KM34(1990)
pag 5 - 27

DATATHEORIE

een geometrische theorie omtrent modellen voor de analyse van variantie, die zelf gebaseerd is op invariantie

REDE

uitgesproken bij de aanvaarding van het
ambt van hoogleraar in de datatheorie
aan de Rijksuniversiteit te Leiden
op vrijdag 16 februari 1990

door

DR. W.J. HEISER

[cette science], s'arrête et se fonde sur la véritable méthode de conduire le raisonnement en toutes choses, que presque tout le monde ignore, et qu'il est si avantageux de savoir, que nous voyons par expérience qu'entre esprits égaux et toutes choses pareilles, celui qui a de la géométrie l'emporte et acquiert une vigueur toute nouvelle.

Pascal, *De l'esprit géométrique*

Mijnheer de Rector Magnificus,

Zeer gewaardeerde toehoorders,

Zoals tien schilders van één model tien afbeeldingen maken, zo beelden tien wetenschappers één werkelijkheid in tien modellen af[1]. Deze bontheid van het wetenschappelijk handwerk vormt de voedingsbodem en het bestaansrecht van de datatheorie, een theorie van modellen. Voor diegenen van u, voor wie de datatheorie tot op heden een gesloten boek was, zal ik trachten vanmiddag een tip van de sluier op te lichten. Voor diegenen die reeds bekend zijn met het gebied, zal ik een paar nieuwe gezichtspunten en begrippen introduceren, die de aard en de positie van de datatheorie verduidelijken.

Punt één is dat de theorie zich niet bezighoudt met data analyse zelf, maar met de modellen en technieken die voor dat doel ontworpen zijn.

Dergelijke modellen en technieken zijn er in vele soorten en maten, maar niet alle vallen ze onder het bereik van de datatheorie. De fundamentele reden hiervoor is, dat om tot een theorie te komen een of ander mechanisme gekozen moet worden dat geabstraheerd wordt, en dat inderdaad verscheidene van die mechanismen onafhankelijk uit te werken zijn. Dit leidt dan soms tot modellen die specifiek onder één theorie vallen, alhoewel – zoals we zullen zien – dit niet noodzakelijk is. Laat ik eerst een contrast schetsen met zo'n andere theorie, binnen het vak waarmee de datatheorie een ambivalente relatie onderhoudt – de statistiek.

Veel, hoewel zeker niet alles, van wat traditioneel tot de statistiek gerekend wordt, is gebaseerd op het mechanisme van de *regressie*. Deze ietwat mistroostige, Darwinistische term herinnert de statisticus er keer op keer aan dat zonen van knappe vaders in 't algemeen een gemiddeld IQ hebben dat lager ligt dan dat van hun vaders. Het basis idee van de regressie theorie is, dat we empirische observaties op *a priori* gronden kunnen indelen in twee of meer klassen, eventueel nog van nadere structuur voorzien, en dat we vervolgens de verdeling *tussen* klassen kunnen vergelijken met de verdeling *binnen* klassen. Variatie tussen klassen is zo per definitie systematische variatie, omdat de klassen a priori gekozen worden en dus op een hypothetische systematiek moeten berusten, terwijl de variatie binnen klassen per definitie als willekeurige variatie wordt aangemerkt. Op dit idee van het vergelijken van "tussen-" en "binnen-" verdelingen stoelen de meeste statistische toetsen, de variantie

analyse, de isotone regressie, gegeneraliseerde lineaire modellen, en zo verder, met als belangrijke uitzondering de discriminant analyse. In de discriminant analyse is ook sprake van een leidinggevende indeling van de observaties in klassen, maar deze is niet geheel *a priori*, en daarom vormt de discriminant analyse een buitenbeentje binnen de klassieke statistiek. Ik zal daar nog op terugkomen.

In de datatheorie is het grondmechanisme dat geabstraheerd wordt een proces van *afstemming*. Aan de theorie kunnen drie facetten onderscheiden worden:

(1) wat moet op wat afgestemd worden?
(2) wat is het medium waarin zich het proces afspeelt?
(3) in welke zin kan er zoal afstemming worden nagestreefd?

Data vormen het eerste facet.

Data

De datatheorie ontleent haar naam aan deze technische term, die in een veel striktere zin begrepen moet worden dan wat het woord in andere wetenschapsgebieden of in het dagelijks leven zoal kan betekenen. Data zijn datgene wat afgestemd wordt, dat wat geanalyseerd wordt; een abstractie dus, die niet samenvalt met dat wat waargenomen is[2], of dat wat bekend is.

De laatste betekenis van het woord data, dat wat bekend is, vindt zijn oorsprong in het bijhouden van een dagtekening, en verwijst naar het idee dat kennis uit elementen bestaat, van welke aard dan ook, die kunnen worden gesorteerd in de tijd. Als we een logboek bijhouden, sorteren we immers feiten, gebeurtenissen, en invallen door er een "datum" bij te zetten, en dit is wezenlijk ook de betekenis van de term "data" die we uit het computerjargon kennen. Deze boekhoudkundige betekenis doet de vraag opkomen of het wellicht mogelijk is om ooit al onze kennis op één standaard manier vast te leggen. Met andere woorden, is een *universele bibliotheek* mogelijk?

De Amerikaanse logicus Willard Quine heeft er op gewezen[3], dat er aan de gedachte van een universele bibliotheek, die "weemoedige fantasie", geopperd door onder anderen de 19e-eeuwse grondlegger van de psychophysica Theodor Fechner, een ultieme absurditeit kleeft. In de universele bibliotheek zouden alle mogelijke boeken staan, in allerlei talen en over allerlei onderwerpen, met begrijpelijke teksten en met onbegrijpelijke teksten, over

dingen die waar zijn en dingen die niet waar zijn; dit alles in één uniform alfabet. Maar zou dit dan niet een onmogelijke, oneindige bibliotheek worden? Het antwoord is: neen. Al naar gelang het aantal letters in het gekozen alfabet kan men voorrekenen wat de bovengrens van het totaal aantal boeken is dat geschreven kan worden door het steeds anders combineren van de verschillende letters uit het alfabet, en hoewel het om een astronomisch groot aantal mogelijke boeken zou gaan is het interessante, dat de universele bibliotheek wel degelijk een *eindig* aantal boeken zou hoeven te bevatten, omdat die bovengrens eindig is.

Dus tussen wartaal en onzin door zou ook de hele waarheid, voor zover die althans in woorden kan worden uitgedrukt, uiteindelijk in Fechner's bibliotheek te vinden zijn! Ook al zouden de boeken de gebruikelijke lengte van enige honderden bladzijden hebben, dan nog zou steeds een ander boek de loop van een redenering weer oppakken; dus door voortdurend een ander boek ter hand te nemen zouden we in staat zijn alles wat gekend kan worden te achterhalen. Bij onze speurtocht naar de waarheid verkeren we weliswaar in het onzekere over welk boek we het eerst moeten pakken, en welk daarna, maar in principe stáát alles er.

