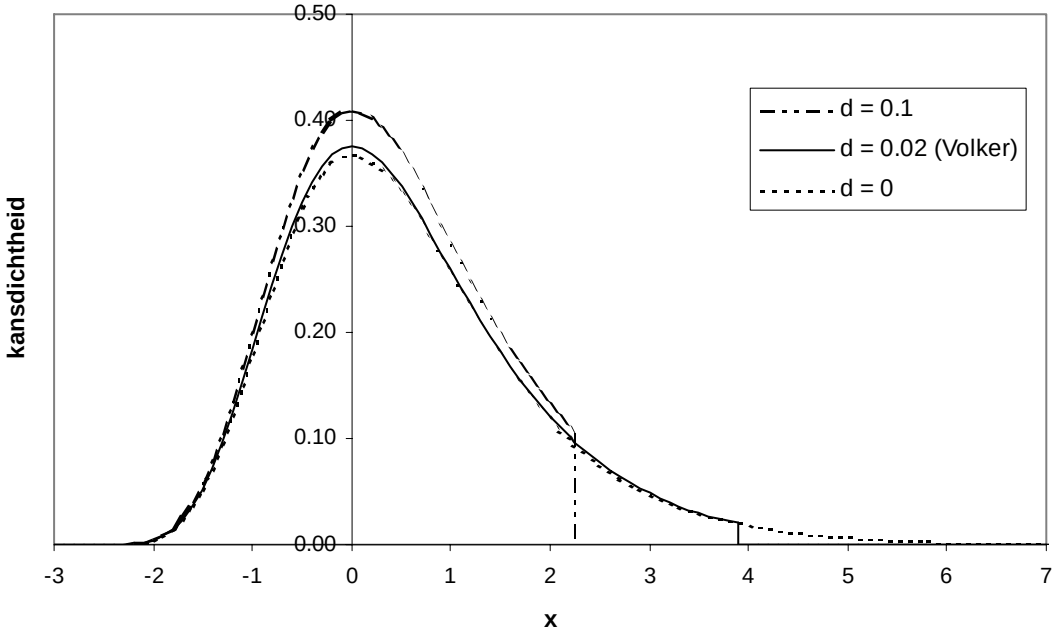
Het kan natuurlijk voorkomen dat voor geen enkele waarde van s een goed correlatiemodel resulteert. Zoals in paragraaf 7.1 al werd opgemerkt is namelijk niet iedere g(q,m) adequaat te beschrijven door het model uit hoofdstuk 3 en 6. Dan dient dus een heel ander model te worden gekozen; zie in dit verband ook hoofdstuk 9.

Tot slot merken we op dat voor zowel het iteratierecept als voor de ML-methode niet is aangetoond dat altijd een waarde van s wordt gevonden. Mogelijk convergeert het iteratierecept niet naar een limietwaarde, mogelijk ook levert de ML-methode niet altijd een oplossing.[8] De indruk van schrijver dezes is dat deze problemen zich alleen voordoen indien het correlatiemodel uit hoofdstuk 3 en 6 zowiezo de data slecht beschrijft – aangetoond is dat echter niet.

[8] Enigszins curieus levert de ML-methode voor de getransformeerde ruimte (die dus niet gebruikt moet worden) wel altijd een eindige, positieve waarde van s als oplossing. Zie stelling A.7.7 uit Appendix A.7. Dat garandeert echter geen goede fit van het model aan de data.

8 Vergelijking met Volkers wind-waterstandstatistiek

In 1987 is door Volker [Volker, 1987] de wind-waterstandstatistiek voor Hoek van Holland afgeleid. Deze statistiek heeft betrekking op een getijperiode en beschrijft de correlatie tussen de windsnelheid U en de zeewaterstand H. Deze statistiek is in het jaar 2002 aangepast aan nieuwe gegevens voor de zeewaterstand, zoals beschreven in [Geerse et al, 2002]. In Volkers oude statistiek werden de kansverdelingen voor de zeewaterstand gegeven door exponentiële verdelingen. In de nieuwe statistiek is dat niet langer het geval. De oude statistiek van Volker vormt een speciaal geval van het in dit stuk behandelde correlatiemodel. In het vervolg wordt toegelicht welke keuzes voor δ, λ(t) en de transformaties moeten worden gemaakt om Volkers formules te krijgen.



Figuur 7.1 Kansdichtheden van (afgeknotte) Gumbelverdelingen. De kansdichtheid met de hoogste top heeft d = 0.1, die met de op een na hoogste top heeft d = 0.02 (Volkers keuze) en de resterende kansdichtheid is de gebruikelijke Gumbelverdeling.

Allereerst volgen wat definities. De zogenaamde afgeknotte Gumbelverdeling wordt gegeven door, voor $0 \le d < 1$,

$$\Lambda_d(x) = \begin{cases} \dfrac{1}{1-d}\exp\{-\exp(-x)\} & , x \le x_d \\ 1 & , x > x_d \end{cases} \tag{8.1}$$

met

$$x_d = -\ln\left(-\ln(1-d)\right) \tag{8.2}$$

Voor d = 0 geeft $\Lambda_0(x)$ de gebruikelijke Gumbelverdeling, terwijl voor d > 0 de verdeling is afgekapt en opnieuw genormeerd voor de hoogste kansbijdrage d. Zie figuur 7.1 voor de kansdichtheden voor de waarden d is 0.1, 0.02 en 0. In Volkers formules is altijd d = 0.02. De kansdichtheid, het gemiddelde en de standaarddeviatie van de afgeknotte verdeling worden gegeven door

$$\lambda_d(x) = \frac{d}{dx}\Lambda_d(x) \tag{8.3}$$

(8.4) $$\mu_d = \int \lambda_d(x)\, x\, dx \cong \mu_0 = 0.577216$$

(8.5) $$\sigma_d = \left[\int (x-\mu_d)^2 \lambda_d(x)\, dx\right]^{1/2} \cong \sigma_0 = \frac{\pi}{\sqrt{6}} = 1.2825$$

De μ_0 en σ_0 geven het gemiddelde en de standaarddeviatie van de niet afgeknotte Gumbelverdeling. Omdat de afknotting weinig effect heeft op het gemiddelde en de standaarddeviatie van de verdeling geldt in benadering (voor d ≤ 2%) dat $\mu_d \cong \mu_0$ en $\sigma_d \cong \sigma_0$. De kansdichtheid voor de zeewaterstand H wordt gegeven door, bij gegeven richting r,

(8.6) $$g(h \mid r) = \frac{1}{B_r} \exp\left(-\frac{h - A_r}{B_r}\right) \quad , h \geq A_r$$

De conditionele verdeling van U, bij gegeven zeewaterstand h en richting r, wordt gegeven door

(8.7) $$F_d(u \mid h, r) = P(U < u \mid h, r) = \Lambda_d\left(\frac{K_r(u) - \rho_r(h - A_r)/B_r}{M_r}\right)$$

Hierin zijn $\rho_r > 0$ en $M_r > 0$, naast A_r en B_r, parameters die voor elk van de richtingen ZW, WZW, …, N gegeven zijn. De functie $K_r(u)$ is voor elk van deze richtingen in tabelvorm gegeven. De kansdichtheid g(u|h,r) volgt door differentiatie van (8.7) naar u. Met (8.6) volgt dan dat g(h,u|r) wordt gegeven door

(8.8) $$g(h,u \mid r) = \frac{1}{B_r} e^{-\frac{h-A_r}{B_r}} \lambda_d\left(\frac{K_r(u) - \rho_r(h - A_r)/B_r}{M_r}\right) \frac{K_r'(u)}{M_r}$$

Deze formule geeft dus voor elke richting ZW t/m N de gezamenlijke kansdichtheid van de stochasten H en U, bij gegeven richting r. Deze kansdichtheid blijkt voor elke beschouwde richting een speciaal geval te zijn van het in hoofdstuk 3 en paragraaf 6.1 beschouwde correlatiemodel. Het is direct te verifiëren dat (8.8) kan worden herschreven als

(8.9) $$g(h,u \mid r) = e^{-J(h)} J'(h) K'(u)\, \lambda\left(K(u) - J(h) - \delta\right)$$

met

(8.10)
$$\lambda(t) = \frac{\rho_r}{M_r} \lambda_d\left(\frac{\rho_r t}{M_r} + \mu_d\right)$$
$$K(u) = \frac{K_r(u)}{\rho_r}$$
$$J(h) = \frac{h - A_r}{B_r}$$
$$\delta = \frac{M_r \mu_d}{\rho_r}$$

De kansdichtheid λ(t) heeft gemiddelde 0 en standaarddeviatie $s = M_r\sigma_d/\rho_r \cong 1.28\, M_r/\rho_r > 0$ en voldoet daarmee aan de in hoofdstuk 3 en 4 gestelde eisen aan λ(t). De verificatie van bovenstaande vergt enig schrijfwerk en wordt aan de lezer overgelaten. Formule (8.9) is van de vorm (6.8) indien in plaats van de stochasten Q en M respectievelijk H en U worden beschouwd. Om te rechtvaardigen dat (8.9) een speciaal geval is van het correlatiemodel uit hoofdstuk 4 dient nog wel te worden geverifieerd dat de functies J(h) en K(u) als transformaties van de vorm (6.2) mogen worden opgevat. Voor J(h) volgt dat eenvoudig uit (3.1) en (6.2). Met betrekking tot K(u) volgt uit (8.9), met Λ(t) de cumulatieve verdelingsfunctie van λ(t),

$$
\begin{aligned}
F_U(u \mid r) &= \int_{-\infty}^{u} \left\{ \int g(h,w \mid r) dh \right\} dw \\
&= \int \left\{ \int_{-\infty}^{u} g(h,w \mid r) dw \right\} dh \\
&= \int \Lambda\left(K(u) - J(h) - \delta\right) e^{-J(h)} J'(h)\, dh \\
&= \int_{0}^{\infty} \Lambda\left(K(u) - x - \delta\right) e^{-x} dx
\end{aligned}
\tag{8.11}
$$

Uit (3.14) volgt dan

$$F_U(u \mid r) = F_Y(K(u)) \tag{8.12}$$

waarmee is aangetoond dat y = K(u) de transformatie geeft van U naar Y. Hiermee is aangetoond dat indien voor het correlatiemodel uit hoofdstuk 3 en 6 de keuzes uit (8.10) worden gemaakt de resulterende kansdichtheid g(h,u|r) gelijk is aan de kansdichtheid volgens de oude statistiek van Volker.

In de nieuwe wind-waterstandstatistiek zijn de verdelingen g(h|r) niet langer exponentieel. Het bleek echter toch mogelijk van Volkers oude formules gebruik te maken. Daarbij werden de parameters A_r en B_r die in (8.7) voorkomen gevonden door de nieuwe g(h|r) te benaderen door exponentiële verdelingen. Overigens diende deze benadering slechts om de genoemde parameters te bepalen; de nieuwe statistiek is wel zo opgesteld dat de marginale verdeling g(h|r) exact de voorgeschreven vorm behoudt. Zie voor verdere details [Geerse et al, 2002]. Achteraf kan worden gesteld dat een plausibeler aanpak voor het bepalen van de nieuwe kansdichtheid g(h,u|r) de in dit stuk beschreven methode zou zijn geweest. Daarbij zou de nieuwe g(h|r) dus zijn getransformeerd naar de standaardexponentiële verdeling. Op het moment dat de nieuwe g(h,u|r) werd afgeleid was dat inzicht echter nog niet voorhanden.

9 Een algemeen bivariaat correlatiemodel

Dit hoofdstuk staat min of meer los van de eerdere hoofdstukken. Het gaat over een zeer algemeen (bivariaat) correlatiemodel. De theorie daarvoor is niet nieuw. Bijvoorbeeld in [Ditlevsen en Madsen, 1996] kan het meeste van het hieronder wordt beschreven gevonden worden.

Paragraaf 9.1 geeft de formules voor het model. Daarin wordt net als in de voorgaande hoofdstukken gewerkt met de transformatie van stochasten Q en M naar stochasten X en Y. Er is dus weer sprake van een 'originele' en van een 'getransformeerde' ruimte. De X en Y hebben (tenzij bijzondere gevallen worden beschouwd) geen verband meer met exponentiele verdelingen. In zekere zin kan het model uit paragraaf 9.1 worden gezien als een manier om uitgaande van een zeer algemene kansdichtheid in de getransformeerde ruimte te komen tot een kansdichtheid in de originele ruimte met voorgeschreven kansdichtheden g(q) en g(m). De formules van het model maken daarnaast duidelijk (voor zover dat de lezer nog niet bekend was) dat er 'oneindig' veel bivariate kansdichtheden g(q,m) bestaan – die zeer verschillend van karakter kunnen zijn – met de voorgeschreven marginale kansdichtheden g(q) en g(m). Paragraaf 9.2 geeft een speciaal geval van het model: de bivariate verdeling in de getransformeerde ruimte is dan gelijk aan de bivariate standaardnormale verdeling met correlatiecoëfficiënt ρ.

