HOOFDSTUK 10

Keuzedata

10.1 Rechthoekige nabijheidsmatrices

In Hoofdstuk 3 en aan het begin van Hoofdstuk 8 hebben we een opsomming gegeven van de mogelijke varianten van het CMDS-probleem. In Hoofdstuk 6 en Hoofdstuk 8 hebben we de gevallen waarin er sprake is van een vierkante nabijheidsmatrix allemaal behandeld. In dit hoofdstuk komen de rechthoekige nabijheidsgegevens aan de orde. Daarbij gaat het in principe om tweeweg/tweemodale en om drieweg/driemodale data. Dat wil zeggen dat de elementen van de verschillende wegen uit verschillende verzamelingen afkomstig zijn of in ieder geval *als zodanig behandeld worden*. Een voorbeeld daarvan zijn we al tegengekomen bij die analyse van asymmetrische data waarin de rijobjecten en de kolomobjecten door afzonderlijke punten werden afgebeeld. Strikt genomen waren de rij- en de kolomobjecten in dat geval wel dezelfde elementen uit één verzameling, maar werden ze geanalyseerd alsof dat niet zo was. In feite was de datamatrix vierkant, maar werd hij behandeld alsof hij rechthoekig was.

Een soortgelijk voorbeeld van een rechthoekige datamatrix is een tabel met als rijen de eerder gebruikte acht Nederlandse steden en als kolommen een aantal, zeg vijf, andere Nederlandse steden. In de cellen van die tabel staat dan bijvoorbeeld de tijd die het kost om per fiets van een rij-stad naar een kolom-stad te reizen. Ook hier komen de rij- en de kolomelementen niet echt uit twee verschillende verzamelingen; het zijn echter wel verschillende deelverzamelingen van alle Nederlandse steden.

Hoewel enigszins gekunsteld, laat dit voorbeeld een van de belangrijkste kenmerken van rechthoekige data duidelijk zien, namelijk dat zo'n matrix in feite een *incomplete* nabijheidsmatrix is. Stel, we nemen als kolomobjecten de steden Alkmaar, Middelburg, Maastricht, Venlo en Leeuwarden. De rechthoekige

datamatrix bevat dan alleen de afstanden tussen deze vijf steden en Amsterdam tot en met Groningen. De rechthoekige matrix bevat echter *niet* de onderlinge afstanden tussen de rijobjecten Amsterdam tot en met Groningen en evenmin de onderlinge afstanden van de kolomobjecten Alkmaar tot en met Leeuwarden. Van de in totaal $(8 + 5)(8 + 5 \times 1)/2 = 78$ afstanden tussen alle dertien steden zijn er in de rechthoekige tabel maar $8 \times 5 = 40$ gegeven. We beschikken dus in dit geval slechts over zo'n vijftig procent van de gegevens die we in principe nodig hebben om een goede kaart van Nederland te construeren. Het zal duidelijk zijn dat deze situatie bij MDS tot problemen *kan* leiden. Het bijzondere is eigenlijk dat MDS van dit soort data *niet altijd* tot problemen leidt. Daarbij zijn verschillende aspecten van belang, met name (de verhouding van) het aantal rijen en het aantal kolommen en de aard en kwaliteit van de data. Hierop zullen we later terugkomen.

Het klassieke voorbeeld van een rechthoekige nabijheidsmatrix is een tabel die per rij de *rangordening* bevat die een persoon (als rijobject) gegeven heeft van zijn of haar voorkeuren voor een aantal kolomobjecten. Omdat deze rangordeningen door verschillende personen gegeven zijn, zijn dit soort gegevens in principe *rijconditioneel*. Immers: ook al vindt zowel persoon A als persoon B spruitjes de lekkerste groente van het drietal spruitjes, bloemkool en andijvie, toch kan persoon A spruitjes veel lekkerder vinden dan persoon B. Persoon A zou bij wijze van spreken wel iedere dag spruitjes willen eten, terwijl persoon B spruitjes eigenlijk alleen maar de minst onaangename groente van het aangeboden drietal vindt. De nabijheid tussen persoon A en spruitjes is dus (veel) groter dan die tussen persoon B en spruitjes. De data zeggen daarom alleen iets over de rangorde van de nabijheden binnen de afzonderlijke rijen. Door deze rijconditionaliteit is het aantal gegevens waarover men beschikt minder dan in het matrixconditionele geval. Aangezien er verder geen informatie is over de onderlinge nabijheden van de rijobjecten en evenmin over de nabijheden van de kolomobjecten onderling, zijn er in het klassieke geval soms te weinig gegevens om een bruikbare oplossing te krijgen. Ook hierop zullen we later (in Hoofdstuk 11) nog terugkomen.

In plaats van rangordeningen kan een rechthoekige tabel met voorkeursdata ook andere getallen bevatten, bijvoorbeeld het aantal uren per week dat men aan bepaalde activiteiten (werk, sport, muziek, huishouden, enzovoort) besteedt of het aantal guldens dat men per maand uitgeeft aan wonen, eten, studie, reizen, vrije tijd, enzovoort. In deze laatste gevallen zijn de data in beginsel niet rijconditioneel; formeel lijken deze data dus meer op het eerstgenoemde stedenvoorbeeld.

Een derde, eveneens klassiek, voorbeeld van een rechthoekige nabijheidsmatrix is een matrix met enen en nullen, waarvan de enen aangeven welke kolomobjecten van toepassing zijn op, of gekozen zijn door de rij-objecten. Frequent voorkomende gevallen zijn:

– een matrix van personen bij eigenschappen (een of meer beoordelaars of de personen zelf hebben op een lijst van eigenschappen die kenmerken aangekruist die op de beoordeelde personen van toepassing zijn);

– een matrix van personen bij objecten (een één geeft aan dat een persoon het desbetreffende object bezit, zou willen hebben, mooi vindt, lelijk vindt, rood vindt, griezelig vindt, enzovoort);
– een matrix van personen bij uitspraken (de enen en nullen geven aan welke personen het met welke uitspraken eens of oneens zijn);
– een matrix van personen bij test-items (een één geeft aan dat de persoon het test-item goed beantwoord heeft).