Merk evenwel op dat als het alfabet vast is, en als het herhaald opslaan van boeken is toegestaan, dat dan een flinke bezuiniging op deze plannen mogelijk is. Immers, dank zij Morse code en computer weten we allemaal dat een alfabet van slechts twee tekens, een punt en een streep, of een nul en een één, voldoende is om een willekeurige boodschap te coderen. En als we toch bereid zijn steeds een ander boek op te slaan, dan kan volstaan worden met de aanschaf van slechts twéé boeken, het ene met uitsluitend nullen erin, het andere met uitsluitend enen erin. Deze aanschaf moet ruimschoots voldoen aan de behoeften van de jonge, energieke wetenschapper die dan tot in lengte van dagen alternerend het boek der enen en het boek der nullen ter hand zal kunnen nemen.

Uit dit mislukte wonder van de eindige, maar universele bibliotheek leren we dat er meer structuur nodig is om zinvol over data te kunnen theoretiseren. We zagen al dat de regressie theorie structuur aanbrengt door een *a priori* indeling van de observaties als extra "gegeven" te veronderstellen. In de datatheorie wordt structuur aangebracht door zich te richten op bepaalde andere patronen van koppeling van de observaties. De data zijn nooit concrete, enkelvoudige elementen, zoals het aanwezig of afwezig zijn van een

attribuut, of een enkel getal dat via een meetinstrument is geregistreerd. In de datatheorie betreft het altijd een *gestructureerde serie* complexe elementen, zoals bijvoorbeeld meerdere classificaties van eenzelfde verzameling objecten, of verschillende rangordeningen van paren van objecten, of een aantal aan elkaar gekoppelde lijsten van getallen. Die patronen van koppeling maken de data *relationeel.*

Over zo'n gestructureerde serie elementen kan worden geredeneerd alsof het om relaties tussen punten, assen, bollen, loodlijnen, krommen en vlakken gaat. Dit wordt gedaan op basis van een centraal *equivalentie principe*, dat inhoudt dat het hetzelfde is om de empirische relaties tussen empirische objecten in de meetkundige ruimte af te beelden in hun volle, laten we zeggen chaotische vorm – dus vóór afstemming – , als om deze relaties af te beelden in een afgestemde vorm. Dit maakt de datatheorie tot een *geometrische theorie*, met een bijzondere aandacht voor invarianties.

Wat zijn invarianties?

Met invariantie wordt aangeduid dat een eigenschap hetzelfde blijft wanneer een bepaalde verandering wordt aangebracht aan de drager van de eigenschap.

In de meetkundige ruimte zijn allerlei typen van invariantie te onderscheiden. Een eenvoudig voorbeeld is de afstand tussen twee punten. Afstand is invariant onder rotaties, reflecties en translaties. Dit wil zeggen dat op de plattegrond van de stad Leiden de afstand tussen bijvoorbeeld de Academie en de Pieterskerk niet verandert, hoe ik de plattegrond ook draai of keer, omhoog schuif of opzij. Een ander voorbeeld van invariantie treedt op bij de projectie van een bol op een vlak. Op iedere dia-avond is er altijd wel iemand die dit empirisch aantoont door een vuist of een hoofd in de lichtbundel tussen projector en projectiescherm te steken. Uit deze ervaring weten we dat, hoe de bol ook tussen lichtbron en scherm wordt gedraaid, het resultaat van de projectie onveranderlijk een cirkelvormige vlek is. Uit de vervolgens vaak losbarstende experimenteerlust van de andere aanwezigen leren we dat de projectie van een plat voorwerp, daarentegen, wel van vorm kan veranderen wanneer z'n oriëntatie verandert.

Door het equivalentieprincipe kunnen we dus redeneren over data alsof we het over meetkundige objecten hebben, en vervolgens kunnen we het proces van

afstemming voorstellen als het uitvoeren van meetkundige operaties die bepaalde, geselecteerde eigenschappen invariant laten. Dit is het tweede facet aan de theorie van het afstemmingsproces: zij voltrekt zich in de meetkundige ruimte, en geeft een belangrijke rol aan concepten als afstand, projectie, rotatie, rechte hoek en kortste pad.

Het derde facet was: wat voor type afstemming wordt er nagestreefd? Ik noem nu vier prominente typen, zonder volledig te kunnen zijn. Ten eerste is er de overgang naar *canonische vorm*, of het vinden van invariante richtingen.

Wanneer we data in canonische vorm brengen komt dit er meestal op neer dat we bepaalde richtingen in de ruimte opzoeken die in één of ander opzicht *bijzonder* zijn. Een klassiek geval is de analyse van schoolprestaties, of – iets algemener – de analyse van psychologische testgegevens. Stel we nemen proefwerk A af; dit ordent de kinderen van dom naar knap op z'n A's. Nu nemen we proefwerk B af; dit ordent dezelfde kinderen van dom naar knap op z'n B's. En zo verder, voor proefwerk C, D, E etcetera. Dit is al een jaar of zeventig een onuitputtelijke bron van data!

Elk proefwerk wordt abstract weergegeven door een pijl. Twee pijlen wijzen meer in dezelfde richting naarmate de twee proefwerken de kinderen op meer overeenkomstige wijze ordenen. Nu kunnen allerlei concrete probleemstellingen geabstraheerd worden. Tapt ieder proefwerk steeds dezelfde vaardigheid af? Dit wordt: wijzen alle pijlen in dezelfde richting? Hoe sterk is A (vaardigheid in het Frans) gerelateerd aan B (vaardigheid in het Engels), gecorrigeerd voor de verschillen in C (vaardigheid in het Nederlands)? Dit wordt: wat is de hoek tussen A en B als je alleen maar parallel aan de richting van C mag kijken? Vervolgens kunnen die geometrische probleemstellingen doorvertaald worden naar kenmerken van operaties in de ruimte. Uiteindelijk komen we dan uit bij het opsporen van invariante richtingen, dat zijn richtingen die de opeenvolgende assen van een ellipsoïde vormen. Omdat de ellipsoïde inmiddels op een groot aantal manieren aan de data geliëerd kan zijn, wordt zo een breed scala van analysemogelijkheden gedekt, bekend onder de verzamelnaam *multivariate analyse*[4].

Een tweede type afstemming bestaat uit het vaststellen van *parallellie* van curven. Een bekend voorbeeld is opnieuw de analyse van de antwoorden op vragen uit een vorderingentest, of van examenresultaten, maar nu op micronivo. De data zijn hier een aantal lijsten met {"goed", "fout"} patronen.