Het in de voorgaande hoofdstukken beschreven correlatiemodel vormt een speciaal geval van het algemene model uit dit hoofdstuk. In eerste instantie lijkt voor toepassingen een zeer algemeen model het meest wenselijk. Dat lijkt maar zo. Bij een eenvoudig model, zoals dat uit de eerdere hoofdstukken, en ook dat uit paragraaf 9.2, is in toepassingen te verifiëren of het model al of niet geschikt is de data te beschrijven. Bij een ingewikkelder model wordt die verificatie veel lastiger.

9.1 Formules voor een algemeen bivariaat correlatiemodel

Als in hoofdstuk 6 gaan we uit van stochasten Q en M met gegeven kansdichtheden $g_Q(q)$ en $g_M(m)$. Na transformatie gaan Q en M over in stochasten X en Y. Het vlak van punten (q,m) zal opnieuw worden aangeduid als de 'orginele ruimte' en het vlak van punten (x,y) als de 'getransformeerde ruimte'. De gezamelijke kansdichtheid van X en Y mag zeer algemeen zijn. Er wordt slechts aangenomen dat deze afhangt van n parameters $a_1, a_2, \ldots, a_n$, die kortweg veelal worden aangeduid als **a**. We schrijven dus

(9.1) $$g_{X,Y}(x,y,a_1,a_2,...,a_n) = g_{X,Y}(x,y,\mathbf{a})$$

Uit deze kansdichtheid kunnen op de gebruikelijke manier de marginale kansdichtheden $g_X(x,\mathbf{a})$ en $g_Y(y,\mathbf{a})$ worden bepaald, evenals de cumulatieve verdelingsfuncties $F_X(x,\mathbf{a})$ en $F_Y(y,\mathbf{a})$. Analoog aan hoofdstuk 6 beschouwen we transformaties

(9.2) $$x = J(q,\mathbf{a})$$

(9.3) $$y = K(m,\mathbf{a})$$

met $J(q,\mathbf{a})$ en $K(q,\mathbf{a})$ vastgelegd door

(9.4) $$F_X\left(J(q,\mathbf{a}),\mathbf{a}\right) = F_Q(q)$$

(9.5) $$F_Y\left(K(m,\mathbf{a}),\mathbf{a}\right) = F_M(m)$$

Door (9.4) naar q te differentiëren volgt

(9.6) $$\frac{\partial J(q,\mathbf{a})}{\partial q} = \frac{g_Q(q)}{g_X\left(J(q,\mathbf{a}),\mathbf{a}\right)}$$

Evenzo volgt uit (9.5)

(9.7) $$\frac{\partial K(m,\mathbf{a})}{\partial m} = \frac{g_M(m)}{g_Y\left(K(m,\mathbf{a}),\mathbf{a}\right)}$$

Wanneer $g_{X,Y}(x,y,\mathbf{a})$ gegeven is, leggen de transformaties (9.2) en (9.3) de kansdichtheid in de originele ruimte vast. De formule om deze te bepalen is standaard, zie bijvoorbeeld Appendix A.6.2. Er volgt met behulp van formule (6.10) uit de appendix en met de zojuist gegeven (9.6) en (9.7)

(9.8)
$$\begin{aligned} g_{Q,M}(q,m,\mathbf{a}) &= g_{X,Y}\left(J(q,\mathbf{a}),K(m,\mathbf{a}),\mathbf{a}\right)\left|\frac{\partial J(q,\mathbf{a})}{\partial q}\frac{\partial K(m,\mathbf{a})}{\partial m}\right| \\ &= \frac{g_{X,Y}\left(J(q,\mathbf{a}),K(m,\mathbf{a}),\mathbf{a}\right)g_Q(q)g_M(m)}{g_X\left(J(q,\mathbf{a}),\mathbf{a}\right)g_Y\left(K(m,\mathbf{a}),\mathbf{a}\right)} \\ &= \text{gezamelijke kansdichtheid van Q en M in de orginele ruimte} \end{aligned}$$

Merk op dat indien X en Y onafhankelijk zijn, dus indien $g_{X,Y}(x,y,\mathbf{a}) = g_X(x,\mathbf{a})\, g_Y(y,\mathbf{a})$, volgt dat $g_{Q,M}(q,m,\mathbf{a}) = g_Q(q)\, g_M(m)$. In dat geval zijn Q en M dus eveneens onafhankelijk. Omgekeerd volgt (behalve eventueel in irreguliere situaties) dat onafhankelijkheid van Q en M onafhankelijkheid van X en Y impliceert. Analoog aan formule (6.16) uit Appendix A.6.2 kan betrekkelijk eenvoudig worden geverifieerd dat $g_{Q,M}(q,m,\mathbf{a})$ de voorgeschreven kansdichtheden $g_Q(q)$ en $g_M(m)$ heeft.

Formule (9.8) vormt een veralgemenisering van het in hoofdstuk 6 en 7 beschouwde correlatiemodel. Door $g_{X,Y}(x,y)$ te kiezen als in (6.6) volgt namelijk, na enige herschrijving, dat de $g_{Q,M}(q,m)$ volgens (6.8) en volgens (9.8) overeenstemmen. Omdat $g_{X,Y}(x,y)$ volgens (9.1) van willekeurig veel parameters $a_1, a_2, \ldots, a_n$ mag afhangen, volgt dus dat het in dit hoofdstuk beschouwde model een meer algemene vorm heeft. We gaan hier niet in op een recept om bij een keuze van $g_{X,Y}(x,y,\mathbf{a})$ op grond van een dataset de parameters $a_1, a_2, \ldots, a_n$ te bepalen. We merken slechts op dat een voor de hand liggende manier de methode van Maximum Likelyhood is.

9.2 De bivariate standaardnormale verdeling als correlatiemodel

In deze paragraaf behandelen we een simpele keuze voor $g_{X,Y}(x,y,\mathbf{a})$, namelijk de bivariate normale verdeling. Dat model is indertijd gebruikt in een probabilistisch model voor de IJsseldelta, beschreven in [Geerse, 2003a], dat in de jaren 1999 en 2000 is geïmplementeerd in een computerprogramma door het bureau HKV. Het bivariate normale correlatiemodel, dat werd gebruikt om de correlatie tussen de IJsselafvoer en het IJsselmeerpeil te beschrijven, bleek daarbij goed werkbaar. De formules voor het correlatiemodel en de vergelijking tussen het model en de data zijn overigens nooit gerapporteerd. Dat werd niet nodig geacht, omdat het probabilistisch model nooit een officiële status heeft gekregen – de Sobeksommen die als invoer voor het model moesten dienen waren namelijk niet op tijd beschikbaar. Inmiddels is het probabilistisch model vervangen door een nieuw probabilistisch model, dat behalve voor de IJsseldelta ook geschikt is voor de Vechtdelta (het oude model was dat niet). Dat laatste model wordt beschreven in [Geerse, 2003b]; het wordt op dit moment geïmplementeerd in een computerprogramma. Het nieuwe model maakt gebruik van het correlatiemodel uit hoofdstuk 3 en 6 van dit rapport, met als keuze voor $\lambda(t)$ de normale verdeling.

We nemen als kansverdeling in de getransformeerde ruimte de bivariate, standaardnormale verdeling. Deze heeft slechts één parameter, namelijk de correlatiecoëfficient ρ. In plaats van $g_{X,Y}(x,y,\rho)$ zullen we in het vervolg gemakshalve $g_\rho(x,y)$ schrijven. Deze heeft de vorm, zie bijvoorbeeld [Casella, 1990],

(9.9) $$g_\rho(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right), \qquad -1<\rho<1$$

We zullen de eendimensionale normale verdeling met gemiddelde μ en standaarddeviatie σ aangeven met $N(\mu,\sigma)$; een stochast Z met verdeling $N(\mu,\sigma)$ heeft dan als dichtheidsfunctie

(9.10) $$g(z) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{z^2}{2\sigma^2}\right\}$$

Het is een bekend feit, zie bijvoorbeeld [Casella, 1990] dat de marginale verdellingen van (9.9) standaardnormaal zijn verdeeld, dus dat geldt

(9.11) $g(x)$ heeft verdeling $N(\mu = 0,\ \sigma = 1)$

(9.12) $g(y)$ heeft verdeling $N(\mu = 0,\ \sigma = 1)$

Verder zijn de conditionele verdelingen van (9.9) eveneens normaal verdeeld, en geldt

(9.13) $g_\rho(x \mid y)$ heeft verdeling $N(\mu = \rho y,\ \sigma = \sqrt{1-\rho^2})$

(9.14) $g_\rho(y \mid x)$ heeft verdeling $N(\mu = \rho x,\ \sigma = \sqrt{1-\rho^2})$

Merk op dat de marginale verdelingen niet van ρ afhangen, terwijl dat voor de conditionele verdelingen wel het geval is. We geven de cumulatieve verdelingsfunctie van de standaardnormale verdeling aan met Φ, dus

(9.15) $$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz$$

De transformaties (9.2) en (9.3) worden dan volgens (9.4) en (9.5) vastgelegd door

(9.16) $\Phi\big(J(q)\big) = F_Q(q)$

(9.17) $\Phi\big(K(m)\big) = F_M(m)$

Volgens (9.8) en (9.9) wordt de kansdichtheid in de orginele ruimte dan gegeven door

(9.18) $$g_{Q,M}(q,m,\rho) = \frac{\exp\left(-\dfrac{J(q)^2 - 2\rho J(q)K(m) + K(m)^2}{2(1-\rho^2)}\right) g_Q(q) g_M(m)}{\sqrt{1-\rho^2}\exp\left(-\dfrac{J(q)^2}{2} - \dfrac{K(m)^2}{2}\right)}$$

Merk op dat de transformaties x = J(q) en y = K(m) niet van ρ afhangen. Deze transformaties kunnen worden uitgevoerd zonder dat de waarde van ρ bekend is. Daarmee is het hier beschouwde model iets eenvoudiger te hanteren dan het correlatiemodel uit hoofdstuk 6 en 7. Voor dat laatste model was een waarde van de spreiding s nodig alvorens K(m) kon worden bepaald. Een voordeel van het model uit hoofdstuk 6 en 7 is dat als conditionele verdeling g(y|x) meerdere kansdichtheden kunnen worden beschouwd, terwijl voor het model uit deze paragraaf g(y|x) altijd een normale verdeling volgt.

Appendix

A.1 Inleiding

Deze appendix geeft allerlei bewijzen die aanvullend zijn op de beweringen in de hoofdstukken 3 tot en met 7. De behandeling wordt beknopt gehouden en is bedoeld voor wiskundig goed onderlegde lezers.

A.2 Stellingen uit analyse en maattheorie

Verderop in de appendix worden enkele bekende stellingen uit de wiskunde gebruikt die voor de duidelijkheid in deze paragraaf worden geformuleerd. We geven eerst enkele betrekkelijk elementaire stellingen uit de analyse. Daarna volgen enkele stellingen uit de maattheorie, waarbij vooral Lebesgue's gedomineerde convergentiestelling van belang is; tevens wordt een versie van Jensen's ongelijkheid gegeven die naast het gebruikelijke geval van 'gewone' convexiteit ook het geval van strikte convexiteit behandelt. De stellingen over maattheorie zijn afkomstig uit [Billingsley, 1995], waarnaar we verwijzen voor verdere details.

Stelling A.2.1 Versie van de Regel van L'Hôpital. [Apostol, 1967; theorem 7.10]. Beschouw twee reële functies f(x) en k(x) en neem aan dat beide functies differentieerbaar zijn voor alle x groter dan een zekere vaste M > 0. Veronderstel dat

$$\begin{aligned} &\lim_{x\to\infty} f(x) = 0 \\ &\lim_{x\to\infty} k(x) = 0 \end{aligned} \tag{2.1}$$

en dat k'(x) $\neq$ 0 voor x > M. Dan geldt, mits de limiet in het rechterlid bestaat,

$$\lim_{x\to\infty} \frac{f(x)}{k(x)} = \lim_{x\to\infty} \frac{f'(x)}{k'(x)} \tag{2.2}$$

Stelling A.2.2 Afgeleide van inverse functie. [Apostol, 1967; theorem 6.7]. Neem aan dat f(x) strikt stijgend en continu is op een interval [a,b] en laat g(y) de inverse zijn van f(x). Als voor een punt x in (a,b) de afgeleide f'(x) bestaat en ongelijk aan nul is, dan bestaat in het met x corresponderende punt y = f(x) de afgeleide g'(y) eveneens welke dan ongelijk aan nul is. Bovendien geldt

$$g'(y) = \frac{1}{f'(x)} \tag{2.3}$$

N.B. Indien f'(x) continu is in x, dan is g'(y) continu in y.