Zo zouden we nog wel even door kunnen gaan, vooral als we bedenken dat de rijobjecten niet alleen maar op personen hoeven te slaan, maar ook dieren, dingen en concepten kunnen zijn.

Preferentiedata

Rechthoekige nabijheidsgegevens worden vaak *preferentiedata* genoemd. Immers: in alle bovenstaande voorbeelden heeft een bepaald rijobject voor sommige kolomobjecten wél een voorkeur en voor andere niet. Bij preferentiedata kunnen we een onderscheid maken tussen *preferentierangordeningen* en *keuzedata*. Preferentierangordeningen zijn observaties die verkregen worden met wat Coombs (1964) een *order*-opdracht noemde, keuzedata zijn observaties, verkregen uit een *pick*-taak.
Zijn er m (kolom)objecten, dan kan men een proefpersoon de opdracht *order k of m* geven, waarbij k kan lopen van twee tot en met $m - 1$. *Order 2 of m* wil zeggen: kies eerst de meest geprefereerde stimulus uit en vervolgens de op één na meest geprefereerde stimulus. *Order (m – 1) of m* geeft een volledige rangordening van alle m stimuli. De opdracht *Pick k of m* vraagt de proefpersoon exact k objecten uit te kiezen. *Pick any of m* laat de proefpersoon vrij in het aantal objecten dat hij of zij kiest.
Zowel voorkeursrangordeningen als keuzedata kunnen op verschillende manieren meerdimensionaal geschaald worden. In de rest van dit hoofdstuk behandelen we de analyse van keuzedata; in Hoofdstuk 11 volgt de MDS van voorkeursrangordeningen.

10.2 MDS VAN KEUZEDATA

Informatie in keuzedata

In Hoofdstuk 3 hebben we Coombs indelingssysteem van data en observaties besproken. Volgens die indeling vallen observaties die uit rechthoekige keuzedata bestaan in kwadrant II. Immers: de observaties hebben betrekking op relaties tussen paren van punten die uit verschillende verzamelingen komen. Daarbij kunnen we dergelijke observaties op twee manieren opvatten: als nabijheidsrelatie (IIb) en als dominantierelatie (IIa). In het eerste geval wordt de observatie dat persoon A object Y kiest als een indicatie van nabijheid opgevat: $o_{AY} = 1$ impliceert dat $d_{AY} < \varepsilon_A$ en $o_{AY} = 0$ betekent dat $d_{AY} > \varepsilon_A$.

Het symbool ε_A staat hier voor een soort *drempelwaarde* of *kritische afstand*. Overschrijdt de afstand tussen persoon en object deze waarde, dan wordt het object niet gewenst, niet gekozen. Is de afstand tussen persoon en object kleiner dan deze kritische afstand, dan wordt het object wél door de persoon gekozen.
In het tweede geval interpreteren we de observatie dat persoon A object Y kiest als een ordenings- of dominantierelatie: $o_{AY} = 1$ impliceert dat $\theta_Y > \theta_A$, wat bijvoorbeeld betekent dat de waarde van object Y groter is dan de minimale waarde of *utiliteit* die een object moet hebben om door persoon A gekozen te worden. Ook hier fungeert θ_A dus als een drempelwaarde. In de moderne psychologische testtheorie, met name in de zogenaamde *item-responstheorie*, gaat men uit van een variant van dit model: een vraag (item) wordt goed beantwoord als de latente vaardigheid θ_A van persoon A de moeilijkheidsgraad θ_Y van item Y overtreft; $o_{AY} = 1$ betekent dan dat $\theta_A > \theta_Y$. De moeilijkheidsgraad van item Y wordt meestal gedefinieerd als de hoeveelheid vaardigheid die nodig is om item Y correct te kunnen beantwoorden[1]. Modellen voor dit soort data komen vooral aan de orde in boeken over moderne testtheorie, bijvoorbeeld Hambleton, Swaminathan en Rogers (1991) en Van der Linden en Hambleton (1997). Enkele specifieke modellen voor keuzedata-als-dominantiegegevens zullen we aan het eind van dit hoofdstuk summier behandelen, terwijl we in Hoofdstuk 11 vooral aandacht zullen besteden aan MDS-analyse van preferentiedata die uit rangordeningen bestaan. Hieronder behandelen we echter eerst de MDS-analyse van keuzedata-als-nabijheden.

10.3 KEUZEDATA ALS NABIJHEDEN

Keuzedata die geïnterpreteerd worden als nabijheden kunnen eendimensionaal en meerdimensionaal geschaald worden door personen en objecten in één gezamenlijke ruimte af te beelden. In dat geval zoeken we een afbeelding waarin de personen dicht bij de door hen gekozen objecten liggen en waarin de objecten dicht bij die personen liggen die hen gekozen hebben.

De *joint plot*

Zoals gezegd willen we in een MDS-analyse van keuzedata een afbeelding maken van personen (rijen) en objecten (kolommen) in één gemeenschappelijke ruimte, een zogenaamde *joint plot*. De personen en objecten zijn zodanig in de ruimte afgebeeld dat hun onderlinge afstanden de keuzedata 'verklaren'. *Idealiter* heeft die afbeelding de volgende eigenschappen:

1 In de item-responstheorie wordt een probabilistische versie van dit model gebruikt: $\theta_A > \theta_Y$ impliceert dat de kans op een correct antwoord groter is dan 50%; dus: $Prob(o_{AY} = 1) > .50$. De moeilijkheidsgraad θ_Y is die waarde van θ waarvoor geldt dat personen met een vaardigheid die gelijk is aan die waarde een kans van 50% hebben om het item goed te beantwoorden, dus: $Prob(o_{AY} = 1 | \theta_A = \theta_Y) = .50$.