Dergelijke lijsten worden vaak geanalyseerd door ze af te zetten tegen een serie parallelle, *stijgende* curven, dat wil zeggen curven die allemaal gelijk zijn aan één enkele stijgende curve, op een verschuiving na. Hiermee worden degenen die de test hebben afgelegd, gemarkeerd langs een lijn; de stijgende curve zegt dat naarmate iemand meer aan de rechterkant van de lijn ligt, de kans groter is om een vraag "goed" te beantwoorden, en de verschuiving van de curven geeft het *faseverschil* tussen de vragen weer: de vragen links zijn gemakkelijker dan de vragen rechts[5].

Een ander voorbeeld in deze klasse wordt gevormd door de analyse van oordelen over de verdedigbaarheid van controversiële uitspraken, of van voorkeuren voor politieke candidaten. Dit soort oordelen komt vaak in aanmerking voor afstemming met behulp van parallelle *enkeltoppige* curven, dat wil zeggen curven die op een verschuiving na gelijk zijn aan één enkel silhouet van een ministerssteek.

Nu zijn het de controversiële uitspraken (of de politieke candidaten) die worden gemarkeerd langs een lijn. De enkeltoppige curve zegt dat er een bepaalde aangrenzende groep uitspraken (of candidaten) is waar een beoordelaar het mee eens is, terwijl betreffende beoordelaar het niet eens is met wat daar links of rechts van ligt. Het faseverschil geeft nu uitdrukking aan het feit dat sommige beoordelaars een "linkser", en andere beoordelaars een "rechtser" standpunt hebben. Precies eenzelfde model wordt trouwens gebruikt voor de groei van verschillende plantensoorten langs wat in dat geval een *ecologische gradient*[6] heet.

Parallellie is het wezenskenmerk van de zogenaamde *één-dimensionale schaaltechnieken*. Voor de zekerheid merk ik nog op dat de term parallellie ook voorkomt als hoeksteen van de klassieke testtheorie, waar men spreekt over parallelle tests indien de test-karakteristieke curven identiek zijn, of althans niet ten opzichte van elkaar verschoven zijn. De kwalificatie "parallel" refereert hier uitsluitend aan het feit dat de ene test in reserve gehouden kan worden om de andere test te vervangen, aangezien beide binnen de grenzen van hun betrouwbaarheid dezelfde scores opleveren. De klassieke testtheorie valt grotendeels buiten het bestek van de datatheorie; zij is sterker geliëerd aan de regressietheorie.

Het derde type afstemming is het vinden van een *inbedding*. Een inbedding is een overgang van de ene ruimte naar een andere ruimte, zodanig dat bepaalde

afstandsrelaties invariant blijven. De reden dat een inbedding gezocht wordt is soms dat de tweede ruimte gemakkelijker te inspecteren is, of door minder parameters gekenmerkt wordt, dan de eerste. Maar het kan ook zijn dat de tweede ruimte een speciale karakteristiek heeft waarin men is geinteresseerd. Dé klassieke inbedding van gelijkenis data is inbedding in de Euclidische, dat wil zeggen de huis-tuin-en-keuken ruimte. Maar ook andere ruimtes komen in aanmerking.

De meetkunde kent vele ruimtes

Omdat de problematiek van de inbedding naar mijn mening een centrale plaats inneemt in de datatheorie, en er tevens veel misverstanden over de ronde doen, wil ik er iets langer bij stil blijven staan. Er is sprake van gelijkenis data wanneer we beschikken over een lijst van paren van objecten, en we van ieder paar kunnen zeggen of het meer of minder onderlinge gelijkenis vertoont dan één van de andere paren; soms ook nog hoeveel meer of minder. Een voorbeeld van dit soort data zou zijn: empirische evidentie dat "zebra" meer op "paard" lijkt dan "lam" lijkt op "leeuw", en zo verder voor andere paren van paren.

Er is nu een stelling die zegt dat vier objecten altijd in het platte vlak ingebed kunnen worden zodanig dat de ordening naar gelijkenis tussen objecten perfect overeenkomt met de ordening naar afstand tussen de punten in de inbedding. Voor vijf objecten lukt het altijd in drie dimensies, voor zes objecten in vier dimensies, en zo verder. Met andere woorden, er is altijd een inbedding van gelijkenis data in één of andere ruimte, hoe groot van dimensie of eventueel vreemd van structuur ook[7].

Nu is het natuurlijk zo dat de Natuur ons pas echt iets meedeelt als we niet vijf, maar *vijfhonderd* objecten in drie dimensies kunnen inbedden, of als de objecten in goed gesepareerde groepen uiteenvallen, of – meer in 't algemeen – als de gelijkenis data bepaalde uitgesproken kenmerken hebben. Deze extra kenmerken leiden tot een rijke familie van afstandsmodellen[8], en daarmee tot een groot aantal speciale inbeddingsproblemen, omdat bepaalde soorten kenmerken een ideaal-typisch karakter hebben gekregen. Ik noem, naast de standaardgevallen van multidimensional scaling, de volgende voorbeelden:

- Wat is de beste *ultrametrische boom* die bij de gelijkenis data past? Dit is het probleem van de hiërarchische cluster analyse.
- Wat is het beste *regelmatig veelvlak* dat bij de gelijkenis data past? Dit is het probleem van de indeling in *k* groepen.
- Wat is de beste lokalisering van de objecten *langs een lijn*? Dit is het seriatieprobleem.
- Wat is de beste lokalisering van de objecten op een serie *concentrische cirkels?* Dit is een vorm van scalogram analyse van partiële ordeningen.
- Wat is de beste allocatie in een *city-block metriek*? Bijvoorbeeld, hoe plaats ik de afdelingen van een organisatie in de kamers van een gebouw zodanig dat de afstanden, langs de gangen gemeten, zo goed mogelijk bij een gewenst patroon van interactie past?

Dit laatste voorbeeld laat zien dat inbedding naast haar theoretische functie ook een heel praktische achtergrond kan hebben. In alle gevallen kunnen we ons voordeel doen met stellingen uit velerlei delen van de wiskunde om deze problemen te systematiseren en op te lossen.

Laat ik nu één van die misverstanden bespreken waar ik zoëven van repte, die over hiërarchische cluster analyse. De bron van de verwarring is de stelling van Holman[9]. Deze zegt dat gelijkenis data die aan de ultrametrische ongelijkheid voldoen, uitsluitend perfect in de Euclidische ruimte kunnen worden ingebed als deze van maximale dimensionaliteit gekozen wordt. Hieruit wordt nogal eens als practische consequentie de conclusie getrokken dat een goed passende inbedding in een ultrametrische boom en een goed passende inbedding in een laag-dimensionale ruimte elkaar uitsluiten. Dit is echter niet waar.