Stelling A.2.3 Beknopte versie impliciete functiestelling. [Marsden en Tromba, 1996; p 226]. Beschouw een functie $H: I_1 \times I_2 \to \mathbb{R}$ met I_1 en I_2 (eventueel onbegrensde) open intervallen in $\mathbb{R}$ die continue partiële afgeleiden heeft. Neem aan dat voor $(s_0, y_0) \in I_1 \times I_2$ geldt

$$\begin{aligned} &H(s_0, y_0) = 0 \\ &\frac{\partial H(s_0, y_0)}{\partial y_0} \neq 0 \end{aligned} \tag{2.4}$$

Dan bestaat er een δ_1-omgeving U van s_0 die bevat is in I_1 en een δ_2-omgeving V van y_0 die bevat is in I_2 zodanig dat er een unieke functie $g: U \to V$ bestaat waarvoor H(s, g(s)) = 0. De functie g(s) is continu differentieerbaar, met afgeleide

(2.5) $$g'(s) = -\left.\frac{\partial H(s,y)/\partial s}{\partial H(s,y)/\partial y}\right|_{y=g(s)}$$

N.B. Merk op dat uniciteit van g(s) impliceert dat indien voor $s \in U$ en $y \in V$ geldt H(s,y) = 0, dat dan y = g(s).

Nu volgen enkele stellingen uit de maattheorie. Waar in het vervolg integratie wordt uitgevoerd is dat altijd in de zin van Lebesgue (en niet van Riemann). Met het begrip maat zal altijd de Lebesguemaat worden bedoeld terwijl met meetbaarheid Lebesguemeetbaarheid wordt bedoeld. Waar geïntegreerd wordt over functies, zoals bijvoorbeeld g(t), f(t), $f_n(t)$, wordt stilzwijgend aangenomen dat het altijd reële Lebesgue-meetbare functies betreft die gedefinieerd zijn voor alle $t \in \mathbb{R}$. Af en toe wordt echter expliciet de toevoeging meetbaar vermeld.

We brengen in herinnering dat een functie f(t) *integreerbaar* is dan en slechts dan als de integraal over |f(t)| eindig is, ofwel indien

(2.6) $$\int |f(t)|\,dt < \infty$$

in welk geval tevens geldt dat de integraal over f(t) eindig is. Merk op dat deze eigenschap niet geldt voor integratie in de zin van Riemann. Verder wordt het begrip *bijna overal* in herinnering gebracht aan de hand van het volgende voorbeeld. Beschouw een rij functies $f_n(t)$, n = 1, 2, 3, ... en een functie f(t) voor reële t. Stel dat voor sommige t geldt

(2.7) $$\lim_{t\to\infty} f_n(t) = f(t)$$

Indien de t-verzameling waarvoor deze limiet bestaat meetbaar is en indien het complement van deze verzameling maat 0 heeft, zegt men dat de bewering (2.7) *bijna overal* geldt. Tevens zegt men dat de bewering bestaat voor *bijna alle t*. Terzijde merken we op dat, zie Theorem 13.4 en Example 16.8 uit [Billingsley, 1995], de t-verzameling waarvoor (2.7) geldt altijd een meetbare verzameling vormt. Indien echter in plaats van een discrete index n een continue indexparameter s wordt gebruikt hoeft de t-verzameling waarvoor (2.7) geldt niet altijd een meetbare verzameling te vormen. Als bijvoorbeeld voor een functie f(t) en een familie $f_s(t)$, s > 0, van functies geldt dat $f_s(t) \to f(t)$ voor $s \to \infty$, dan kan het voorkomen dat de t-verzameling waarvoor deze limiet bestaat géén meetbare verzameling vormt. Dit feit speelt een rol in de formulering van onderstaande stellingen. Daarom definiëren we hier het begrip *volledige maat* als volgt: een (Lebesgue)meetbare verzameling A heeft volledige maat indien het complement van deze verzameling in $\mathbb{R}$, aangeduid als $\mathbb{R} \setminus A$, maat 0 heeft. In de situatie dat in onderstaande stellingen sprake is van een familie van functies met continue indexparameter zal de verzameling A de rol spelen die het begrip 'bijna overal' vervult voor de rij van functies met discrete indexparameter. Nu volgen twee versies voor Lebesgue's gedomineerde convergentiestelling, verder kortweg aangeduid als 'LDC-stelling', de eerste voor een discrete indexverzameling en de tweede voor een continue indexverzameling. In het vervolg van de appendix wordt het aan de lezer overgelaten welke versie van de LDC-stelling van toepassing is omdat dat uit de context altijd voldoende duidelijk zal zijn.

Stelling A.2.4 Lebesgue's gedomineerde convergentiestelling voor discrete indexverzameling (LDC-stelling) [Billingsley, 1995; Theorem16.4]. Stel dat voor een rij functies $f_n(t)$, n = 1, 2, 3, ... en voor een integreerbare g(t) geldt

(2.8) $$|f_n(t)| \le g(t)$$ bijna overal, voor $n = 0, 1, 2, \ldots$

en dat tevens voor een functie f(t) geldt

(2.9) $$\lim_{n\to\infty} f_n(t) = f(t)$$ bijna overal

Dan zijn $f_n(t)$ en f(t) integreerbaar en geldt

(2.10) $$\lim_{n\to\infty} \int f_n(t)\,dt = \int f(t)\,dt$$

Stelling A.2.5 Lebesgue's gedomineerde convergentiestelling voor continue indexverzameling (LDC-stelling) [Billingsley, 1995; pagina 211]. Stel dat voor een familie van functies $f_s(t)$, met s uit een continue indexverzameling, en voor een integreerbare g(t) geldt, voor een verzameling A met volledige maat,

(2.11) $\left|f_s(t)\right| \le g(t) \qquad t \in A$, alle s

en dat tevens voor een functie f(t) geldt, voor $s_0 \in \mathbb{R}$,

(2.12) $\lim_{s \to s_0} f_s(t) = f(t) \qquad t \in A$

Dan zijn $f_s(t)$ en f(t) integreerbaar en geldt

(2.13) $\lim_{s \to s_0} \int f_s(t)\,dt = \int f(t)\,dt$

Voor de limiet in (2.12) en (2.13) mag eveneens de linker- of de rechterlimiet worden beschouwd. Tevens mag voor s_0 ook $-\infty$ of ∞ worden genomen.

Nu volgt een stelling die betrekking heeft op de wijze waarop continuïteit en differentieerbaarheid onder het integraalteken zich vertalen in continuïteit en differentieerbaarheid van de integraal.

Stelling A.2.6 Differentiëren en continuïteit onder het integraalteken [Billingsley, 1995; Theorem 16.8]. Beschouw een reële functie f(t,z) die integreerbaar is in t voor iedere z in (a,b) en laat φ(z) gegeven zijn door

(2.14) $\varphi(z) = \int f(t,z)\,dt$

Dan geldt het volgende.

1. Stel dat voor een verzameling A van volledige maat voor iedere $t \in A$ geldt dat f(t,z) continu is in $z_0 \in$ (a,b). Neem tevens aan dat voor een integreerbare functie g(t) geldt, voor zekere $\delta(z_0) > 0$,

(2.15) $\left|f(t,z)\right| \le g(t)$ voor $t \in A$ en alle z waarvoor $\left|z - z_0\right| < \delta(z_0)$

   waarbij $\delta(z_0)$ onafhankelijk is van t en dusdanig dat het interval $(z_0 - \delta(z_0), z_0 + \delta(z_0))$ bevat is in (a,b). Dan is φ(z) continu in z_0.

   In plaats van het open interval (a,b) in bovenstaande mag ook het gesloten interval [a,b] worden genomen. Indien $z_0 = a$ dient de voorwaarde $|z - z_0| < \delta(z_0)$ dan vervangen te worden door $0 \le z - z_0 < \delta(z_0)$; indien $z_0 = b$ dient de voorwaarde $|z - z_0| < \delta(z_0)$ dan vervangen te worden door $0 \le z_0 - z < \delta(z_0)$.

2. Stel dat voor een verzameling A van volledige maat voor iedere $t \in A$ geldt dat de afgeleide $\partial f(t,z_0)/\partial z_0$ bestaat voor een beschouwde z_0 in (a,b). Neem verder aan dat er een integreerbare $g(t,z_0)$ bestaat waarvoor, voor zekere $\delta(z_0) > 0$,

(2.16) $\left|\frac{\partial f(t,z)}{\partial z}\right| \le g(t,z_0)$ voor $t \in A$ en alle z waarvoor $\left|z - z_0\right| < \delta(z_0)$

   waarbij $\delta(z_0)$ onafhankelijk is van t en dusdanig dat het interval $(z_0 - \delta(z_0), z_0 + \delta(z_0))$ bevat is in (a,b). Dan wordt de afgeleide van φ(z) in z_0 gegeven door

(2.17) $\varphi'(z_0) = \int \frac{\partial f(t,z_0)}{\partial z_0}\,dt$

Tot slot van deze paragraaf wordt Jensens ongelijkheid behandeld. Die ongelijkheid wordt gewoonlijk geformuleerd voor willekeurige maten. In het vervolg wordt de iets beperktere situatie beschouwd waarin de

maat kan worden gerepresenteerd door een kansdichtheid. In dat geval kan in Jensens ongelijkheid *strikte* ongelijkheid worden bewezen in het geval dat de in de ongelijkheid beschouwde functie *strikt* convex is in plaats van gewoon convex.

Een reële functie f(t) op een open interval I (begrensd of onbegrensd), wordt *convex* genoemd wanneer

(2.18) $$f(\beta t_1 + (1-\beta)t_2) \le \beta f(t_1) + (1-\beta) f(t_2)$$

voor alle t_1 en t_2 in I met $t_1 \neq t_2$ en $0 < \beta < 1$. De functie wordt *strikt convex* genoemd indien in (2.18) strikte ongelijkheid geldt.

Stelling A.2.7 Jensens ongelijkheid.
Beschouw een convexe functie f(t) op $\mathbb{R}$ en een willekeurige kansdichtheid $\lambda(t)$, dat wil zeggen een niet-negatieve Lebesgue-integreerbare functie met integraal gelijk aan 1. Neem aan dat met betrekking tot $\lambda(t)$ de verwachtingswaarden E(T) en E(f(T)) beiden bestaan als eindige reële getallen. Dan geldt

(2.19) $$f(E(T)) \le E(f(T))$$

Bovendien geldt ingeval f(t) strikt convex is

(2.20) $$f(E(T)) < E(f(T))$$

Het geval (2.19) is overbekend, zie bijvoorbeeld [Billingsley, 1995; pagina 276 en Appendix A33]. Op soortgelijke wijze als in de genoemde referentie zal nu (2.20) worden bewezen. Beschouw daartoe $\lambda(t)$ en eindige verwachtingswaarden E(T) en E(f(T)) als beschouwd in (2.20) en neem aan dat f(t) strikt convex is. Volgens [Billingsley, 1995; pagina 276 en Appendix A33] is er dan een raaklijn k(t) = a + bt, dat wil zeggen een lijn die volledig onder de grafiek van f(t) ligt of hooguit daaraan raakt, die gaat door het punt $(t_0, f(t_0))$ met $t_0 =$ E(T), waarvoor

(2.21) $$k(t) = a + bt \le f(t) \qquad \text{alle} \, t$$

We zullen laten zien dat vanwege strikte ongelijkheid tevens geldt

(2.22) $$k(t) = a + bt < f(t) \qquad \text{alle } t \neq t_0$$

Neem aan dat (2.22) onjuist is. Dan is er $t_1 \neq t_0$ waarvoor $k(t_1) = f(t_1)$. Neem vooreerst aan dat $t_1 > t_0$. Uit de convexiteit van f(t) volgt dan, voor $0 < \beta < 1$,

(2.23)
$$\begin{aligned} f(\beta t_0 + (1-\beta)t_1) &\le \beta f(t_0) + (1-\beta) f(t_1) \\ &= \beta k(t_0) + (1-\beta) k(t_1) \\ &= k\left(\beta t_0 + (1-\beta)t_1\right) \end{aligned}$$

wat vanwege (2.21) impliceert dat f(t) op het interval (t_0, t_1) gelijk is aan k(t). Een functie met een lineair deel in zijn grafiek is echter niet strikt convex, zodat een dergelijke $t_1 > t_0$ niet kan bestaan. Voor $t_1 < t_0$ volgt uiteraard dezelfde conclusie door het interval (t_1, t_0) te beschouwen. Bewering (2.22) moet dus juist zijn. Op grond van deze bewering moet dan gelden, zie desgewenst [Billingsley, 1995; Theorem 15.2],

(2.24) $$\int \left(f(t) - a - bt \right) \lambda(t)\, dt > 0$$

en derhalve

(2.25) $$E(f(T)) > a + b\,E(T) = f(E(T))$$

waarmee (2.20) is bewezen.