1 personen liggen dicht bij de door hen gekozen objecten, althans dichter bij de gekozen objecten dan bij de objecten die zij niet gekozen hebben;
2 objecten liggen dicht bij de personen door wie ze gekozen zijn, althans dichter bij die personen dan bij de personen door wie ze niet gekozen zijn;
3 personen met sterk overeenkomstige keuzepatronen, dat wil zeggen, personen die voor een groot deel dezelfde objecten kiezen, liggen dichter bij elkaar in de buurt dan personen die merendeels verschillende objecten kiezen;
4 objecten die merendeels door dezelfde groep personen gekozen worden, liggen dichter bij elkaar dan objecten die door verschillende groepen personen gekozen worden.

Elk van deze eigenschappen volgt logischerwijs uit elke andere eigenschap; het zijn vier verschillende manieren om de informatie in de afbeelding samen te vatten.

De vraag is nu natuurlijk: hoe komen we aan zo'n afbeelding? Laten we eens veronderstellen dat we weten welke posities de personen in de afbeelding innemen. Voor iedere persoon zijn dan de coördinaten bekend op de dimensies van de ruimte. Stel vervolgens dat twee personen, *A* en *B*, allebei object *Y* gekozen hebben, en dat dit de enige twee personen zijn door wie *Y* gekozen is. Waar zouden we object *Y* nu het best kunnen afbeelden? Het antwoord is: zo dicht mogelijk in de buurt van zowel persoon *A* als persoon *B*, dus in de *centroïde* van de posities van *A* en *B*. De centroïde bepaalt men door de coördinaten van de personen per dimensie te middelen; de gemiddelde coördinaatwaarden zijn dan de coördinaten van de centroïde. Zo'n centroïde heeft de eigenschap dat de som van de gekwadrateerde afstanden tot de bijbehorende punten minimaal is. Als we *Y* in de centroïde van *A* en *B* plaatsen, heeft *Y* dus de kleinste gemiddelde (gekwadrateerde) afstand tot *A* en *B*. Kortom, dit is het beste dat we kunnen doen als we voor *Y* een positie moeten bepalen.

Omgekeerd, stel dat de positie van de objecten van tevoren bekend was en dat het erom gaat persoon *P* af te beelden van wie bekend is dat hij uitsluitend object *Y* en *Z* gekozen heeft. Ook in dit geval is het de beste keuze om persoon *P* in de centroïde van de objecten *Y* en *Z* te lokaliseren.

Het echte MDS-probleem bij keuzedata is ingewikkelder, omdat noch de coördinaten van de personen noch de coördinaten van de objecten van tevoren bekend zijn. Het gaat er juist om de coördinaten van beide verzamelingen te vinden. Dat kan weer op een iteratieve manier, door gebruik te maken van de centroïden-aanpak die hierboven geschetst is. Het betreffende algoritme gaat als volgt:

Stap 1: kies voor alle n personen willekeurige coördinaten op één dimensie; deze coördinaten duiden we aan met x_{p1}. Standaardiseer de coördinaten zodanig dat ze een gemiddelde van nul en een kwadratensom gelijk aan n hebben. Dus: $\sum_p x_{p1} = 0$, $\sum_p x^2_{p1}/n = 1.0$.

Stap 2: bepaal voor ieder object j ($j = 1,..., m$) de centroïde van die personen die het betreffende object gekozen hebben. We hebben nu dus coördinaten voor de objecten die als volgt berekend zijn:

$$y_{j1} = \frac{\sum_{p=1}^{n} o_{pj} x_{p1}}{\sum_{p=1}^{n} o_{pj}} \qquad [10.1]$$

waarbij y_{j1} de coördinaat van object j op dimensie 1 is. Het symbool o_{pj} staat voor de keuze van persoon p voor object j en kan alleen de waarden 1 (gekozen) en 0 (niet gekozen) aannemen. In de noemer van Formule [10.1] staat het totaal aantal personen dat object j gekozen heeft; in de teller staat de som van de coördinaten van die personen.

Stap 3: bepaal voor iedere persoon de centroïde van de objecten die hij gekozen heeft. We krijgen zo dus nieuwe coördinaten voor de personen, volgens

$$x_{p1} = \frac{\sum_{j=1}^{m} o_{pj} y_{j1}}{\sum_{j=1}^{m} o_{pj}} \qquad [10.2]$$

In de noemer van Formule [10.2] staat het totaal aantal objecten dat door persoon p gekozen is; in de teller staat de som van de coördinaten van die objecten.

Stap 4: herschaal de nieuwe coördinaten van de personen zodanig dat ze opnieuw een kwadratensom gelijk aan n krijgen. Ga nu naar Stap 2 en herhaal de Stappen 2, 3 en 4, net zolang tot de nieuwe coördinaten van de personen en objecten niet meer verschillen van de coördinaten uit de vorige iteratie.

Stap 5: kies een nieuw stel willekeurige coördinaten voor de personen. Dit zijn hun coördinaten x_{p2} op de tweede dimensie. Kies de $\{x_{p2}\}$ op zo'n manier dat ze een gemiddelde van nul en een kwadratensom van n hebben en dat ze *ongecorreleerd* (dus orthogonaal) zijn met de uiteindelijke coördinaten op de eerste dimensie. Dus: $\sum_p x_{p2} = 0$, $\sum_p x^2_{p2} = n$ en $\sum_p x_{p1} x_{p2} = 0$.