Bij een ultrametrische boom kunt u zich het best een stamboom voorstellen. Er is nu een stelling die ik samen met Frank Critchley heb opgesteld die er op neer komt dat een stamboom wel degelijk in een laag-dimensionale Euclidische ruimte kan worden afgebeeld, jazelfs dat dit kan langs een lijn[10]. Het bewijs dat dit zo is berust op twee elementaire inzichten. Ten eerste is een boom *zelfgelijkvormig*, dat wil zeggen dat een onderdeel van een boom, zoals een tak, alle eigenschappen van een boom heeft. Zelfgelijkvormigheid maakt een vereenvoudiging van het bewijs mogelijk: we kunnen ons eigenlijk tot één vertakkingssituatie van vier punten beperken. Ten tweede kunnen we er van uitgaan dat gelijke data waarden, zgn. "ties", in de inbedding niet per se tot gelijke afstanden zouden hoeven te leiden, zolang het maar zo is dat kleinere

gelijkenis tot grotere afstand leidt. Dit principe staat bekend onder de naam *primary approach to ties*.

Welnu, als we de stam van de boom nemen, waar de objecten voor het eerst in twee groepen worden opgesplitst, laten we zeggen de Friezen en de Hollanders, dan hoeven we er alleen maar voor te zorgen dat de kleinste afstand tussen een Fries en een Hollander groter is dan de grootste afstand tussen een willekeurige Fries en een andere Fries, of een willekeurige Hollander en een andere Hollander. Langs een lijn is dit altijd mogelijk door de Friezen ver genoeg naar links, of de Hollanders ver genoeg naar rechts te verschuiven. Dankzij zelfgelijkvormigheid geldt precies hetzelfde argument binnen de Friese tak en binnen de Hollandse tak van de boom, en dankzij de "primary approach to ties" is het niet noodzakelijk alle Friezen boven op elkaar te leggen om aan de ultrametrische ongelijkheid te voldoen. We kunnen ze, als 't moet, *willekeurig dicht* bij elkaar leggen, zoals dat in de wiskunde heet, wat toch iets anders is dan òp elkaar.

Er is dus helemaal geen grote tegenstelling tussen een ultrametrische boom structuur en lage dimensionaliteit in een continue ruimte, zoals in de praktijk trouwens ook vaak blijkt. We moeten ons de dimensie alleen anders voorstellen dan gebruikelijk is bij het denken in hiërarchiën. Tussen twee haakjes merk ik nog op dat zelfgelijkvormigheid een interessant type invariantie vertegenwoordigt, namelijk invariantie van structuur onder verandering van schaal. In dit opzicht lijkt de boom op een spiraal. Een spiraal waar we op "inzoomen" blijft een spiraal.

Een tweede veel gehoord misverstand dat over inbedding bestaat is het idee dat gelijkenis data dan misschien wel fundamenteel of elegant zijn, maar dat ze in de harde praktijk van het empirisch onderzoek te ingewikkeld zijn, of te duur om te verzamelen. Daarom zouden inbeddingstechnieken thuis horen bij de curiosa; interessant voor liefhebbers, maar dan ook voor hen alleen. Tegen deze misvatting is zóveel in te brengen, dat ik er nog regelmatig door in verwarring wordt gebracht, niet wetend waar de bestrijding te beginnen. Laat ik me daarom hier beperken tot één tegenargument, en dat is dat veel data analyse technieken ook – dus niet alleen, maar óók – kunnen worden gezien als oplossingen van een inbeddingsprobleem, als men zich maar rekenschap geeft van het feit dat de gelijkenissen *afgeleid* kunnen zijn van de observaties, en daarom zelf niet empirisch verzameld hoeven te zijn.

Welke sociale wetenschapper is niet bekend met het histogram[11]? Een histogram is een tekening van gediscretiseerde observaties; het geeft het resultaat weer van een discretisatie probleem. Dit discretisatie probleem, op haar beurt, is een speciaal geval van een cluster analyse probleem, namelijk hoe de observatiewaarden in *k* groepen te verdelen zodanig dat de maximale afstand binnen groepen een bepaalde drempelwaarde niet te boven gaat. En een cluster analyse probleem is een inbeddingsprobleem, waarbij hier de rol van gelijkenissen wordt gespeeld door de onderlinge afstanden langs de schaal van de observaties.

Een iets ingewikkelder, maar toch eigenlijk nog steeds zeer basaal geval betreft de *homogeniteitsanalyse* van een kruistabel die de associatie tussen twee categorische variabelen vastlegt, bijvoorbeeld het "beroep van de moeder" en het "beroep van de dochter" in een populatie van families. Natuurlijk kan men ook vele andere dingen doen met zo'n kruistabel, en binnen dat wat bekend staat als homogeniteitsanalyse, of "l'analyse des correspondances", bestaan weer vele uitgangspunten om de techniek te beschrijven – veelal geliëerd aan het vinden van een bepaalde canonische vorm. Maar waar het om gaat is dat deze situatie óók beschreven kan worden als een *dubbel* inbeddingsprobleem, met als gelijkenis data een afgeleide maat voor de overeenkomst tussen ieder paar rijen van de tabel, en tussen ieder paar kolommen van de tabel.

Hierbij vergelijken we de verdeling over beroepscategoriën van alle dochters van, bijvoorbeeld, "agrariërs" met de verdeling van alle dochters van "lagere employés", en eisen van de inbedding dat het punt voor "agrariërs" dichter bij het punt voor "lagere employés" ligt naarmate die twee verdelingen meer op elkaar lijken. Dit principe bepaalt de inbedding van de moeders. Analoog bepalen we een inbedding van de beroepscategoriën van de dochters door steeds de verdeling van herkomst te vergelijken. Als er totaal geen sociale mobiliteit zou zijn, verwachten we inbeddingen te vinden in hoog-dimensionale ruimtes, met punten die op regelmatige veelvlakken liggen. Als de sociale mobiliteit zich voltrekt in kleine stapjes langs één beroepsvolgorde, moeten we een inbedding kunnen vinden in één dimensie. Er zijn ook nog allerlei interessante dingen te zeggen over de betekenis van bepaalde verschillen tussen de moeder-ruimte en de dochter-ruimte, maar dat zou ons bij deze gelegenheid te ver voeren.

Ik heb nu drie typen van afstemming besproken: overbrenging naar canonische vorm, het uitgangspunt voor de multivariate analyse; vaststelling

van parallellie van curven, het basismotief van de één-dimensionale schaaltechnieken; en inbedding van de ene ruimte in de andere, leidend tot multidimensional scaling, cluster analyse, en aanverwante afstandsproblemen. Het vierde, en laatste, type dat ik bespreek is in die zin van een iets andere aard, dat het gecombineerd kan worden met elk van de voorafgaande. Het betreft *optimal scaling*, een moeilijk in het Nederlands om te zetten term.