A.3 Details hoofdstuk 3

In het volgende wordt de equivalentie van de voorwaarden (3.9) en (3.17) uit hoofdstuk 3 bewezen. In het bewijs wordt gebruik gemaakt van een versie van de Regel van L'Hôpital die werd gegeven in stelling A.2.1. Neem aan dat voldaan is aan (3.9) uit hoofdstuk 3, dus dat geldt

(3.1) $$E(e^t) = \int e^t \lambda(t)\, dt < \infty$$

We zullen aantonen dat dan (3.17) geldt uit hoofdstuk 3, dus dat dab voldaan is aan

(3.2) $$\int e^t \overline{\Lambda}(t)\, dt < \infty$$

Uit (3.1) volgt dat $e^t\lambda(t)$ naar 0 gaat in de limiet $t \to \infty$. De Regel van L' Hôpital levert dan, omdat vanaf zekere vaste waarde $M > 0$ de functie $\overline{\Lambda}(t)$ differentieerbaar is,

(3.3) $$\lim_{t\to\infty} \overline{\Lambda}(t)e^t = \lim_{t\to\infty} \frac{\overline{\Lambda}(t)}{e^{-t}} = \lim_{t\to\infty} \frac{\lambda(t)}{e^{-t}} = \lim_{t\to\infty} \lambda(t)e^t = 0$$

We brengen in herinnering dat bij aanname $\lambda(t)$ hooguit eindig veel discontinuïteitspunten heeft. Neem vooreerst aan dat $\lambda(t)$ geen enkel discontinuïteitspunt heeft. Met partiële integratie volgt dan

(3.4) $$\int_{-\infty}^{y} e^t \overline{\Lambda}(t)\, dt = \left[e^t \overline{\Lambda}(t) \right]_{-\infty}^{y} + \int_{-\infty}^{y} e^t \lambda(t)\, dt$$

Omdat de stokterm vanwege (3.3) gelijk aan 0 wordt in de limiet $y \to \infty$, volgt vanwege (3.1) dat dan voldaan is aan (3.2). Neem nu aan dat (3.2) geldt. Dan volgt onmiddelijk dat $\overline{\Lambda}(t)e^t$ tot 0 nadert als $t \to \infty$. Uit (3.4) volgt dan dat tevens is voldaan aan (3.1). Hiermee is de equivalentie van (3.1) en (3.2) bewezen. Merk op dat terloops is bewezen dat de integraal in (3.2) dan gelijk is aan $E(e^T)$. Indien $\lambda(t)$ eindig veel discontinuïteitspunten heeft kan de integratie in (3.4) worden opgesplitst in eindige veel trajecten waarop $\lambda(t)$ continu is. Het is eenvoudig te verifiëren dat formule (3.4) dan blijft gelden.

A.4 Details hoofdstuk 4

Lemma A.4.1 Neem aan dat $E(Te^T) < \infty$. Dan geldt

(4.1) $$\mu_\gamma = -e^{\delta_0} \int \lambda(t) t e^t\, dt < 0$$

Bewijs Er dient te worden bewezen dat

(4.2) $$\int \lambda(t) t e^t\, dt > 0$$

Wanneer men zich niet bekommert om de wiskundige rechtvaardiging van een en ander is het bewijs eenvoudig en gaat het als volgt. Beschouw de 'hulpfunctie' $\varphi(z)$ gegeven door

(4.3) $$\varphi(z) = \int e^{zt} t \lambda(t)\, dt \quad , 0 \le z \le 1$$

Omdat $\lambda(t)$ gemiddelde 0 heeft geldt $\varphi(0) = 0$. Verder is $\varphi(1)$ gelijk aan het linkerlid van (4.2) zodat bewezen moet worden dat $\varphi(1) > 0$. De afgeleide van $\varphi(z)$ volgt door te differentiëren onder de integraal en is gelijk aan

(4.4) $$\varphi'(z) = \int e^{zt} t^2 \lambda(t)\, dt > 0$$

waarbij het groter dan teken volgt uit het feit dat de integrand (op het argument t = 0 na) strikt positief is. Omdat $\varphi(z)$ tussen 0 en 1 strikt stijgend is en omdat $\varphi(0) = 0$ moet dan $\varphi(1) > 0$ zijn waarmee het bewijs geleverd is.

De wiskundige rechtvaardiging van bovenstaande kan geschieden met behulp van de stelling A.2.6 en brengt het nodige schrijfwerk met zich mee. Omdat deze stelling elders in deze appendix diverse malen wordt gebruikt zal hier ter illustratie het gebruik van de stelling in detail worden gegeven, terwijl dat elders in de appendix voornamelijk aan de lezer zal worden overgelaten.

We zullen eerst aantonen dat $\varphi(z)$ continu is op [0, 1]. Definieer f(t,z) door

(4.5) $\quad f(t,z) = e^{zt} t\lambda(t) \quad , 0 \le z \le 1$

Voor $t \ge 0$ geldt $f(t,z) \le te^t\lambda(t)$ en voor $t < 0$ geldt $f(t,z) \le |t|\lambda(t)$. Dus

(4.6) $\quad f(t,z) \le g(t) \equiv (|t|+1)e^t\lambda(t) \quad , 0 \le z \le 1$, alle t

Omdat het gemiddelde van $\lambda(t)$ bestaat moet $|t|\lambda(t)$ integreerbaar zijn. Omdat $e^t \le e + |t|e^t$ volgt uit $E(Te^T) < \infty$ dat $E(e^T) < \infty$. Dan volgt

(4.7) $$\int g(t)\,dt \le \int_{-\infty}^{0} (|t|+1)\lambda(t)dt + \int_{0}^{\infty} (t+1)e^t\lambda(t)dt < \infty$$

waaruit blijkt, omdat g(t) niet-negatief is, dat g(t) integreerbaar is. Het is nu eenvoudig de voorwaarden uit stelling A.2.6 voor de verzameling $A = \mathbb{R}$ die volledige maat heeft. Beschouw eerst $z_0 \in (0, 1)$ en $\delta(z_0) = \min\{z_0, 1\text{-}z_0\} > 0$. Dan is

(4.8) $\quad f(t,z) \le g(t) \quad , |z - z_0| < \delta(z_0)$, alle t

Beschouw nu $z_0 = 0$ en $\delta(z_0) = 1/2$. Dan geldt

(4.9) $\quad f(t,z) \le g(t) \quad , 0 \le z < \delta(z_0)$, alle t

Verder geldt voor $z_0 = 1$ en $\delta(z_0) = 1/2$ dat

(4.10) $\quad f(t,z) \le g(t) \quad , \delta(z_0) < z \le 1$, alle t

Dan volgt uit stelling 3 punt 1 dat $\varphi(z)$ continu is op [0, 1].

We zullen nu bewijzen met behulp van stelling A.2.6 dat de afgeleide van $\varphi(z)$ op (0,1) wordt gegeven door het middelste lid van (4.4). Beschouw daartoe $z_0 \in (0, 1)$ en $\delta(z_0) = \min\{z_0, 1\text{-}z_0\}/2 > 0$. Dan is betrekkelijk eenvoudig te verifiëren dat

(4.11) $$\left|\frac{\partial f(t,z)}{\partial z}\right| \le g(t,z_0) \equiv \left(t^2 + t^2 e^{[z_0+\delta(z_0)]t}\right)\lambda(t) \quad , |z - z_0| < \delta(z_0), \text{ alle } t$$

Vanaf zekere t_0 zal gelden, omdat $z_0+\delta(z_0) < 1$, dat $g(t,z_0) < e^t$, zodat

(4.12) $$g(t,z_0) \le \left(t^2 + t^2 e^{[z_0+\delta(z_0)]t_0} + e^t\right)\lambda(t) \quad , \text{alle } t$$

Omdat de standaarddeviatie van $\lambda(t)$ bestaat moet de integraal over $t^2\lambda(t)$ eindig zijn, zodat $g(t,z_0)$ integreerbaar moet zijn. Met $A = \mathbb{R}$ volgt dan uit stelling A.2.6 dat de $\varphi'(z)$ inderdaad wordt gegeven door het middelste lid van (4.4). Het groter dan teken in (4.4) volgt omdat de integrand groter dan 0 is op een verzameling van positieve maat (namelijk op $\mathbb{R} \setminus \{0\}$ met Lebesguemaat ∞).

Dus φ(z) is continu op [0, 1] en differentieerbaar op (0, 1) waarbij φ(0) = 0. Uit de middelwaardestelling volgt dan dat voor een $z_0 \in (0, 1)$ moet gelden

$$(4.13) \qquad \varphi'(z_0) = \frac{\varphi(1) - \varphi(0)}{1 - 0} > 0$$

zodat geldt φ(1) = φ'(z_0) > 0 hetgeen bewezen moest worden.

A.5 Details hoofdstuk 5

In het vervolg worden formules (5.3) en (5.6) uit hoofdstuk 5 afgeleid. Het betreft slechts de berekening van relatief eenvoudige integralen die hier voor de volledigheid toch worden gegeven. Omdat de tweede bewering uit (5.3) uit hoofdstuk 5 direct volgt uit definitie (3.20) uit hoofdstuk 3, dienen slechts $E(e^T)$, μ_γ en $\overline{F}_Y(y)$ te worden berekend. Er geldt, met λ(t) gegeven door (5.1) uit hoofdstuk 5,

$$(5.1) \qquad E(e^t) = \int e^t \lambda(t) dt = \int_{-a}^{a} e^t \frac{1}{2a} dt = \frac{e^a - e^{-a}}{2a}$$

Met γ(t) gegeven door (5.4) uit hoofdstuk 5 volgt

$$(5.2) \qquad \mu_\gamma = \int \gamma(t) t \, dt = \frac{e^{\delta_0}}{2a} \int_{-a}^{a} t e^{-t} dt = \frac{e^{\delta_0}}{2a} \left[-(t+1)e^{-t} \right]_{-a}^{a} = \frac{e^{\delta_0}}{2a} \left(e^a (1-a) - e^{-a}(1+a) \right)$$

Hiermee is (5.3) uit hoofdstuk 5 aangetoond. We beschouwen nu (5.6) uit hoofdstuk 5. Voor $y \geq a + \delta$ volgt de juistheid van (5.6) uit hoofdstuk 5 onmiddellijk uit (3.22) en (5.1) uit hoofdstuk 3 en 5, terwijl voor $y \leq \delta - a$ de juistheid van (5.6) uit hoofdstuk 5 eenvoudig volgt uit (3.15) uit hoofdstuk 3. Beschouw nu $\delta - a < y < \delta + a$. Om gebruik te kunnen maken van (3.15) uit hoofdstuk 3 berekenen we eerst met behulp van partiële integratie de volgende integraal.

$$\begin{aligned} & \int_{-\infty}^{y-\delta} e^t \overline{\Lambda}(t) dt = \int_{-a}^{-a} e^t dt + \int_{-a}^{y-\delta} e^t \overline{\Lambda}(t) dt \\ (5.3) \qquad & = e^{-a} + \left[e^t \overline{\Lambda}(t) \right]_{-a}^{y-\delta} + \int_{-a}^{y-\delta} e^t \frac{1}{2a} dt \\ & = e^{y-\delta} \overline{\Lambda}(y-\delta) + \frac{1}{2a} \left(e^{y-\delta} - e^{-a} \right) \end{aligned}$$

Omdat $\overline{\Lambda}(t) = (1 - t/a)/2$ voor –a < t < a volgt nu met (3.15) uit hoofdstuk 3

$$(5.4) \qquad \overline{F}_Y(y) = \frac{1}{2} - \frac{y-\delta}{2a} + \frac{1}{2a} - \frac{1}{2a} e^{-y+\delta-a}$$

Hiermee is (5.6) uit hoofdstuk 5 aangetoond.

A.6 Details hoofdstuk 6

In paragraaf A.6.1 en A.6.3 wordt uitgegaan van de specifieke keuze dat delta gelijk is aan δ(s). Men mag echter in de beweringen zonder bezwaar δ(s) vervangen door een willekeurig getal δ < 0. Hieronder wordt (alleen in paragraaf A.6.1 maar niet in paragraaf A.6.2 en A.6.3) de grootheid $K_s(m)$ gebruikt, die feitelijk pas in (7.11) van hoofdstuk 7 wordt gedefinieerd. Deze grootheid is echter gelijk aan de K(m) uit hoofdstuk 6 indien een specifieke keuze voor λ(t) wordt gemaakt.