Stap 6: ga naar Stap 2 en herhaal de Stappen 2 tot en met 4 voor de coördinaten op de tweede dimensie. Zorg er iedere keer voor dat de nieuwe coördinaten van de personen op de tweede dimensie ongecorreleerd zijn met hun coördinaten op de eerste dimensie. Dat kan door elke keer de regressie van de nieuwe $\{x_{p2}\}$ op de $\{x_{p1}\}$ te berekenen en de coördinaten op de tweede dimensie te vervangen door het residu $x_{p2} - (b.x_{p1} + a)$.
Wil men drie of meer dimensies, dan kan men dit proces net zo vaak herhalen als men nodig vindt. Om ongecorreleerde coördinaten op de derde dimensie te krijgen moet men de residuen nemen van de multipele regressie van die coördinaten op de coördinaten van de eerste twee dimensies:

$$x_{p3} - (b_1 x_{p1} + b_2 x_{p2} + a).$$

Bovenstaande methode is bedacht door Richardson en Kuder in 1933 en door hen de *method of reciprocal averaging* genoemd. Richardson beschikte uiteraard niet over een computer. Hij voerde het iteratieve proces uit met behulp van een soort ponskaarten die hij steeds opnieuw op stapeltjes legde. Tegenwoordig

beschikken we wel over computers en over verschillende computerprogramma's die bovenstaand algoritme snel kunnen uitvoeren. Een van die programma's is *HOMALS* (Gifi, 1990) dat ook in SPSS is opgenomen.

10.4 HOMALS

HOMALS is een acroniem voor *homogeneity analysis by alternating least squares.* De naam homogeniteitsanalyse berust op het feit dat HOMALS groepen respondenten probeert te onderscheiden die homogeen zijn in hun antwoordpatronen en dus dicht bij elkaar moeten worden afgebeeld. HOMALS berust op een aanpak die door verschillende onderzoekers in verschillende landen op verschillende tijdstippen is (her)ontdekt (zie ook Gifi, 1990; Tenenhaus & Young, 1985). Hoewel HOMALS niet speciaal ontworpen is voor keuzedata die uit enen en nullen bestaat, is het daar goed voor te gebruiken. In Blok 10.1 bespreken we een HOMALS-analyse van keuzedata die in empirisch onderzoek verkregen zijn. Hieronder bekijken we eerst een klein, artificieel, voorbeeldje van een HOMALS-oplossing.

Stel dat we zes studenten twee vragen voorleggen die ze met ja of nee mogen beantwoorden. Vraag 1 luidt 'Denk je dat je erin zult slagen je studie in vier jaar af te ronden?' en Vraag 2: 'Denk je dat je binnen drie maanden na je afstuderen een betrekking zult hebben gevonden?'. Stel vervolgens dat we de antwoorden krijgen die in Tabel 10.1 zijn weergegeven.

Tabel 10.1 De antwoorden van zes respondenten op twee vragen

Respondent	Vraag 1	Vraag 2
1	1[a]	1
2	1	1
3	1	2
4	2	2
5	2	2
6	2	1

[a] De cijfers zijn coderingen van de antwoordcategorieën: 1 = ja; 2 = nee.

Door middel van HOMALS proberen we een meerdimensionale oplossing te vinden waarin alle proefpersonen die dezelfde antwoorden gegeven hebben dezelfde coördinaten krijgen. Dat betekent voor Vraag 1 dat enerzijds de respondenten 1, 2 en 3 dezelfde coördinaten moeten krijgen evenals, anderzijds, de respondenten 4, 5 en 6. Bovendien moeten de coördinaten van de laatste drie personen liefst maximaal verschillend zijn van die van de eerste drie. Voor Vraag 2 moeten de personen 1 , 2 en 6 dezelfde coördinaten krijgen

die liefst maximaal verschillen van de coördinaten van respondenten 3, 4 en 5. Zo'n oplossing is echter onmogelijk: persoon 3 en persoon 6 kunnen niet tegelijkertijd dezelfde coördinaten als persoon 1 en 2 én dezelfde coördinaten als persoon 4 en 5 krijgen.

De tweedimensionale HOMALS-oplossing[2] van dit probleem staat in Tabel 10.2. In deze tabel staan de coördinaten van de personen op twee dimensies. De eerste dimensie is een compromis: persoon 3 en persoon 6 krijgen een coördinaat waardoor zij precies tussen de posities van persoon 1 en 2 enerzijds en persoon 3 en 4 anderzijds in komen te liggen. Hoewel deze oplossing optimaal is in één dimensie, is hij niet perfect, omdat respondent 3 en respondent 6 samenvallen terwijl zij tegengestelde antwoordpatronen hebben. Nemen we nu een tweede dimensie, dan zoekt HOMALS voor de personen coördinaten die orthogonaal zijn met hun coördinaten op de eerste as. Persoon 1, 2, 4 en 5 krijgen op deze dimensie allemaal dezelfde positie (in het midden) maar respondenten 3 en 6 komen op deze nieuwe as maximaal uit elkaar te liggen. Gezamenlijk laten de twee dimensies dus een perfecte oplossing zien: de vier verschillende antwoordpatronen van de respondenten worden nu perfect onderscheiden.