Optimale calibratie

Calibratie, afgeleid van kaliber, is een begrip uit de scheikunde, en betekent dat een glazen buis van merktekens wordt voorzien om eenheden van inhoud aan te geven. Calibreren betekent, meer specifiek, de onregelmatigheden van een buis onderzoeken alvorens de maatverdeling aan te brengen, en dit doel maakt het tot een treffend beeld voor optimal scaling. Als de buis naar beneden toe spits toeloopt, zal de afstand tussen de maatstreepjes steeds groter moeten worden om gelijke inhoud af te meten, maar hoe precies is niet altijd *a priori* bekend. Evenzo is het soms in de data analyse niet bekend of we een variabele in z'n oorspronkelijke eenheden, of in log-eenheden, of in nog andere, onregelmatige eenheden moeten uitdrukken.

Optimale calibratie betekent nu, dat de merktekens zó gekozen worden dat één van de afstemmingscriteria geoptimaliseerd wordt. In het sociale mobiliteits-voorbeeld zou dit kunnen betekenen dat we schaalwaarden voor de beroeps-categoriën van moeders en dochters willen vinden die optimaal zijn in de zin van *maximale correlatie*. Maar evengoed kunnen we schaalwaarden voor de beroepscategoriën van moeders en dochters willen vinden die de families het best onderscheiden naar religieuze gezindte, of schaalwaarden die daar juist voor gecorrigeerd zijn. Het is dus altijd: optimaal met betrekking tot een bepaald criterium, en daarom lijkt optimal scaling op een ijkingsproces.

Er kunnen meer of minder restricties aan de calibratie worden opgelegd, al naar gelang een conditie die met een zeer ongelukkige term wordt aangeduid als het *meetnivo* van een variabele. Uiteraard denkt iedere oningewijde dat met "meetnivo" iets wordt bedoeld dat volgt uit de wijze waarop een variabele gemeten is, en dit is ook de betekenis van het woord binnen de *meettheorie*, maar niet de betekenis die bedoeld wordt binnen de optimal scaling. Stel dat uit vorig onderzoek, of uit de sociologische theorie, al numerieke waarden voor beroepscategoriën bekend zijn. Dan kunnen we de optimale calibratie in

die zin bepalen dat de nieuwe merktekens gelijk moeten zijn aan de oude, op een vermenigvuldiging met een schaalfactor na. Of we kunnen – liberaler – de optimale calibratie beperken door te eisen dat de nieuwe merktekens alleen *in volgorde* gelijk moeten zijn aan de oude. Het moge duidelijk zijn dat zo'n restrictie zelf niet uit de metingen volgt, maar gebaseerd is op *geaccumuleerde* kennis, en dat het beter ware om te spreken van *maatnivo*, of *calibratienivo*, dan over meetnivo.

De kritische toehoorder zal nu opmerken: 'ja, daar zie je dat de vergelijking met calibratie mank gaat, want bij nominale variabelen willen we zelfs geen volgorde aannemen, om niet-monotone verbanden op te sporen. Naar analogie zou je dus zoiets als negatieve inhoud moeten hebben om op die buis een maatverdeling te krijgen die niet-monotoon is met de natuurlijke ordening van de getallen'. Het antwoord op deze tegenwerping is als volgt. Nominaal maatnivo impliceert dat we de buis vullen met standaardeenheden van verschillende vloeistoffen waarvan we het soortelijk gewicht niet kennen. Hierdoor kan een later ingeschonken vloeistof *onder* een eerder ingeschonken vloeistof terechtkomen. Wist u dat op dit principe een speciaal soort cocktail is gebaseerd, de zogenaamde "Pousse Café"? Hierbij drijven dranken van verschillende kleur onder elkaar. Wat dacht u van het volgende recept:

Havana Rainbow

1/7 Grenadinesiroop
1/7 Anisette
1/7 Parfait Amour
1/7 Crème de Menthe
1/7 Orange Curaçao
1/7 Groene Chartreuse
1/7 Jamaica Rum

Schenk de ingrediënten in de gegeven volgorde in een likeurpijpje.

Als we het dus niet in de gegeven volgorde doen, is het idee, dan vormt zich na enige tijd toch dezelfde regenboog. Dit verschijnsel illustreert invariantie van de analyse onder niet-monotone transformaties.

De gedachte waarop optimale calibratie is gebaseerd is afkomstig uit de discriminant analyse. Eerder heb ik al opgemerkt dat deze techniek, die zich richt op het voorspellen van een classificatie variabele, een speciale plaats

inneemt in de klassieke statistiek. De reden is, grof gezegd, dat wanneer men aan een niet-numerieke variabele normaal verdeelde fouten toevoegt, het resultaat niet meer niet-numeriek is. Er moeten dus speciale modellen voor classificatiefouten opgesteld worden, en hiermee worden de grenzen van de normale regressietheorie overschreden. Inmiddels is het wel allemaal in orde gekomen, binnen het raamwerk van de logistische regressie[12]. Als we twee niet-numerieke response variabelen hebben, kunnen we proberen de groepen van de eerste variabele te discrimineren op basis van de tweede, en de groepen van de tweede variabele te discrimineren op basis van de eerste. Dit idee van een *wederkerige* discriminant analyse vormt een alternatieve rationale voor de afleiding van homogeniteitsanalyse, de optimale calibratietechniek pur sang.

Het wordt tijd om de balans op te maken

Ik heb u de essentie van de gedachtenwereld proberen uit te leggen op grond waarvan de datatheorie uitspraken doet over methoden van data analyse. Hoewel de theorie door haar meetkundig karakter een algemene strekking heeft, toont zich in de typen van afstemming die centraal zijn gezet het sociaal-wetenschappelijk erfgoed, en dit is iets om trots op te zijn. Het is een intrigerend fenomeen dat we vaak in andere disciplines – zoals de biologie en de economie – dezelfde technieken zien ontstaan, waarbij in een groot aantal gevallen geconstateerd kan worden dat de psychometrie en sociometrie voorop lopen in abstractie en degelijkheid[13].

Ik heb u voorgehouden dat de datatheorie uitspraken doet over modellen en methoden van data analyse, niet over de praktijk van de data analyse zelf. Wat moeten we de mensen vertellen dat ze moeten doen?