A.6.1 Dragers stochasten en continuïteit en differentieerbaarheid

Ten behoeve van de hoofdstukken 6 en 7 wordt aangenomen dat de dragers van de stochasten Q en M open intervallen vormen en dat de kansdichtheden g(q) en g(m) continu zijn op hun dragers. Van een stochast Z met kansdichtheid g(z) zal de drager worden aangegeven met D_Z, welke gedefinieerd is door

$$(6.1) \qquad D_Z = \{ z \in \mathbb{R} \mid g(z) > 0 \}$$

We brengen in herinnering dat volgens hoofdstuk 3 de drager van λ(t) werd aangegeven met $D = (y_b, y_e)$ en dat, voor s > 0, Y als drager het interval $D_{Y,s} = (y_b + \delta(s), \infty)$ heeft. De dragers van λ(t), X, Y, Q en M hebben dan de vorm

$$\begin{aligned} D &= (y_b, y_e) \\ D_X &= (0, \infty) \\ (6.2) \qquad D_{Y,s} &= (y_b + \delta(s), \infty) \\ D_Q &= (q_b, q_e) \\ D_M &= (m_b, m_e) \end{aligned}$$

De intervallen D, D_Q en D_M kunnen naar boven zowel als beneden begrensd dan wel onbegrensd zijn terwijl $D_{Y,s}$ naar beneden begrensd dan wel onbegrensd kan zijn. We geven nu beknopt wat eigenschappen van diverse grootheden. Die eigenschappen kunnen geverifieerd worden met behulp van stelling A.2.2 en aan de hand van hoofdstuk 3. We brengen in herinnering dat een reële functie die gedefinieerd is op een interval (a,b) *continu differentieerbaar* wordt genoemd indien in ieder punt x uit (a,b) de afgeleide f '(x) bestaat en continu is in dat punt. De inverse van f(x) zal (indien deze bestaat) worden aangegeven met $f^{-1}(x)$. Waar in het volgende de inverse van een functie wordt beschouwd kan altijd betrekkelijk eenvoudig worden aangetoond dat deze inverse bestaat; de functies waarvan de inverses worden beschouwd zijn namelijk altijd continu en strikt stijgend.

Er is aangenomen dat g(q) continu is op D_Q. Dan geldt dat F_Q: $D_Q \rightarrow (0, 1)$ zowel als F_Q^{-1}: $(0,1) \rightarrow D_Q$ continu differentieerbaar zijn. Evenzo is aangenomen dat g(m) continu is op D_M. Dan geldt dat F_M: $D_M \rightarrow (0, 1)$ zowel als F_M^{-1}: $(0,1) \rightarrow D_M$ continu differentieerbaar zijn. Voor de transformatie J(q) geldt dat $J : D_Q \rightarrow D_X$ zowel als zijn inverse $J^{-1} : D_X \rightarrow D_Q$ continu differentieerbaar zijn. Voor de afgeleide van J(q) volgt door differentiatie van (6.5) uit paragraaf 6.2

$$(6.3) \qquad J'(q) = \frac{g(q)}{1 - F_Q(q)} > 0 \qquad , q \in D_Q$$

Beschouw nu een vaste s > 0. Dan is $F_{Y,s}$: $D_{Y,s} \rightarrow (0, 1)$ zowel als $F_{Y,s}^{-1}$: $(0,1) \rightarrow D_{Y,s}$ continu differentieerbaar. Verder geldt dat $F_{Y,s}$: $\mathbb{R} \rightarrow [0, 1)$ continu differentieerbaar is waarbij $F_{Y,s} = 0$ voor $y \leq y_b + \delta(s)$. De transformatie $K_s(m)$ als functie van m is van de vorm K_s: $D_M \rightarrow D_{Y,s}$ en wordt vastgelegd door

$$(6.4) \qquad K_s(m) = F_{Y,s}^{-1}\left(F_M(m)\right) \qquad , m \in D_M$$

Deze functie is continu differentieerbaar. Door de tweede bewering in (6.2) uit paragraaf 6.2 te differentiëren naar m volgt voor de afgeleide van (6.4) naar m

(6.5) $$K_s'(m) = \frac{dK_s(m)}{dm} = \frac{g(m)}{g_{Y,s}\left(K_s(m)\right)} > 0 \quad , m \in D_M$$

Met als domein $D_{Y,s}$ bestaat de inverse van $K_s(m)$ met als bereik D_M, ofwel K_s^{-1}: $D_{Y,s} \rightarrow D_M$.

Volgens Lemma A.7.1 is daarnaast voor iedere vaste $m \in D_M$ de functie $s \rightarrow K_s(m)$, ofwel $K_{(.)}(m) : (0, \infty) \rightarrow D_{Y,s}$ vastgelegd door

(6.6) $$K_s(m) = F_{Y,s}^{-1}\left(F_M(m)\right) \quad , s > 0$$

continu differentieerbaar. Merk op dat de eventuele discontinuïteitspunten in $\lambda(t)$ niet tot uitdrukking komen in de hier beschouwde grootheden. Ondanks dat in $\lambda(t)$ sprongen mogen voorkomen zijn de hier beschouwde grootheden alle zeer netjes, in de zin van differentieerbaar met continue afgeleiden.

A.6.2 Transformatie van de bivariate kansdichtheid

We geven hier het bewijs van (6.7) en (6.8) uit hoofstuk 6. Daarna wordt geverifieerd dat g(q,m) als marginale verdelingen g(q) en g(m) heeft.

In de getransformeerde ruimte wordt de gezamenlijke kansdichtheid gegeven door

(6.7) $$g_{X,Y}(x,y) = g_X(x)\lambda(y - x - \delta)$$

De drager $D_{X,Y}$ van deze gezamenlijke kansdichtheid wordt gegeven door

(6.8) $$D_{X,Y} = \left\{(x,y) \in \mathbb{R}^2 \,\middle|\, g_{X,Y}(x,y) > 0\right\}$$

De formule om de kansdichtheid $g_{X,Y}(x,y)$ te transformeren naar $g_{Q,M}(q,m)$ is standaard, zie bijvoorbeeld [Billingsley, 1995; p. 261] en [Casella, 1990; p. 148]. De drager van $g_{Q,M}(q,m)$ wordt gegeven door, waarbij J^{-1} en K^{-1} volgens paragraaf A.6.1 de inverses van J en K aangeven,

(6.9)
$$\begin{aligned} D_{Q,M} &= \left\{(q,m) \in \mathbb{R}^2 \,\middle|\, g_{Q,M}(q,m) > 0\right\} \\ &= \left\{(q,m) \in \mathbb{R}^2 \,\middle|\, q = J^{-1}(x) \text{ en } m = K^{-1}(y) \text{ voor een } (x,y) \in D_{X,Y}\right\} \end{aligned}$$

Voor $(q,m) \in D_{Q,M}$ geldt dan

(6.10) $$g_{Q,M}(q,m) = g_{X,Y}\left(J(q), K(m)\right)\left|D\right|$$

waarbij D de Jacobiaan voorstelt, die in dit geval gegeven wordt door

(6.11) $$D = \begin{vmatrix} \frac{\partial J(q)}{\partial q} & 0 \\ 0 & \frac{\partial K(m)}{\partial m} \end{vmatrix} = J'(q)K'(m) > 0$$

De positiviteit van D volgt uit (6.3) en (6.5). Voor punten $(q,m) \notin D_{Q,M}$ is uiteraard $g_{Q,M}(q,m) = 0$. Terzijde merken we op dat de wiskundige rechtvaardiging voor het toepassen van (6.10) eenvoudig geverifieerd kan worden op basis van de met paragraaf A.6.1 te checken feiten dat voor $q \in D_{Q,M}$ ten eerste J(q) en K(m) beiden continu differentieerbaar zijn en ten tweede dat de afbeelding die aan $(x,y) \in D_{X,Y}$ het punt $(J^{-1}(x), K^{-1}(y)) \in D_{Q,M}$ toevoegt bijectief is. Merk op dat formule (6.10) algemeen geldig is los van de specifieke vorm die (6.7) heeft. Uit (6.7), (6.10) en (6.11) volgt nu

(6.12) $$g_{Q,M}(q,m) = g_X\left(J(q)\right)\lambda\left(K(m) - J(q) - \delta\right)J'(q)K'(m)$$

Vanwege (6.2) geldt

(6.13) $$g_Q(q) = F_Q'(q) = \frac{d}{dq}F_X\left(J(q)\right) = g_X\left(J(q)\right)J'(q)$$

zodat

(6.14) $$g_{Q,M}(q,m) = g_Q(q)\lambda\left(K(m) - J(q) - \delta\right)K'(m)$$

Na deling door $g_Q(q)$ volgt dat g(m|q) inderdaad wordt gegeven door de tweede bewering in (6.7) uit hoofdstuk 6. De eerste bewering volgt uit (6.13) omdat $g_X(x) = e^{-x}$. Hiermee is (6.7) uit hoofstuk 6 bewezen. Bewering (6.8) van hoofdstuk 6 volgt uit (6.7) van hoofdstuk 6 en de definitie van $\gamma(t)$ gegeven in (4.4) van hoofdstuk 4.

We zullen tenslotte verifiëren dat g(q,m) de juiste marginale verdelingen heeft. In feite is dat laatste een direct gevolg van (6.10) en (6.11) waarbij de precieze vorm van J(q) en K(m) irrelevant is. We zullen slechts verifiëren dan de integraal van g(q,m) over alle q gelijk is aan g(m), omdat het bewijs dat de integraal van g(q,m) over alle m gelijk is aan g(q) analoog verloopt. Beschouw daartoe een vaste $m \in D_M$. Neem eerst aan dat $\lambda(t)$ geen enkel discontinuïteitspunt heeft. In dat geval volgt dat voor iedere $x \in D_X$ de grootheid

(6.15) $$P(Y < y \mid x) = \Lambda\left(y - x - \delta\right)$$

differentieerbaar is naar y voor elke $y \in \mathbb{R}$. Dan volgt uit (6.10) en (6.11), met behulp van de substititie x = J(q) en (6.2) uit hoofdstuk 6

$$\begin{aligned}
\int g(q,m)\,dq &= \int_{D_X} g_{X,Y}\left(x, K(m)\right)K'(m)\,dx \\
&= \int_{D_X} g_X(x)\,g_{Y|x}\left(K(m) \mid x\right)K'(m)\,dx \\
&= \int_{D_X} g_X(x)\,\frac{d}{dm}P(Y < K(m) \mid x)\;dx \\
&= \frac{d}{dm}\int_{D_X} g_X(x)\;P(Y < K(m) \mid x)\;dx \qquad (6.16) \\
&= \frac{d}{dm}F_Y\left(K(m)\right) \\
&= \frac{d}{dm}F_M(m) \\
&= g_M(m)
\end{aligned}$$

In feite dient in bovenstaande de stap waarin de afgeleide naar m buiten de integraal wordt gehaald gerechtvaardigd te worden. Dat kan gedaan worden, indien voor P(Y<y|x) de expliciete vorm uit (6.15) wordt ingevuld, met behulp van Stelling A.2.6; de details worden aan de lezer overgelaten. Neem nu aan dat $\lambda(t)$ eindig veel discontinuïteitspunten t_i, i = 1, 2,..., n heeft. Dan is P(Y<K(m)|x) differentieerbaar in het (vaste) punt m, uitgezonderd voor de waarden $x_i = K(m) - \delta - t_i$, i = 1, 2,..., n. De integratie over x in (6.16) kan dan worden gesplitst in de trajecten $(0, x_1)$, (x_1, x_2),...., (x_n, ∞). Het is te verifiëren dat op elk van deze trajecten de afgeleide naar m buiten de integraal gehaald mag worden. Vervolgens kunnen deze trajecten weer worden samengevoegd (omdat differentiëren een lineaire operatie vormt) tot de afgeleide naar m van één integraal die gelijk is aan $F_Y(K(m))$. Dan blijkt dat ook indien $\lambda(t)$ discontinuïteitspunten heeft formule (6.16) zijn geldigheid behoudt.