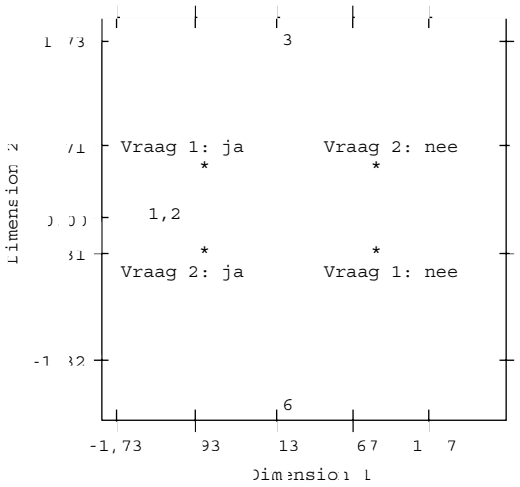
Merk op dat op beide dimensies de coördinaten een gemiddelde van nul hebben en dat hun kwadratensom gelijk aan zes is. Als coördinaten van de antwoordcategorieën geeft HOMALS op de eerste dimensie de waarden (-1.22 – 1.22 + 0.00)/3 = -.813 voor ja en (1.22 + 1.22 + 0.00)/3 = +.813 voor nee op de eerste vraag. Op de tweede dimensie zijn de coördinaten van deze antwoordcategorieën respectievelijk (0.00 + 0.00 + 1.73)/3 = +.577 en (0.00 + 0.00 – 1.73)/3= -.577. Op analoge manier berekend, zijn de antwoordcategorieën van de tweede vraag -.813 en +.813 op de eerste en -.577 en +.577 op de tweede dimensie. De afbeelding van deze HOMALS-oplossing is weergegeven in Figuur 10.1.

Tabel 10.2 De coördinaten van zes respondenten op twee HOMALS-dimensies

Respondent	Dimensie 1	Dimensie 2
1	-1.22	0.00
2	-1.22	0.00
3	0.00	1.73
4	1.22	0.00
5	1.22	0.00
6	0.00	-1.73

2 Deze oplossing is verkregen met de SPSS-commando's:

```
  DATA LIST FREE/VRAAG1 VRAAG2.
BEGIN DATA.
        1 1 1 1 1 2 2 2 2 2 2 1
END DATA.
        HOMALS VARIABLES VRAAG1 VRAAG2 (2)/PRINT DEFAULT OBJECT.
```



Figuur 10.1 Afbeelding van de HOMALS-oplossing uit Tabel 10.2 van het voorbeeld van Tabel 10.1

Goodness-of-fit van een HOMALS-oplossing

Een aspect waar we het nog niet over gehad hebben, is de vraag hoe goed deze HOMALS-oplossing nu eigenlijk de data beschrijft. De *goodness-of-fit* wordt in HOMALS uitgedrukt in termen van *discriminatiematen*. Het idee achter HOMALS is het onderscheiden van homogene groepen, dat wil zeggen, dat personen met verschillende antwoordpatronen zover mogelijk uit elkaar geplaatst worden en dat personen met dezelfde antwoorden in principe zouden moeten samenvallen. HOMALS berekent nu per vraag en per dimensie een discriminatiemaat, door op de coördinaten van de personen een *variantieanalyse* (ANOVA) uit te voeren. De *totale kwadratensom* (SS_T) van de persooncoördinaten wordt opgesplitst in een *tussen-groepen kwadratensom* (SS_B; de B komt van between) en een *binnen-groepen kwadratensom* (SS_W; de W komt van within). Voor alle vragen en alle dimensies geldt dat $SS_T = n$ (in dit geval: 6.00) omdat de persooncoördinaten zo gestandaardiseerd zijn. De groepen waar het om gaat zijn de proefpersonen die hetzelfde antwoord hebben gegeven. In ons voorbeeld heeft Vraag 1 twee groepen. De eerste groep bestaat uit de eerste drie personen, die ja hebben geantwoord; de tweede groep bestaat uit persoon 4, 5 en 6, die nee hebben gezegd. Deze groepen hebben op de eerste dimensie respectievelijk de gemiddelde coördinaten -.813 en +.813 (NB: deze gemiddelden zijn de betreffende categoriecoördinaten). De SS_B van deze groepen is $3 \times (-.813)^2 + 3 \times (+.813^2) = 6 \times .661 = 3.966$. De bijbehorende discriminatiemaat is nu gelijk aan $SS_B/SS_T = 3.966/6.00 = .661$. Ook de tweede vraag heeft op de eerste dimensie een discriminatiemaat van .661; de gemiddelde discriminatiemaat van alle

vragen op deze dimensie is dus ook .661. Dit gemiddelde is een maat voor de *fit* van de eerste dimensie in zijn geheel. Op de tweede dimensie zijn de SS_B's van beide vragen gelijk aan $6 \times .577^2 = 1.998$. De *fit* van de tweede dimensie in zijn geheel is dus 1.998/6 = .333. We zien hieraan dat de eerste dimensie belangrijker is dan de tweede. We zien ook dat de gezamenlijke *fit* gelijk is aan .661 + .333 = .994. De maximale *fit* bij een oplossing in twee dimensies is gelijk aan 2.00, het aantal dimensies. We hebben dus vijftig procent van de variatie in de antwoorden 'verklaard'.

HOMALS-terminologie

Voor we verdergaan, iets over de terminologie van het HOMALS-programma. De kolommen van de datamatrix die in HOMALS wordt ingevoerd, worden in een HOMALS-analyse *variabelen* genoemd. Elke variabele heeft een aantal *categorieën*; de coördinaten van de categorieën heten *categoriekwantificaties*. De rijen van de geanalyseerde matrix heten bij HOMALS *objecten*; hun coördinaten of kwantificaties heten *objectscores*. De discriminatiematen zijn we hierboven al tegengekomen. De totale *fit* van een dimensie correspondeert met een zogenaamde *eigenwaarde*. Zie voor meer details over HOMALS bijvoorbeeld Van de Geer (1988) of de handleiding van *SPSS Categories* (1994).