Dit is de vraag naar de statistische, of data analytische consultatie. Data analytische consultatie is mijns inziens een vak apart, waarvoor men op z'n minst een goede talenknobbel moet hebben, om zich in de Babylonische spraakverwarring een weg te banen. Een niet gering probleem is dat er tegenstellingen zijn gecreëerd die de verwarring alleen maar groter maken, zoals die tussen data analyse *versus* statistiek, of exploratie *versus* confirmatie. Zo speelde zich twee jaar geleden een opmerkelijk debat af tussen de Groningse hoogleraar Molenaar en de Leidse hoogleraar de Leeuw, onder de provocerende titel "Formeel Gronings of informeel Leids?". De beide opponenten bestookten elkaar met uitspraken als:

> 'Hoe kan een lezer de conclusies weerleggen of generaliseren die een Leidse priester trekt uit een plaatje dat eensklaps wordt voortgebracht door het orakel van Gifi?'. Aldus Molenaar[14].

> 'Als we de tijd en de gelegenheid hebben, kunnen we proberen deze persoon zorgvuldig uit te leggen dat de criteria die hij of zij gebruikt om de bruikbaarheid van data analyse technieken te evalueren niet op het terrein van de wetenschap liggen, maar beter thuishoren op godsdienstige bijeenkomsten'. Aldus de Leeuw[15].

Twee toonaangevende geleerden die elkaar bestrijden door de ander van bedenkelijke religieuze praktijken te beschuldigen: een waarachtig Hollands tafereel! Ik ben niet tegen polemiek, maar het lijkt mij dat de tegenstelling "formeel" versus "informeel" slechts op oppervlakkige stijlverschillen berust. Vooral echter lijkt het me een ernstige vergissing om statistische consultatie als een *ideologische* kwestie te behandelen.

Ik heb vanmiddag de datatheorie afgezet tegen de regressietheorie, niet met de bedoeling de één boven de ander te laten prevaleren, maar om duidelijke conceptuele lijnen te kunnen trekken. Veel technieken – zoals ze, om zo te zeggen, "in de natuur" voorkomen – dragen de sporen van beider invloed, en dat valt soms gelukkiger uit dan anders[16].

Ik heb ook de term "afstemming" geintroduceerd als het centrale thema waar de datatheorie zich mee bezighoudt. Niet geheel toevallig roept dit het beeld op van een data analysetechniek als een *radio*, die in principe van alles uit de ether kan opvangen, als je maar op de juiste wijze aan de knoppen draait. Uiteraard horen we alleen maar iets als er werkelijk een boodschap te ontvangen is, en mits we het frequentiebereik goed gekozen hebben. En natuurlijk moet de data analytisch adviseur in zijn verkenningsvaartuig niet volstaan met een radio. Als er vondsten afgewogen moeten worden komt de schaalbalans van de regressietheorie goed van pas; en voor het sturen zijn we blij de gyroscoop van de systeemtheorie aan boord te hebben. We kunnen ook van minder gespecialiseerde hulpmiddelen gebruik maken. In ieder geval lijkt consultatie mij een vak waarin de diagnoseproblematiek – d.w.z., wat wil de klant nu werkelijk weten? – centraal staat, en daarbij past een zo zakelijk mogelijke houding.

Wat we vanmiddag hebben gezien is dat het mogelijk is een theorie op te stellen omtrent modellen voor de analyse van variantie[17], die zelf gebaseerd is op invariantie. Deze voorliefde voor invarianties is een klassieke benaderingswijze in de wiskunde, maar het bijzondere is dat dit nu wordt toegepast op minder klassieke objecten, namelijk relationele datastructuren.

In het Leidse onderzoek zijn bijdragen geleverd op alle terreinen die ik de revue heb laten passeren, en het is de bedoeling dat dit zo blijft. Hierbij zal wel meer dan voorheen de nadruk liggen op *robuustheid* van technieken. De uitdaging is om dit begrip, dat meestal geformuleerd wordt in regressie termen, een geschikte plaats te geven binnen de datatheorie. Ook zal het onderzoek naar *afstandsmodellen* geïntensiveerd worden, omdat hier recentelijk interessante nieuwe aanknopingspunten zijn aangebracht door leden van de vakgroep. Ik denk aan de niet-lineaire biplot van John Gower[18], de afstandsbenadering van de multivariate analyse van Jacqueline Meulman[19], en het gebruik van de Kemeny afstand tussen rangordeningen van Rian van Blokland[20]. Tenslotte zullen modellen waarin de *tijd* een rol speelt in belang winnen. Er zijn de afgelopen acht jaar binnen het Leidse VF programma Datatheorie vier proefschriften over dit onderwerp tot stand gekomen, zodat gerust gesproken kan worden van een succesvolle onderzoekslijn[21]. Maar de introductie van functies van de tijd in de ruimtelijke theorie biedt nog vele onbenutte mogelijkheden, zodat op dit terrein interessante nieuwe activiteiten te verwachten zijn.

Zeer gewaardeerde toehoorders,

Aan het eind van mijn rede gekomen, wil ik al diegenen bedanken die zich ingezet hebben voor het voortbestaan van deze leerstoel. Ook ben ik dankbaar voor het vertrouwen dat aan mij is geschonken.

Hooggeleerde van de Geer, beste John,

Het kan je niet ontgaan zijn dat ik een woord uit de wereld van de muziek gebruikt heb om het centrale thema van de datatheorie mee aan te duiden. Je kunt muziekinstrumenten op elkaar afstemmen, of het tempo van spelen afstemmen op de vereiste sfeer. Omdat jij naast de datatheorie een passie voor muziek hebt, hoop ik dat je de keuze van deze centrale term als huldeblijk voor

jouw pioniersrol zult kunnen waarderen. Wat ik van jou boven alles heb geleerd is koersvastheid: het idee dat er een juiste lijn te trekken is, ook daar waar voorschriften en voorbeelden ontbreken. Het is het soort denken op een vakgebied dat principiëel is, maar niet vervalt in betweterij en principes met een grote P. Dank hiervoor.

Hooggeleerde de Leeuw, beste Jan,

Hoewel niet lijfelijk aanwezig, ben je toch ook zozeer mijn leermeester dat ik nu het woord tot je wil richten. Jij hebt me vele technische aspecten van het vak bijgebracht en me de ogen geopend voor die groeidiamant die we kortweg "scaling" noemen. Daarnaast heeft jouw brede visie ertoe geleid dat ik, om van hetzelfde panorama te kunnen genieten, haast onwillekeurig me op grotere hoogten heb gewaagd dan ik eerst van plan was. Het is goed om te zeggen dat zonder jou het onderzoek van de vakgroep nooit zo'n internationaal aanzien had verworven als wat nu is bereikt. Je hebt een hoge standaard achtergelaten.