A.6.3 Eigenschappen van δ(s)

Lemma A.6.1 Neem aan dat $E_s(e^T) < \infty$. Dan is de functie δ(s) continu differentieerbaar op s > 0. Tevens geldt

$$\delta'(s) = \frac{\int te^{st}\lambda_1(t)\,dt}{\int e^{st}\lambda_1(t)\,dt} < 0 \tag{6.17}$$

$$\lim_{s\downarrow 0}\delta(s) = 0 \tag{6.18}$$

$$\lim_{s\downarrow 0}\frac{\delta(s)}{s} = \lim_{s\downarrow 0}\delta'(s) = 0 \tag{6.19}$$

Bewijs Wanneer men zich niet druk maakt om de wiskundige rechtvaardiging van een en ander is het bewijs van (6.17), (6.18) en (6.19) niet zo moeilijk. Hier wordt beknopt aangegeven hoe een precies bewijs gegeven kan worden met behulp van stelling A.2.6. Het domein van δ(s) is (0, ∞) hetgeen directe toepassing van de genoemde stelling onmogelijk maakt. Daarom definiëren we een hulpfunctie $\tilde{\delta}(s)$ die als domein heel $\mathbb{R}$ heeft. Voor s, t $\in \mathbb{R}$ definiëren we

$$\begin{aligned} f(t,s) &= \left[\chi_{(-\infty,0]}(s)(1+st) + \chi_{(0,\infty)}(s)e^{st}\right]\lambda_1(t) \\ I(s) &= \int f(t,s)dt \geq 1 \\ \tilde{\delta}(s) &= -\ln\left(I(s)\right) \leq 0 \end{aligned} \tag{6.20}$$

Dan geldt, omdat $\lambda_1(t)$ gemiddelde nul heeft,

$$\tilde{\delta}(s) = \begin{cases} \delta(s) & , s>0 \\ 0 & , s \leq 0 \end{cases} \tag{6.21}$$

Voor I(s) kan op basis van Stelling A.2.6 met het nodige schrijfwerk worden aangetoond dat I(s) continu differentieerbaar is op heel $\mathbb{R}$ waarbij de afgeleide wordt gegeven door

$$\begin{aligned} I'(s) &= \int\left[\chi_{(-\infty,0]}(s)t + \chi_{(0,\infty)}(s)te^{st}\right]\lambda_1(t)\ dt \\ &= \begin{cases} \int te^{st}\lambda_1(t)\,dt & , s>0 \\ 0 & , s \leq 0 \end{cases} \end{aligned} \tag{6.22}$$

Analoog aan Lemma A.4.1 kan worden bewezen dat I'(s) > 0 voor s > 0. Omdat I(s) ≥ 1 voor s $\in \mathbb{R}$ volgt dan uit de definitie van $\tilde{\delta}(s)$ in (6.20) dat deze grootheid op heel $\mathbb{R}$ differentieerbaar is, met afgeleide

$$\tilde{\delta}'(s) = -\frac{I'(s)}{I(s)} \tag{6.23}$$

De continuïteit van deze grootheid volgt uit het feit dat de teller en noemer in het rechterlid beide continu zijn terwijl de noemer altijd positief is. Voor s > 0 volgt uit (6.20), (6.21) en (6.22) en omdat I'(s) > 0 dat (6.17) moet gelden. Bewering (6.18) volgt uit de continuïteit van $\tilde{\delta}(s)$ en omdat $\tilde{\delta}(0) = 0$. Bewering (6.19) volgt uit het feit dat $\tilde{\delta}'(s)$ continu is in 0 en uit het feit dat

$$0 = \tilde{\delta}'(0) = \lim_{s\downarrow 0}\frac{\tilde{\delta}(s)}{s} = \lim_{s\downarrow 0}\frac{\delta(s)}{s} \tag{6.24}$$

Hiermee zijn alle beweringen uit het lemma bewezen.

A.7 Details hoofdstuk 7

Hieronder volgen een aantal lemma's en stellingen waaruit de beweringen in hoofdstuk 7 volgen. Omwille van het overzicht volgen de bewijzen pas ná de formuleringen van de lemma's en stellingen.

Schrijf als alternatieve notatie $F(s,y) = F_{Y,s}(y)$ en schrijf voor de de partiële afgeleide naar y $g(s,y) = g_{Y,s}(y)$. Het domein van deze functies wordt gegeven door $D = (0,\infty)\times\mathbb{R}$.

Lemma A.7.1 Beschouw $F : D \to \mathbb{R}$ en $g : D \to \mathbb{R}$ als zojuist gedefinieerd.

1. Dan is F continu en de partiële afgeleiden van F bestaan en zijn continu.
2. Voor iedere vaste $m \in D_M$ is de functie $s \to K_s(m)$ continu differentieerbaar op $(0, \infty)$ met afgeleide

(7.1) $$\frac{\partial K_s(m)}{\partial s} = -\frac{\partial F(s,y)/\partial s}{g(s,y)}\bigg|_{y=K_s(m)}$$

Lemma A.7.2 Voor iedere vaste $m \in D_M$ geldt

(7.2) $$\lim_{s\downarrow 0} K_s(m) = -\ln\left(1 - F_M(m)\right)$$

Lemma A.7.3 Beschouw een vaste $x_0 \geq 0$ en een vaste $t \in \mathbb{R}$. Dan geldt

(7.3) $$\lim_{s\downarrow 0} F(s, st + x_0 + \delta(s)) = 1 - e^{-x_0}$$

en

(7.4) $$\lim_{s\downarrow 0} F(s, x_0) = 1 - e^{-x_0}$$

Lemma A.7.4 Beschouw een vaste $x_0 \geq 0$ en $m \in D_M$. Dan geldt

(7.5) $$\lim_{s\to\infty} \frac{K_s(m) - x_0 - \delta(s)}{s} = \Lambda_1^{-1}\left(F_M(m)\right)$$

Stelling A.7.5 Er geldt, voor $q \in D_Q$ en $m \neq m(q)$

(7.6) $$\lim_{s\downarrow 0} F(m \mid q) = \begin{Bmatrix} 1 & , m > m(q) \\ 0 & , m < m(q) \end{Bmatrix}$$

waarbij m(q) wordt gegeven door (7.31) uit hoofdstuk 7. Verder geldt, voor $q \in D_Q$ en $m \in \mathbb{R}$

(7.7) $$\lim_{s\downarrow 0} g_s(q,m) = g(q)\delta_D\left(m - m(q)\right)$$

met $\delta_D(t)$ de Dirac delta functie met locatieparameter 0.

Stelling A.7.6 Stel dat $\lambda_1(t)$ een eindig aantal van n discontinuïteitspunten heeft die tot de drager ($y_b(1)$, $y_e(1)$) van $\lambda_1(t)$ behoren, waarbij n eventueel 0 mag zijn. Geef voor n > 0 deze aan met $t_1, t_2, \ldots, t_n$. Definieer m_i door

(7.8) $$m_i = F_M^{-1}\left(\Lambda_1(t_i)\right) \ , i = 1,2,...,n$$

Merk op dat n = 0 indien $\lambda_1(t)$ continu is of indien $\lambda_1(t)$ slechts $y_b(1)$ en/of $y_e(1)$ als discontinuïteitspunt heeft. Er geldt nu

(7.9) $$\lim_{s\to\infty} F_s(m \mid q) = F_M(m) \quad , q \in D_Q \text{ en } m \in \mathbb{R}$$

(7.10) $$\lim_{s\to\infty} g_s(q,m) = g(q)g(m) \quad , q \in D_Q \text{ en } m \in \mathbb{R} \text{ met } m \neq m_i, \ i = 1,2,...,n$$

Indien n = 0 geldt de laatste bewering voor alle $q \in D_Q$ en $m \in \mathbb{R}$.

Stelling A.7.7

1. Beschouw $L_{tr}(s)$ volgens (7.36) uit hoofdstuk 7. Dan geldt

(7.11) $$\lim_{s\to\infty} L_{tr}(s) = 0$$

Neem aan dat voor de dataset $(q_i, m_i) \in D_Q \times D_M, i = 1,2,...,N,$ geldt dat $F_Q(q_i) \neq F_M(m_i)$, ofwel dat geen enkel datapunt op de lijn van gelijke kansen ligt, hetgeen met kans 1 het geval is met betrekking tot $g_s(q,m)$ voor elke s > 0. Dan geldt

(7.12) $$\lim_{s\downarrow 0} L_{tr}(s) = 0$$

Verder geldt voor minstens één $s_0 \in (0,\infty)$ dat $L_{tr}(s_0) > 0$.

2. Ga uit van dezelfde voorwaarden aan de dataset als ten behoeve van (7.12). Dan neemt, indien $\lambda_1(t)$ continu is $L_{tr}(s)$ zijn maximum aan op het interval (0,∞).

Met betrekking tot de voorgaande stelling maken we de volgende opmerkingen. In een concrete toepassing zal men in de regel de waarde van s waarvoor $L_{tr}(s)$ maximaal wordt bepalen door te beginnen met een kleine waarde van s en deze dan steeds iets ophogen totdat blijkt voor welke waarde van s de functie $L_{tr}(s)$ maximaal wordt. Het is (nog niet) aangetoond dat er slechts één oplossing voor s bestaat die $L_{tr}(s)$ maximaliseert. In pricipe is denkbaar dat meerdere oplossingen bestaan, hoewel dat schrijvers dezes gezien de aard van het probleem in ieder geval in praktische toepassingen onwaarschijnlijk lijkt. Verder valt er de wiskundig-technische opmerking te maken dat een discontinue $L_{tr}(s)$, die resulteert als λ_1 discontinu is, niet altijd zijn maximum hoeft aan te nemen. Bijvoorbeeld de functie

(7.13) $$f(s) = \begin{cases} s & , s < 1 \\ 0 & , s \geq 1 \end{cases}$$

heeft supremum 1 terwijl door de discontinuïteit in s = 1 altijd f(s) < 1. Met betrekking tot de praktijk betreft het hier meer een spitsvondigheid. Indien bijvoorbeeld s door ‘trial and error’ wordt bepaald zal men mogelijk tot de conclusie komen dat f(s) maximaal wordt door s = 0.99, of voor s = 0.995 of iets dergelijks. De gevonden waarde kan dan in de praktijk verder worden gebruikt. Voor de praktijk is de gesignaleerde spitsvondigheid dus niet zeer relevant.

Samenvattend kan het volgende gesteld worden. Indien aan de genoemde ‘spitsvondigheid’ voorbij wordt gegeaan kan op basis van (7.11) en (7.12) worden gesteld dat $L_{tr}(s)$ zijn maximum aanneemt voor een eindige s’> 0. In principe kunnen meerdere oplossingen bestaan, maar dat lijkt voor praktijksituaties niet erg waarschijnlijk.

Bewijs Lemma A.7.1 Om aan te tonen dat F continu is op D beschouwen we een rij (s_n, y_n) in D die convergeert naar $(s, y) \in D$. We dienen dan te laten zien dat

(7.20) $\lim_{n\to\infty} F(s_n, y_n) = F(s, y)$

met volgens (7.7) uit hoofdstuk 7

(7.21) $$F(s,y) = \int_0^\infty e^{-x} \Lambda_1\left(\frac{y - x - \delta(s)}{s}\right) dx$$

Omdat Λ_1 continu is en omdat $\delta(s)$ continu is op $(0,\infty)$ kan met behulp van de LDC-stelling (zie Stelling A.2.4) vrij eenvoudig worden aangetoond dat (7.20) geldt. Dus F is continu op D.

Beschouw nu de partiële afgeleide van F naar y, die volgens (3.16) uit hoofdstuk 3 kan worden geschreven als

(7.22) $$g(s,y) = \frac{\partial F(s,y)}{\partial y} = \Lambda_1\left(\frac{y - \delta(s)}{s}\right) - F(s,y)$$

Omdat de laatste twee leden beiden continu zijn op D, volgt dat g(s,y) continu is op D.