HOMALS op keuzedata

Wat is de relatie van bovenstaand voorbeeld met de keuzedata waar het in dit hoofdstuk om gaat? Met andere woorden: hoe kunnen we HOMALS toepassen op de enen-en-nullen-data die aangeven of iemand iets gekozen heeft of niet? Daarbij zijn twee aspecten van belang. In de eerste plaats gaat het erom hoe we de matrix met keuzedata in HOMALS invoeren: getransponeerd of ongetransponeerd. In de tweede plaats gaat het om de betekenis van de nullen in de analyse.
Variabelen en objecten. In een matrix met keuzedata is het gebruikelijk dat de rijen corresponderen met 'kiezers', meestal personen, en de kolommen met 'objecten', de dingen die al of niet gekozen worden. In ons voorbeeldje waren de rijen inderdaad personen en waren de kolommen de vragen die beantwoord moesten worden. Voert men deze matrix ongetransponeerd in HOMALS in, dan krijgen we een oplossing waarin de coördinaten van de rijen (de objectscores) op elke dimensie een kwadratensom gelijk aan *n* hebben en de antwoordcategorieën van de variabelen in de centroïden van de rijen worden afgebeeld. Het idee is dus: een object wordt afgebeeld te midden van de personen door wie het gekozen is.
We kunnen dit idee omdraaien door een kiezer af te beelden te midden van de objecten die hij gekozen heeft. In dat geval moeten we de matrix met keuzedata eerst transponeren voor hij in HOMALS wordt ingevoerd. De gekozen objecten zijn dan de rijen van de matrix (de 'objecten' bij HOMALS) en de kiezers zijn de kolommen (de 'variabelen' bij HOMALS). In dat geval krijgen de objecten op elke dimensie een kwadratensom gelijk aan *m* en worden de kiezers afgebeeld

in de centroïden van de gekozen objecten. Voor de *fit* maakt het niet uit, het is alleen een andere manier van afbeelden.

Nullen als ontbrekende gegevens. In het voorbeeldje hierboven hebben we gezien dat HOMALS de ja-zeggers op een vraag gezamenlijk zo dicht mogelijk afbeeldt bij het punt voor de ja-categorie van die vraag en de nee-zeggers zo dicht mogelijk bij de nee-categorie. Dus ook alle mensen die *niet* de ja-categorie hebben gekozen, worden idealiter boven op elkaar afgebeeld. Bij keuzedata die uit enen en nullen bestaan willen we dat meestal niet. Het enige dat we willen, is de personen die een bepaald object gekozen hebben zo dicht mogelijk bij elkaar en zo dicht mogelijk bij het gekozen object plaatsen. De *ideaalpunten* van deze personen liggen in principe in de buurt van een gekozen object. Er is echter geen dwingende reden om alle personen die dat object *niet* gekozen hebben eveneens dicht bij elkaar af te beelden. Als iemand een object niet gekozen heeft, wil dat alleen maar zeggen dat (het ideaalpunt) van deze persoon een betrekkelijk grote afstand tot het object heeft. Het zegt verder niets over de positie van die persoon in de ruimte. De ene niet-kiezer kan aan de ene kant van de ruimte liggen, een andere niet-kiezer aan de andere kant. Het enig belangrijke is dat ze allebei relatief ver van het niet-gekozen object komen te liggen. Bij de analyse van keuzedata willen we onze afbeelding dus eigenlijk alleen maar baseren op de enen in de matrix en niet op de nullen. In HOMALS gebeurt dit automatisch, omdat nullen door HOMALS als ontbrekende gegevens (missing values) beschouwd worden. Wil men de nullen wel in de analyse betrekken, dan moeten de data eerst gehercodeerd worden, bijvoorbeeld, door van de nullen enen en van de enen tweeën te maken.
Ter illustratie van een HOMALS-analyse met nullen als ontbrekende data (in Blok 10.1 wordt een groter voorbeeld behandeld) zijn de tweeën uit de gegevens uit Tabel 10.1 gehercodeerd tot nullen (zie Tabel 10.3). Merk op dat er daardoor voor twee respondenten geen enkele valide gegevens meer beschikbaar zijn. Het ligt voor de hand om deze respondenten uit de analyse weg te laten. Omdat we alleen maar weten welke objecten ze niet gekozen hebben, hebben we immers geen enkele informatie over de plaats van hun ideaalpunten in de ruimte. Als we ze niet weglaten, hebben ze overigens geen invloed op de oplossing. Ze worden door HOMALS automatisch in de oorsprong van de configuratie afgebeeld, hun objectscores zijn dan nul op alle dimensies. Aan de data in Tabel 10.3 kunnen we zien dat de vier respondenten die wel valide gegevens hebben drie verschillende antwoordpatronen vertonen. Het grootste verschil is er tussen Respondent 3 (1, 0) en Respondent 6 (0, 1); de respondenten 1 en 2 (1, 1) lijken qua antwoordpatroon zowel op 3 als op 6. We verwachten daarom dat HOMALS de respondenten 3 en 6 tegenover elkaar in de ruimte afbeeldt met 1 en 2 midden tussen hen in.

Tabel 10.3 Antwoorden van zes respondenten en hun objectscores in de HOMALS-oplossing, met nullen als missing values

Respondent	Vraag 1	Vraag 2	Objectscores	Categorie-kwantificatie
1	1	1	.00	Vraag 1:
2	1	1	.00	ja: .67
3	1	0	2.00	nee: .00
4	–	–	–	Vraag 2:
5	–	–	–	ja: -.67
6	0	1	-2.00	nee: .00

De resultaten van de HOMALS-analyse staan naast de data in Tabel 10.3. Het eerste dat opvalt, is dat er maar één reeks objectscores vermeld is. Door het volledig uitvallen van twee respondenten en door de ontbrekende antwoorden bij twee andere respondenten zijn er in dit kleine voorbeeldje onvoldoende gegevens over om twee (laat staan meer!) dimensies te zoeken. In een praktijkgeval zullen er meestal wel voldoende gegevens overblijven om een meerdimensionale oplossing te krijgen.