Hooggeleerde van der Kamp, beste Leo,

Ik herinner me nog heel goed hoe jij me de beginselen van de multivariate analyse hebt uitgelegd. Het was na de cursus psychometrische testtheorie, in één van die donkere collegezalen op Stationsplein 242. Ik vroeg je wat factor analyse was, en je antwoord bestond uit het tekenen van twee pijlen op het bord, met een hoek van ongeveer 45 graden. 'Dit zijn twee tests', zei je erbij, 'en die hoek is hun correlatie'. Ik kon het niet geloven; ik was ontsteld! De rest van je betoog ging volledig langs me heen. Twee pijlen en een hoek: mysterieus maar ook transparant. Het was duidelijk, maar was dit alles? 'Ja, dit is eigenlijk wel alles', zei je met een bemoedigende glimlach.

Leo, veel mensen zouden deze anecdote aan John toeschrijven, maar juist daarom vertel ik hem: we zijn allen behept met dezelfde hartstocht.

Geachte medewerkers van de vakgroep Datatheorie en Wetenschapsstudies,

Sedert 1 januari jl. is er een nieuwe vakgroep die twee secties kent. In de eerste plaats wil ik het woord richten tot de medewerkers van de sectie Wetenschapsstudies. U behoort tot de meest actieve en enthousiaste gebruikers

van inbeddingstechnieken, in uw onderzoek op het terrein van de sociale netwerk structuur van wetenschappelijke activiteiten. Daarnaast zijn er ook theoretische aanknopingspunten tussen ons via de theorie van de fractals, dat is de meetkunde van zelfgelijkvormige objecten – zoals de hiërarchische bomen waarover ik vanmiddag sprak. Ik weet dat deze theorie uw warme belangstelling heeft, vanwege haar potentie om "eilandvorming" in de wetenschappelijke communicatie te verklaren, en ik vertrouw daarom op een alleszins vruchtbare samenwerking.

De sectie Datatheorie vormt de andere poot van de nieuwe vakgroep. Beste teamgenoten, samen met jullie heb ik een spannende periode achter de rug. Het was een tijd waarin vacaturestop, fusieplannen, en wisseling van de wacht samenvielen met het opstarten van nieuwe onderzoeksprojecten en het voltooien van twee belangrijke producten: het Gifi boek, dat door Wiley uitgegeven wordt, en de Gifi software, die door SPSS wordt gedistribueerd. Zoiets vereist nèt die extra geestdrift, waardoor het buitengewone zich onderscheidt van het gewone. Ik ben dankbaar voor jullie grote inzet en loyaliteit. Ik reken er op dat het onderzoek nu in een rustiger vaarwater terechtkomt, waar het eigenlijk ook thuishoort. De tijd is aangebroken om daarvan te gaan profiteren, en daarbij kan een ieder op mijn volledige steun rekenen.

Dames en Heren Studenten,

Mijn bemoeienissen met het eerste fase onderwijs zullen gering zijn; des te meer zal ik mij met het tweede fase onderwijs bemoeien. Een student heet in die fase, ondanks doctorandus titel, assistent-in-opleiding; en dat in een tijd dat een jongste bediende *assistent vice-president* heet! Moge AiO een geuzennaam worden voor u die in belangrijke mate de vernieuwing van de wetenschap moeten dragen. Ik zal trachten mij de raadgeving van Erasmus aan te trekken, die hij de geleerden van zijn tijd voorhield: 'Onafhankelijke en nobele karakters laten zich gaarne beleren, maar verdragen geen dwang'[22].

Ik heb gezegd.

NOTEN

[1] Deze vergelijking is ingegeven door het lezen van een beschouwing van Bas Heijne over de wereld van de schrijver en de werkelijkheid ("Gelijkenis", *Vrij Nederland/Boekenbijlage*, 16 December 1989). Het letterlijke citaat is: 'Zoals tien schilders een model op tien verschillende manieren afbeelden, zo kun je ook van vader Reve tien verschillende romanfiguren maken.'

[2] Het betreft hier eenzelfde soort primitieve term als het begrip stochastische grootheid, dat in de mathematische statistiek gebruikt wordt om uitspraken over een waarneming te doen.

[3] W.V. Quine, *Quiddities, An Intermittently Philosophical Dictionary*, Harvard University Press, 1987. Vertaald in Nederland uitgegeven bij Bert Bakker (Amsterdam, 1989), onder de titel "Quintessenties, Een bij vlagen filosofisch abc".

[4] De abstract geometrische aanpak van de multivariate analyse die hier bedoeld wordt is te vinden in, bijvoorbeeld, F. Cailliez en J.P. Pagès (*Introduction à l'Analyse des Données*, Paris: SMASH, 1976), en J.P. van de Geer (*Introduction to Linear Multivariate Data Analysis*, Leiden: DSWO Press, 1986).

[5] Parallellie aan een stijgende moedercurve is het gezamenlijk kenmerk van de Bogardus-schaal, de Guttman-schaal, de Mokken-schaal, de Rasch-schaal, en nog vele andere vormen van cumulativiteit. Hoewel fijnproevers hier een grote verscheidenheid menen te kunnen waarnemen, zijn vanuit het oogpunt van de datatheorie de overeenkomsten toch groter dan de verschillen.

[6] De combinatie van parallellie en enkeltoppigheid vormt één mogelijke definitie van het zgn. unfolding model, waarin substitutie de rol overneemt van cumulativiteit. Het ordenen van substitutie patronen leidt tot een grotere combinatorische complexiteit dan ordening naar monotonie. Vgl. W.J. Heiser (1981), *Unfolding Analysis of Proximity Data*. Academisch proefschrift, Leiden, en W.J. Heiser (1987), Joint ordination of species and sites: the unfolding technique, in P. Legendre *et al.* (Eds.), *Developments in Numerical Ecology*, New York: Springer.

[7] Meestal is deze "universele" inbedding niet-uniek, hoog-dimensionaal, en soms niet-Euclidisch, maar het is dan ook meestal niet de inbedding waar het

uiteindelijk om gaat. Vermelding van enkele Leidse bijdragen is hier op z'n plaats. Een uitgebreidere behandeling van de theorie van inbedding wordt gegeven in J. de Leeuw en W.J. Heiser (1982), Theory of multidimensional scaling (in L. Kanal en P.R. Krishnaiah (Eds.), *Handbook of Statistics, Volume II*; Amsterdam: North-Holland). Restricties worden nader uiteen gezet, naast de referenties vermeld in Noten 6 en 8, in J. de Leeuw en W.J. Heiser (1980), Multidimensional scaling with restrictions on the configuration (in P.R. Krishnaiah (Ed.), *Multivariate Analysis, Vol. V*, Amsterdam: North-Holland); W.J. Heiser en J.J. Meulman (1983), Analyzing rectangular tables with joint and constrained multidimensional scaling, *Journal of Econometrics*, *22*, 139-167; J.J. Meulman en W.J. Heiser (1984), Constrained multidimensional scaling: more directions than dimensions (in T. Havránek *et al.* (Eds.), *COMPSTAT 1984*, Wenen: Physica Verlag); en J. de Leeuw en J.J. Meulman (1986), Principal component analysis and restricted multidimensional scaling (in W. Gaul en M. Schader (Eds.), *Classification as a Tool of Research*, Amsterdam: North-Holland).