Beschouw nu de partiële afgeleide van F naar s. Indien $\lambda_1(t)$ continu zou zijn, zou het bestaan en de continuïteit van $\partial F(s,y)/\partial s$ vrij eenvoudig volgen uit stelling A.2.4 en A.2.6. De $\lambda_1(t)$ mag echter eindig veel discontinuïteitspunten hebben, wat de situatie compliceert. We zullen aannemen dat $\lambda_1(t)$ één discontinuïteitspunt heeft. De situatie dat $\lambda_1(t)$ meerdere discontinuïteitspunten heeft is niet wezenlijk ingewikkelder en wordt aan de lezer overgelaten. Neem dus aan dat $\lambda_1(t)$ één discontinuïteitspunt α heeft dat dan behoort tot $[y_b(1), y_e(1)]$. Beschouw nu een vaste $(s, y) \in D$ en h met $0 < |h| < s/2$. Schrijf nu

(7.23) $$I_h(x) = \frac{1}{h}\left(\Lambda_1\left(\frac{y - x - \delta(s+h)}{s+h}\right) - \Lambda_1\left(\frac{y - x - \delta(s)}{s}\right)\right)$$

Dan volgt

(7.24) $$\frac{F(s+h,y) - F(s,y)}{h} = \int \chi_{(0,\infty)}(x) e^{-x} I_h(x)\ dx$$

Definieer nu $A = \mathbb{R} \setminus \{y - \delta(s) - \alpha s\}$. Dan volgt voor $x \in A$, omdat $\lambda_1(t) < C$ voor zekere constante C, uit (7.6) uit hoofdstuk 7

(7.25) $$|I_h(x)| \le C \left| \frac{\frac{y - x - \delta(s+h)}{s+h} - \frac{y - x - \delta(s)}{s}}{h} \right|$$

Bedenk nu dat vanwege de keuze van h altijd $s + h \in (s/2, 3s/2)$ met uiteraard $s/2 > 0$. Omdat $f(t) = [y - x - \delta(t)]/t$ differentieerbaar is voor $t > 0$, volgt dan met de middelwaardestelling dat de term tussen absoluutstrepen in het rechterlid van (7.25) gelijk moet zijn aan $f'(s + \theta h)$ voor een $\theta \in (0,1)$. Omdat $\delta(t)$ volgens lemma A.6.1 continu differentieerbaar is op $(0,\infty)$, is $f'(t)$ continu op $[s/2, 3s/2]$. Derhalve bestaat een constante B waarvoor $|f'(s + \theta h)| < B$ voor alle beschouwde h en θ. Dus moet gelden

(7.26) $|I_h(x)| \le BC$, $x \in A$ en $0 < |h| < s/2$

Als $x \in A$ volgt $[y - x - \delta(s)]/s \neq \alpha$, wat inhoudt dat voor $x \in A$ de afgeleide van $k(t) = \Lambda_1([y - x - \delta(t)]/t)$ bestaat in $t = s$. Uit (7.23) volgt dan voor $x \in A$

(7.27) $$\lim_{h\to 0} I_h(x) = \lambda_1\left(\frac{y-x-\delta(s)}{s}\right)\left\{\frac{-s\delta'(s)-y+x+\delta(s)}{s^2}\right\}$$

Uit (7.24), (7.26) en (7.27) volgt dan met behulp van de LDC-stelling, zie Stelling A.2.5, dat

(7.28) $$\frac{\partial F(s,y)}{\partial s} = \int \chi_{(0,\infty)}(x)e^{-x}\lambda_1\left(\frac{y-x-\delta(s)}{s}\right)\left\{\frac{-s\delta'(s)-y+x+\delta(s)}{s^2}\right\}dx$$

We dienen nog te bewijzen dat $\partial F(s,y)/\partial s$ continu is op D. Beschouw daartoe een vaste $(s,y)\in D$ en een rij (s_n, y_n) waarvoor $s_n \in [s/2, 3s/2]$ en $y_n \in [y-1, y+1]$ die convergeert naar (s, y). We zullen aantonen dat dan noodzakelijkerwijs

(7.29) $$\lim_{h\to\infty}\frac{\partial F(s_n,y_n)}{\partial s_n} = \frac{\partial F(s,y)}{\partial s}$$

wat betekent dat $\partial F(s,y)/\partial s$ continu is op D. Bewering (7.29) kan worden bewezen met de LDC-stelling. Definieer daartoe

(7.30) $$f_n(x) = \chi_{(0,\infty)}(x)e^{-x}\lambda_1\left(\frac{y_n-x-\delta(s_n)}{s_n}\right)\left\{\frac{-s_n\delta'(s_n)-y_n+x+\delta(s_n)}{s_n^2}\right\}$$

Omdat δ(s) continu differentieerbaar is volgt dat de functie $(s, y)\to[-s\delta'(s)-y+x+\delta(s)]/s^2$ continu is op $[s/2, 3s/2]\times[y-1, y+1]$. Dan volgt dat de term tussen accolades in (7.30) begrensd moet zijn. Omdat $\lambda_1(t)$ tevens begrensd is volgt dan dat $|f_n(x)|$ wordt begrensd door een integreerbare functie. Voor $x\in A$ geldt verder dat $f_n(x)$ convergeert naar de integraal in (7.28) als $n\to\infty$. De LDC-stelling laat dan zien dat (7.29) moet gelden. Hiermee is punt 1 van het Lemma bewezen.

Punt 2 uit het Lemma volgt vrij eenvoudig uit de impliciete functiestelling, zie Stelling A.2.3. Hier volgen de details.

Beschouw een vaste $s_0>0$ en een vaste $m\in D_M$. Definieer y_0 door

(7.31) $$y_0 = K_{s_0}(m) \in D_{Y,s_0}$$

Schrijf $I_1 = (s_0/2, 3s_0/2)$ en $I_2 = \mathbb{R}$ en definieer H: $I_1 \times I_2 = \mathbb{R}$ door

(7.32) $$H(s,y) = F(s,y) - F_M(m)$$

Uit (7.10) en (7.11) uit hoofdstuk 7 volgt dan met (7.31)

(7.33) $$H(s_0, y_0) = 0$$

terwijl (7.31) tevens impliceert

(7.34) $$\frac{\partial H(s_0,y_0)}{\partial y_0} = g(s_0,y_0) > 0$$

Uit stelling A.2.3 volgt nu dat de afgeleide van de functie $s\to K_s(m)$ in s_0 gelijk is aan (7.1) met daarin s vervangen door s_0. Bovendien geeft de stelling aan dat deze afgeleide continu is in s_0. Omdat $s_0 > 0$ willekeurig is gekozen, is hiermee punt 2 van het lemma bewezen.

Bewijs Lemma A.7.2 Beschouw een vaste $m \in D_M$ en definieer, voor s > 0, h(s) door

(7.35) $\quad h(s) = K_s(m) - \delta(s)$

Uit (7.21) en (7.11) uit hoofdstuk 7 volgt dan

(7.36)
$$\begin{aligned} F_M(m) &= \int_0^{\infty} e^{-x} \Lambda_1\left(\frac{h(s)-x}{s}\right) dx \\ &= e^{-h(s)} \int_{-\infty}^{h(s)} e^t \Lambda_1\left(\frac{t}{s}\right) dt \end{aligned}$$

Beschouw nu een rij $s_n \to 0$ waarvoor $h(s_n) \to \alpha$ voor een zekere $\alpha \in [-\infty, \infty]$. Hieronder zal worden aangetoond dat dan altijd $\alpha \in (0, \infty)$. Neem daarom nu aan dat $\alpha \in (0, \infty)$. Omdat Λ_1 continu is geldt

(7.37) $\quad \lim_{s \downarrow 0} \Lambda_1\left(\frac{t}{s}\right) = \begin{cases} 1 & , t > 0 \\ \Lambda_1(0) & , t = 0 \\ 0 & , t < 0 \end{cases}$

Formeel kan nu in (7.36) de limiet s ↓ 0 worden genomen via de rij s_n, met als resultaat

(7.38) $\quad F_M(m) = e^{-\alpha} \int_0^{\alpha} e^t dt = 1 - e^{-\alpha}$

hetgeen gerechtvaardigd kan worden met de LDC-stelling. Deze rechtvaardiging wordt aan de lezer overgelaten. Uit (7.38) volgt dan

(7.39) $\quad \alpha = -\ln(1 - F_M(m)) \in (0,1)$

Aldus is aangetoond dat in iedere rij $s_n \downarrow 0$ waarvoor $h(s_n)$ convergeert geldt dat $h(s_n)$ convergeert naar α gegeven door (7.39). Maar dat houdt in dat h(s) convergeert naar α gegeven door (7.39) als s ↓ 0. Omdat $\delta(s) \to 0$ als s ↓ 0 volgt (7.2) dan met behulp van (7.35).

Er rest nog aan te tonen dat $h(s_n) \to \alpha$ indien $s_n \downarrow 0$ impliceert dat $\alpha \in (0, \infty)$. We zullen dat doen door te laten zien dat de gevallen $\alpha = 0$, $\alpha \in [-\infty, 0)$ en $\alpha = \infty$ tot een tegenstrijdigheid leiden. Neem eerst aan $\alpha = 0$. Schrijf als afkorting

(7.40) $\quad I_n(a,b) = e^{-h(S_n)} \int_a^b e^t \Lambda_1\left(\frac{t}{s_n}\right) dt$

Kies een vaste $\varepsilon > 0$. Voor alle n zo groot dat $|h(s_n)| < \varepsilon$ volgt met behulp van (7.36)

(7.41) $\quad F_M(m) = I_n(-\infty, h(s_n)) = I_n(-\infty, -\varepsilon) + I_n(-\varepsilon, h(s_n))$

Met de LDC-stelling kan geverifieerd worden dat $I_n(-\infty, -\varepsilon) \to 0$ als $n \to \infty$. Verder geldt, indien n zo groot dat $|h(s_n)| < \varepsilon$, dat $|I_n(\varepsilon, h(s_n))| \leq 2\varepsilon$. Dan volgt uit (7.41)

(7.42) $\quad F_M(m) \leq \limsup_{n \to \infty} I_n(-\varepsilon, h(s_n)) \leq 2\varepsilon$

Omdat ε willekeurig is volgt dan $F_M(m) = 0$, hetgeen strijdig is met het feit dat $m \in D_M$. Dus geldt $\alpha \neq 0$. De aanname $\alpha \in [-\infty, 0)$ leidt eveneens, zoals betrekkelijk eenvoudig te verifiëren valt, tot $F_M(m) = 0$. Dus $\alpha \notin [-\infty, 0)$. De aanname $\alpha \neq \infty$ blijkt te leiden tot de conclusie $F_M(m) = 1$, hetgeen eenvoudig te verifiëren valt

met behulp van het tweede lid van (7.36). Ook deze conclusie is strijdig met $m \in D_M$. Dus alleen $\alpha \in (0,\infty)$ is mogelijk. Hiermee is het bewijs van het lemma voltooid.

Bewijs Lemma A.7.3 Uit de definitie van F(s,y), zie bijvoorbeeld (7.21), volgt

(7.43) $$F\left(s, st + x_0 + \delta(s)\right) = \int f_s(x)\,dx$$

met

(7.44) $$f_s(x) = \chi_{(0,\infty)}(x)e^{-x}\Lambda_1\left(t + \frac{x_0 - x}{s}\right) \le \chi_{(0,\infty)}(x)e^{-x}$$

Voor $x \neq x_0$ geldt

(7.45) $$\lim_{s\downarrow 0} f_s(x) = \chi_{(0,x_0)}(x)e^{-x}$$

Met behulp van de LDC-stelling kan nu worden geverifieerd dat

(7.46) $$\lim_{s\downarrow 0} F\left(s, st + x_0 + \delta(s)\right) = \int_0^{x_0} e^{-x}dx = 1 - e^{-x_0}$$

waarmee (7.3) bewezen is. Bewering (7.4) kan op analoge wijze bewezen worden, door gebruik te maken van het feit dat δ(s)/s →0 volgens (6.19). De details worden aan de lezer overgelaten. Hiermee is het Lemma bewezen.

Bewijs Lemma A.7.4 Beschouw een vaste $x_0 \ge 0$ en een vaste $m \in D_M$. Volgens (7.21) en daarnaast (7.11) uit hoofdstuk 7 geldt

(7.47) $$F_M(m) = \int_0^{\infty} e^{-x}\Lambda_1\left(\frac{K_s(m) - x - \delta(s)}{s}\right)dx$$

Beschouw nu een rij $s_n \to \infty$ die de eigenschap heeft dat

(7.48) $$\lim_{n\to\infty} \frac{K_{s_n}(m) - \delta(s_n)}{s_n} = \alpha \in [-\infty,\infty]$$

Omdat Λ_1 continu is volgt dan voor iedere $x \ge 0$

(7.49) $$\lim_{n\to\infty} \Lambda_1\left(\frac{K_{s_n}(m) - x - \delta(s_n)}{s_n}\right) = \Lambda_1(\alpha)$$

Door nu formeel in (7.47) de limiet $s \to \infty$ te nemen via de rij s_n volgt dan

(7.50) $$F_M(m) = \int_0^{\infty} e^{-x}\Lambda_1(\alpha)dx = \Lambda_1(\alpha)$$

Dat deze limiet inderdaad onder het integraal teken genomen mag worden kan worden geverifieerd met de LDC-stelling. Dus volgt

(7.51) $$\alpha = \Lambda_1^{-1}(F_M(m)) \in (0,1)$$

De conclusie is dat iedere rij s_n waarvoor de limiet in het linkerlid van (7.48) bestaat, leidt tot een en dezelfde limietwaarde α gegeven door (7.51). Maar dat impliceert dat

$$(7.52) \quad \lim_{s\to\infty} \frac{K_s(m)-\delta(s)}{s} = \Lambda_1^{-1}(F_M(m))$$

Hieruit volgt (7.5), waarmee het lemma is bewezen.

Bewijs Lemma A.7.5 Beschouw om (7.6) te bewijzen een vaste $q \in D_Q$ en $m \neq m(q)$. Uit (7.9) en (7.31) uit hoofdstuk 7 volgt

$$(7.53) \quad J(q) = -\ln\left(1-F_M(m(q))\right)$$

Volgens (7.13) uit hoofdstuk 7 geldt

$$(7.54) \quad F_s(m \mid q) = \Lambda_1\left(\frac{K_s(m)-J(q)-\delta(s)}{s}\right)$$

Neem vooreerst aan dat m < m(q). Dan is $F_M(m) < F_M(m(q))$ en volgt met behulp van Lemma A.7.2 en (7.53), omdat $x \to -\ln(1-x)$ stijgend is op (0,1),

$$(7.55) \quad \lim_{s\downarrow 0} K_s(m) - J(q) = -\ln\left(1-F_M(m)\right) - J(q) < -\ln\left(1-F_M(m(q))\right) - J(q) = 0$$

Omdat volgens (6.19) $\delta(s)/s \to 0$ als $s \downarrow 0$ volgt hieruit dat

$$(7.56) \quad \lim_{s\downarrow 0}\left(\frac{K_s(m)-J(q)-\delta(s)}{s}\right) = -\infty$$

zodat uit (7.54) volgt dat $F_s(m|q) \to 0$ als $s \downarrow 0$. Voor m > m(q) volgt analoog dat $F_s(m|q) \to 1$ als $s \downarrow 0$, waarmee (7.6) is bewezen.