In de rechterkolom van Tabel 10.3 staan de categoriekwantificaties van de antwoorden op de twee vragen. Het ja-antwoord op Vraag 1 is gegeven door de respondenten 1, 2 en 3. Hun gemiddelde objectscore is $(0 + 0 + 2.00)/3 = .67$ en deze waarde wordt gekozen als de coördinaat voor de categorie ja op Vraag 1. Het ja-antwoord op Vraag 2 is gekozen door respondenten 1, 2 en 6, van wie de gemiddelde objectscore -.67 bedraagt. De niet-gekozen antwoordcategorieën, die hier als ontbrekende gegevens zijn opgevat, krijgen automatisch nullen als kwantificaties.

Het tweede dat opvalt, is dat de kwadratensom van de objectscores niet gelijk aan 4.00 is, maar aan 8.00. Dat komt doordat HOMALS de objectscores zodanig standaardiseert dat $\sum_i v_i x_{is} = 0$ en $\sum_i v_i x_{is}^2 = n.m$, waarbij v_i het aantal *valide*, dat wil zeggen, niet-ontbrekende keuzen van persoon i is. Dit betekent dat sommige objectscores (heel) groot kunnen worden, wat tot problemen bij de interpretatie kan leiden (zie de paragraaf over gedegenereerde oplossingen). In het geval er geen missing values zijn, is v_i altijd gelijk aan m, het aantal vragen of objecten, zodat $\sum_i v_i x_{is}^2 = m \sum_i x_{is}^2 = m.n$ en dus $\sum_i x_{is}^2 = n$. Is v_i gemiddeld kleiner dan m, dan wordt $\sum_i x_{is}^2$ groter dan n. Een derde verschil ten opzichte van HOMALS zonder ontbrekende data betreft de discriminatiewaarden. In dit geval zijn voor beide vragen de discriminatiewaarden gelijk aan $(3 \times .67^2)/n = 1.333/4 = .333$. Het probleem van deze discriminatiewaarde is dat hij niet langer tussen 0 en 1 hoeft te liggen. Als sommige respondenten zeer extreme objectscores krijgen, dan kunnen de kwantificaties van de door hen gekozen categorieën eveneens extreem worden, waardoor de discriminatiewaarden

groter dan 1 kunnen worden. Daardoor zijn deze coëfficiënten niet meer als percentage verklaarde variantie te interpreteren[3].
Inhoudelijk komt de oplossing van Tabel 10.3 erop neer dat de twee respondenten (3 en 6) met maximaal verschillende antwoordpatronen (1-0 en 0-1) maximaal van elkaar onderscheiden worden. De respondenten 1 en 2, die beiden het antwoordpatroon 1-1 hebben, komen vanzelfsprekend tussen de twee extreme personen in te liggen. Inhoudelijk is daar veel voor te zeggen: deze respondenten hebben als het ware geen voorkeur voor het ene ja-antwoord boven het andere.

Uitbijters en gedegenereerde oplossingen

In de paragraaf 'keuzedata als nabijheid' is een aantal eigenschappen genoemd die een afbeelding van keuzedata idealiter zou moeten bezitten. Omdat een persoon dicht bij de objecten afgebeeld wordt die hij of zij gekozen heeft, kan het niet anders dan dat iemand die (zeer) veel objecten gekozen heeft, in het midden van de configuratie terechtkomt. Hetzelfde geldt voor een object dat door (zeer) veel personen gekozen is.
Het omgekeerde is niet automatisch waar. Een persoon die maar één object gekozen heeft, kan best in het midden terechtkomen, als tenminste het betreffende object in het midden van de afbeelding ligt. Dit object is dan waarschijnlijk door veel anderen gekozen. Ook een object dat maar door één persoon gekozen is, kan best in het midden liggen als die ene persoon daar is gelokaliseerd. Die persoon heeft dan waarschijnlijk ook nog veel andere objecten gekozen.
Naarmate de objecten die een persoon gekozen heeft door minder anderen gekozen zijn, zal de desbetreffende persoon meer aan de rand van de configuratie terechtkomen. Idem: naarmate een object minder vaak samen met andere objecten gekozen is, zal dat object meer aan de buitenkant worden geplaatst. Dit kan aanleiding geven tot het volgende probleem.
Als één of enkele personen één of enkele objecten gekozen hebben die niet ook door anderen gekozen zijn, dan worden deze (unieke) personen en objecten extreem ver weg van de rest van de groep geplaatst (dat geeft in HOMALS namelijk de grootste discriminatiewaarden). Dit kan op zo'n extreme manier gebeuren dat een bepaalde dimensie alleen maar het onderscheid weergeeft tussen deze unieke personen en de rest. Alle verdere structuur in de keuzedata gaat dan als het ware verloren. In zo'n geval spreken we van *uitbijters* die een *gede-*

3 Een betere maat voor het discriminerend vermogen van het ja-antwoord op vraag 1 zou in dit geval zijn: $(3 \times .667^2)/(0^2 + 0^2 + 2.00^2)$, dat wil zeggen, het door de categoriekwantificatie verklaarde deel van de som van de gekwadrateerde objectscores van alle personen die op Vraag 1 antwoord ja gekozen hebben. Als de groep ja-kiezers volstrekt homogeen is, dat wil zeggen, geen onderlinge fluctuaties vertoont, dan wordt bovenstaande coëfficiënt gelijk aan 1; zijn ze volstrekt heterogeen, dan ligt de categoriekwantificatie in het nulpunt en wordt deze coëfficiënt gelijk aan nul. In dit voorbeeld is de uitkomst identiek aan de discriminatiewaarde uit HOMALS (.333) maar dat hoeft niet altijd zo te zijn.

genereerde oplossing veroorzaken. Wat men in zo'n geval kan doen is de uitbijter(s) verwijderen, of de bijbehorende objecten verwijderen (eventueel: samenvoegen met andere objecten). Daarna moet men de HOMALS-analyse toepassen op de overgebleven data en hopen dat er niet opnieuw personen met unieke keuzepatronen tussen zitten die de oplossing verstoren.