[8] Dergelijke hypothetische kenmerken kunnen vertaald worden in restricties op de puntenconfiguratie, of in restricties op de afstanden zelf; meestal komt dit op hetzelfde neer (vgl. W.J. Heiser & J.J. Meulman (1983), Constrained multidimensional scaling, including confirmation. *Applied Psychological Measurement*, *7*, 381-404).

[9] E.W. Holman (1972), The relation between hierarchical and Euclidean models for psychological distances. *Psychometrika*, *37*, 417-423.

[10] F. Critchley & W.J. Heiser (1988), Hierarchical trees can be perfectly scaled in one dimension. *Journal of Classification*, *5*, 5-20.

[11] Het voorbeeld is niet geheel willekeurig gekozen. Veel problemen die te maken hebben met beschrijving of manipulatie van concrete verdelingen zijn bij uitstek treffend te beschrijven als inbeddingsproblemen.

[12] Voor een recent overzicht, zie J.C. van Houwelingen en S. le Cessie (1988), Logistic regression, a review. *Statistica Neerlandica*, *42*, 215-232. Ook de overige artikelen in deze aflevering nr. 4 zijn interessant, zie Noot 16.

[13] Binnen de Sociale Wetenschappen kunnen reële verschillen in onderzoekstradities tot een interessante diversiteit in methoden leiden. Een aardig

voorbeeld hiervan vormt de ontwikkeling van de *latente trek theorie* en de *latente klassentheorie*, die beide gericht zijn op de analyse van meerdere binaire items. De eerste spruit voort uit de psychologie, waar men gewend is met veel items en relatief kleine steekproeven te werken, zodat het begrijpelijk is dat het meest populaire model – het Rasch model – zich tot de bivariate marginalen beperkt. De tweede spruit voort uit de sociologie, waar men gewend is met weinig items en relatief grote steekproeven te werken, zodat het begrijpelijk is dat de loglineaire analyse centraal kwam te staan, die bij uitstek multivariaat is. Toch is het geen enorme stap om je af te vragen of de latente klassen soms één-dimensionaal geordend zijn, en is het gezond vakmanschap om naar de residuën ten opzichte van hogere-orde marginalen te kijken. Het recente boek van R. Langeheine en J. Rost (Eds.) (*Latent trait and latent class models*. New York and London: Plenum Press, 1988) laat zien dat hier de onvermijdelijke integratie in volle gang is gekomen. In Leiden konden we dit al zien aankomen dankzij het proefschrift van A. Mooijaart (1978), *Latent Structure Models*. De ontwikkelingen gaan nu echter zeer snel, zie bijv. de discussie artikelen van D. Edwards (Hierarchical interaction models), en van N. Wermuth en S.L. Lauritzen (On substantive research hypotheses, conditional independence graphs en graphical chain models) in de *Journal of the Royal Statistical Society*, Series B (1990), *52*, No. 1, pp. 3-72.

[14] I.W. Molenaar (1988), Formal statistics and informal data analysis, or why laziness should be discouraged. *Statistica Neerlandica*, *42*, 83-90. De vertaling van de Engelse neerslag van de discussie is van mij.

[15] J. de Leeuw (1988), Models and techniques. *Statistica Neerlandica*, *42*, 91-98. Zie ook noot 14.

[16] In bepaalde schaalmodellen, bijvoorbeeld, zijn zowel persoons- als itemparameters conceptueel niet stochastisch. Het correcte domein van generalisatie is in zo'n geval gedefineerd per vaardigheidsgroep maal itemmoeilijkheidsklasse combinatie. Door in navolging van R.D. Bock en M.Lieberman (Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *40* (1970), 5-32) eenvoudigweg te marginaliseren over personen verlaat men in feite het oorspronkelijke model, en gooit zo het kind met het badwater weg. Of een dergelijke specificatiefout in de praktijk veel voorkomt kan ik niet beoordelen. Sinds kort winnen relatief eenvoudige herwogen kleinste kwadraten methoden gelukkig weer terrein. Hun

bruikbaarheid, ook op klassieke criteria zoals zuiverheid en efficiëntie, wordt onder meer bewezen in N. Verhelst en I.W. Molenaar (1988), Logit based parameter estimation in the Rasch model. *Statistica Neerlandica, 42*, 273-295.

[17] Omwille van de eenvoud duid ik hier de modellen die bij uitstek door de datatheorie worden gedekt aan als modellen voor verklaring van variantie. Deze term wordt niet gebruikt zoals in de "analysis of variance" in de stricte betekenis; gedoeld wordt op allerlei soorten spreiding rond allerlei soorten van gemeenschappelijke structuur.

[18] J.C. Gower en S.A. Harding (1988), Nonlinear biplots. *Biometrika, 75*, 445-455.

[19] J.J. Meulman (1986), *A Distance Approach to Nonlinear Multivariate Analysis*. Academisch proefschrift, Leiden: DSWO Press.

[20] A.W. van Blokland-Vogelesang (1989), Unfolding and consensus ranking: a prestige ladder for technical occupations (in G. de Soete *et al.* (Eds.), *New Developments in Psychological Choice Modeling*, Amsterdam: North-Holland).

[21] In het VF programma werkt de vakgroep Datatheorie en Wetenschapsstudies samen met de vakgroepen Methoden & Technieken van Psychologische Research, en de Vakgroep Algemene Pedagogiek. Het betreft de proefschriften van R.A. Visser (*On Quantitative Longitudinal Data in Psychological Research*, 1982), P.G.M. van der Heijden (*Correspondence Analysis of Longitudinal Categorical Data*, 1987), C.C.J.H. Bijleveld (*Exploratory Linear Dynamic Systems Analysis*, 1989) en S. van Buuren (*Optimal Scaling of Time Series*, 1990). Met enige goede wil zou ook P.M. Kroonenberg (*Three-mode Principal Component Analysis*, 1983) in deze rij in te passen zijn.

[22] Brief aan paus Leo X, vanuit Leuven, gedateerd 13 september 1520. In: *Luther is mij volkomen vreemd, uit de brieven van Erasmus*. Vertaald en ingeleid door J. Trapman. Weesp: De Haan, 1983.

Ontvangen: 26-02-1990
Geaccepteerd: 26-02-1990