In het bewijs van (7.7) wordt er van uitgegaan dat de lezer bekend is met de Dirac delta functie en met testfuncties. Beschouw een continue begrensde testfunctie k(m). Om (7.7) te bewijzen dient te worden aangetoond dat

$$(7.57) \quad \lim_{s\downarrow 0} \int k(m) g_s(m \mid q)\, dm = k\left(m(q)\right)$$

Met behulp van (7.12) uit hoofdstuk 7 kan worden geschreven

$$\begin{aligned} &\int k(m) g_s(m \mid q)\, dm \\ (7.58) \quad &= \int k(m) \lambda_1\left(\frac{K_s(m)-J(q)-\delta(s)}{s}\right)\frac{K_s'(m)}{s}\, dm \\ &= \int k\left(K_s^{-1}(sy+J(q)+\delta(s)\right)\lambda_1(y)\, dy \end{aligned}$$

waarbij K_s^{-1} volgens paragraaf A.6.1 wordt gegeven door

$$(7.59) \quad K_s^{-1}(y) = F_M^{-1}\left(F(s,y)\right) \quad , y \in D_{Y,s} = (y_b(s)+\delta(s), \infty)$$

In de integraal in (7.58) hoeft slechts geïntegreerd te worden over alle y waar voor $\lambda_1(y) > 0$. Het is dan, omdat J(q) > 0, eenvoudig te verifiëren dat sy + J(q) + δ(s) bevat is in $D_{Y,s}$, zodat het argument van de functie k in de

integraal in het laatste lid van (7.58) dus goed gedefinieerd is. Omdat F_M^{-1} continu is op (0,1) volgt uit (7.59) met behulp van (7.3) en (7.53)

(7.60) $$\lim_{s\downarrow 0} K_s^{-1}\left(sy + J(q) + \delta(s)\right) = F_M^{-1}\left(1 - e^{-J(q)}\right) = m(q)$$

Met behulp van de LDC-stelling kan worden geverifieerd dat de limiet s ↓ 0 in (7.58) onder het integraalteken mag worden genomen, zodat uit (7.60) volgt, omdat k(m) continu is, dat (7.57) moet gelden. Hiermee is stelling A.7.5 bewezen.

Bewijs Stelling A.7.6 Om (7.9) en (7.10) te bewijzen laten we eerst de volgende opmerkingen vooraf gaan. De drager $D_{Q,M}$ van $g_s(q,m)$ werd gegeven in (6.9). Omdat deze drager van s afhangt zullen we deze hier aangeven met $D_{Q,M;s}$. Met enig schrijfwerk, dat aan de lezer wordt overgelaten, kan worden aangetoond dat

(7.61) $$D_{Q,M;s} \subset D_Q \, x \, D_M$$

Dit resultaat is overigens nogal plausibel. Als $(q,m) \in D_{Q,M;s}$ is $g_s(q,m) > 0$ en ligt het voor de hand dat zowel g(q) als g(m) positief zijn, zodat dan $(q,m) \in D_Q \, x \, D_M$. In het bijzonder volgt uit (7.61) dat

(7.62) $$g_s(q,m) = 0 \quad \text{voor } (q,m) \text{ met } q \in D_Q \text{ en } m \notin D_M$$

Beschouw om (7.9) te bewijzen een vaste $q \in D_Q$. We brengen in herinnering dat $D_M = (m_b, m_e)$. Neem eerst aan dat m_b en m_e beiden eindig zijn. Met behulp van (7.62) kan eenvoudig worden aangetoond dat voor $m \le m_b$ dan geldt $F_s(m|q) = 0$ en voor $m \ge m_e$ dan geldt $F_s(m|q) = 1$. Dus voor dergelijke m is aan (7.9) voldaan. We hoeven (7.9) dus slechts te bewijzen voor $(q,m) \in D_Q \, x \, D_M$. Het is eenvoudig in te zien dat we ons bij een naar links en/of rechts onbegrensde D_M eveneens mogen beperken tot $(q,m) \in D_Q \, x \, D_M$. Uit (7.61) volgt dat we ons in het bewijs van (7.10) eveneens mogen beperken tot dergelijk *(q,m)*: indien $(q,m) \notin D_Q \, x \, D_M$ zijn zowel het linker- als het rechterlid van (7.10) triviaal gelijk aan 0.

Beschouw nu een vaste $(q,m) \in D_Q \, x \, D_M$. Dan zijn $K_s(m)$ en $J(q) > 0$ goed gedefinieerd. Volgens (7.13) uit hoofdstuk 7 en Lemma A.7.4 volgt dan, omdat Λ_1 continu is

(7.63) $$\lim_{s\to\infty} F_s(m \mid q) = \lim_{s\to\infty} \Lambda_1\left(\frac{K_s(m) - J(q) - \delta(s)}{s}\right) = F_M(m)$$

waarmee (7.9) bewezen is. Neem om (7.10) te bewijzen aan dat $(q,m) \in D_Q \, x \, D_M$ met $m \neq m_i$, i = 1, 2, …, n. Er volgt uit (6.5) en (7.1) en (7.12) uit hoofdstuk 7 dat (bedenk $g(s,y) = g_{Y,s}(y)$)

(7.64) $$g_s(m \mid q) = \frac{\lambda_1\left(\frac{K_s(m) - J(q) - \delta(s)}{s}\right) g(m)}{s\, g(s, K_s(m))}$$

Uit de definitie van g(s,y) volgt direct dat

(7.65) $$sg(s, K_s(m)) = \int_0^\infty e^{-x} \lambda_1\left(\frac{K_s(m) - x - \delta(s)}{s}\right) dx$$

Omdat λ_1 begrensd is kan met de LDC-stelling dan geverifieerd worden, met behulp van Lemma A.7.4, en omdat λ_1 continu is in het punt $\Lambda^{-1}(F_M(m))$

$$(7.66)\qquad \lim_{s\to\infty} sg(s,K_s(m)) = \int_0^\infty e^{-x}\lambda_1\left(\Lambda_1^{-1}(F_M(m))\right)dx$$
$$= \lambda_1\left(\Lambda_1^{-1}(F_M(m))\right)$$

Door nogmaals gebruik te maken van Lemma A.7.4 volgt dan uit (7.64) dat (7.10) moet gelden. Hiermee is de stelling bewezen.

Bewijs Stelling A.7.7 Volgens (7.1), (7.35) en (7.36) uit hoofdstuk 7 kan voor $L_{tr}(s)$ worden geschreven

$$(7.67)\qquad L_{tr}(s) = \prod_{i=1}^{N}\left\{\frac{1}{s}e^{-x_i}\lambda_1\left(\frac{K_s(m_i)-x_i-\delta(s)}{s}\right)\right\}$$

Omdat λ_1 begrensd is geldt voor een zekere C dat $|\lambda_1| < C$, zodat de i–de factor uit het produkt in het rechterlid wordt begrensd door $Cexp(-x_i)/s$. Dan volgt onmiddellijk dat (7.11) moet gelden. Om (7.12) te bewijzen merken we eerst op dat volgens Lemma A.7.2 en (6.15) en (7.34) uit hoofdstuk 6 en 7 geldt

$$(7.68)\qquad \lim_{s\downarrow 0} K_s(m_i) - x_i - \delta(s) = -\ln\left(1-F_M(m_i)\right) + \ln\left(1-F_Q(q_i)\right) \neq 0$$

waarbij het ongelijkteken in het laatste lid volgt uit het feit dat bij aaname (q_i, m_i) niet op de lijn van gelijke kansen ligt. Schrijf nu

$$(7.69)\qquad t_i(s) = K_s(m_i) - x_i - \delta(s)$$

Dan volgt $|t_i(s) / s| \to \infty$ als $s \downarrow 0$ zodat voor s voldoende klein uit voorwaarde (3.8) uit hoofdstuk 3 volgt, voor $c > 1$,

$$(7.70)\qquad \frac{1}{s}\lambda_1\left(\frac{t_i(s)}{s}\right) = \frac{K_s^{\,c-1}}{\left(t_i(s)\right)^c}$$

Uit (7.68) volgt dat het rechterlid naar 0 gaat als $s \downarrow 0$, zodat

$$(7.71)\qquad \lim_{s\downarrow 0}\frac{1}{s}\lambda_1\left(\frac{t_i(s)}{s}\right) = 0$$

Uit (7.67) volgt dan dat (7.12) moet gelden.

We dienen nog te laten zien dat voor een $s_0 \in (0,\infty)$ geldt dat $L_{tr}(s_0) > 0$. Daartoe merken we op dat volgens Lemma A.7.4 voor een voldoend grote $s_0 > 0$ moet gelden dat $t_i(s_0)/s_0$ tot de drager van λ_1 behoort voor alle i = 1, 2, …, n. Uit (7.67) volgt dan onmiddellijk $L_{tr}(s_0) > 0$. Alle beweringen uit punt (1) van de stelling zijn hiermee bewezen.

Punt (2) van de stelling is eenvoudig te bewijzen. Neem aan dat λ_1 continu is. Dan is $L_{tr}(s)$ continu op $s > 0$. Vanwege (7.11) en (7.12) geldt dan dat $\sup\{L_{tr}(s) \mid s > 0\}$ eindig is. Tevens volgt dat voor een interval $[s_1,s_2]$ met $s_1 > 0$ dit supremum gelijk is aan $\sup\{L_{tr}(s) \mid s \in [s_1, s_2]\}$. Omdat een continue functie zijn supremum aanneemt op een compact (dat wil zeggen gesloten en begrensd) interval is er dus een $s' \in [s_1, s_2]$ waarvoor $L_{tr}(s') = \sup\{L_{tr}(s) \mid s \in [s_1, s_2]\}$. Deze $L_{tr}(s')$ is dan gelijk aan het maximum van $L_{tr}(s)$ op $s > 0$.

Referenties

[Apostol, 1967]
Calculus – Volume 1, second edition. T. M. Apostol. John Wiley & Sons, New York, 1967.

[Beijk en Geerse, 2004]
Bivariate correlatiemodellen gebaseerd op exponentiële marginale verdelingen. Beschrijving en werking MATLAB-programmatuur. V.A.W. Beijk en C.P.M. Geerse. RIZA – werkdocument 2004.109X, mei 2004.

[Billingsley, 1995]
Probability and Measure. P. Billingsley. John Wiley & Sons, New York, 1995.

[Cassella, 1990]
Statistical Inference. G. Cassella, R.L. Berger. Duxbury Press, Belmont, California, 1990.

[Ditlevsen en Madsen, 1996]
Structural reliability analysis. O. Ditlevsen, H.O. Madsen. John Wiley & Sons, Chichester, England 1996.

[Geerse et al, 2002]
Wind-waterstandstatistiek Hoek van Holland. , C.P.M. Geerse (RIZA), M.T. Duits (HKV), H.J. Kalk (HKV), I.B.M. Lammers, (HKV). RIZA/HKV rapport, Lelystad, juli 2002.

[Geerse, 2003a]
Probabilistisch model voor de IJsseldelta. C.P.M.Geerse. RIZA – werkdocument 2003.091X. RIZA Lelystad, mei 2003. (Het betreft een nauwelijks gewijzigde definitieve versie van het gelijknamige concept van 27 juni 2000.)

[Geerse, 2003b]
Probabilistisch model hydraulische randvoorwaarden IJssel- en Vechtdelta. C.P.M.Geerse. RIZA – werkdocument 2003.129X. RIZA Lelystad, september 2003.

[Marsden en Tromba, 1996]
Vector Calculus. (Fourth edition.) J.E. Marsden, A.J. Tromba. W.H. Freeman and Company, New York, 1996.

[Volker, 1987]
Statistiek van wind en waterstanden in Hoek van Holland; tweede concept. W.F. Volker. 20 mei 1987.

[Vrouwenvelder *et al*, 1999]
Theoriehandleiding PC-Ring; deel B: Statistische modellen, 3e concept. A.C.W.M. Vrouwenvelder, H.M.G.M. Steenbergen, K.Slijkhuis. TNO-rapport 98-CON-R1431. TNO-Bouw, 31 januari 1999.