BLOK 10.1 EEN HOMALS-ANALYSE VAN KEUZEDATA

Vierentwintig studenten die een cursus meerdimensionale schaaltechnieken volgden, is onderstaand lijstje met twaalf boeken uit de Nederlandse literatuur voorgelegd:

- G.K. van het Reve: *De avonden*
- W.F. Hermans: *De donkere kamer van Damocles*
- H. Mulisch: *Het stenen bruidsbed*
- J. Wolkers: *Kort Amerikaans*
- R. Giphart: *Giph*
- J. Zwagerman: *Gimmick*
- C. Palmen: *De vriendschap*
- M. de Moor: *Eerst grijs dan wit dan blauw*
- N. Noordervliet: *De naam van de vader*
- A.F.Th. van der Heijden: *Asbestemming*
- L. de Winter: *De ruimte van Sokolov*
- A. van Dis: *Indische duinen.*

De studenten werd gevraagd aan te willen kruisen welke boeken zij gelezen hadden, dat wil zeggen, *pick any of 12*. Hun antwoorden zijn verzameld in een personen × objecten matrix met enen (gelezen) en nullen (niet gelezen). Deze matrix is via de volgende commando's met HOMALS geanalyseerd.

```
data list free/stud reve herm muli wolk  giph zwag palm moor
            noor heij wint vdis.
begin data.
 1 1 1 1 1 0 0 0 1 0 0 0 1   2 0 0 1 1 1 1 0 0 1 0 0 1
 3 1 1 0 1 0 0 0 0 0 0 0 0   4 0 1 1 1 0 0 1 1 0 0 0 0
 5 1 1 0 1 0 0 0 1 0 1 0 0   6 0 1 1 1 0 1 0 0 1 0 1 0
 7 1 0 1 1 0 0 1 1 0 1 0 0   8 1 1 0 1 0 1 1 0 0 0 0 0
 9 1 1 0 1 0 0 1 1 0 0 0 0  10 1 1 1 1 0 1 0 1 0 1 1 0
11 1 1 0 1 0 0 1 0 0 0 1 0  12 1 0 1 1 0 0 0 0 0 0 1 1
13 1 1 1 0 0 0 0 0 1 0 0 1  14 1 0 1 1 0 0 0 0 0 0 1 1
15 1 1 1 0 0 0 1 0 0 1 0 0  16 1 1 1 0 0 0 0 1 0 0 0 0
```

```
17 1 0 1 0 1 0 1 0 1 0 0 0  18 0 0 0 1 1 1 1 1 0 1 0 1
19 1 1 0 0 0 0 0 0 0 1 1 1  20 1 1 0 0 0 0 1 1 0 0 1 0
21 1 1 0 1 0 0 1 0 0 0 1 0  22 0 1 1 1 0 0 0 0 0 0 1 0
23 1 1 0 0 0 1 0 0 1 1 0 0  24 0 1 1 0 0 0 1 1 0 0 0 1
end data.
recode reve to vdis (0=2).
value labels stud 1 `1' 2 `2' 3 `3' 4 `4' 5 `5' 6 `6' 7 `7' 8 `8'
   9 `9' 10 `10' 11 `11' 12 `12' 13 `13' 14 `14' 15 `15' 16 `16'
 17 `17' 18 `18' 19 `19' 20 `20' 21 `21' 22 `22' 23 `23' 24 `24'.
value labels reve 1 `re' 2 `0' /herm 1 `he' 2 `0'.
value labels muli 1 `mu' 2 `0' /wolk 1 `wo' 2 `0'.
value labels giph 1 `gi' 2 `0' /zwag 1 `zw' 2 `0'.
value labels palm 1 `pa' 2 `0' /moor 1 `mo' 2 `0'.
value labels noor 1 `no' 2 `0' /heij 1 `hy' 2 `0'.
value labels wint 1 `wi' 2 `0' /vdis 1 `vd' 2 `0'.

homals variables stud (24) reve to vdis (1)
    /analysis reve to vdis
    /dimension 2
    /print default object
    /plot default object (stud reve herm muli wolk giph zwag
                  palm moor noor heij wint vdis).
```

De opdracht `recode reve to vdis (0=2)` is niet nodig voor de analyse, maar is handig omdat HOMALS daardoor duidelijkere grafieken oplevert. Er is gekozen voor een tweedimensionale afbeelding. De belangrijkste uitvoer van HOMALS wordt hieronder afgedrukt.

```
MARGINAL FREQUENCIES
====================
VARIABLE  MISSING  CATEGORIES
             1
REVE         6     18
HERM         6     18
MULI        10     14
WOLK         8     16
GIPH        21      3
ZWAG        18      6
PALM        13     11
MOOR        14     10
NOOR        19      5
HEIJ        17      7
WINT        15      9
VDIS        16      8
```

```
*** Warning: Variable REVE has all valid objects (18) in the same category (1).
*** Warning: Variable HERM has all valid objects (18) in the same category (1).
*** Warning: Variable MULI has all valid objects (14) in the same category (1).
*** Warning: Variable WOLK has all valid objects (16) in the same category (1).
*** Warning: Variable GIPH has all valid objects (3)  in the same category (1).
*** Warning: Variable ZWAG has all valid objects (6)  in the same category (1).
*** Warning: Variable PALM has all valid objects (11) in the same category (1).
*** Warning: Variable MOOR has all valid objects (10) in the same category (1).
*** Warning: Variable NOOR has all valid objects (5)  in the same category (1).
*** Warning: Variable HEIJ has all valid objects (7)  in the same category (1).
*** Warning: Variable WINT has all valid objects (9)  in the same category (1).
*** Warning: Variable VDIS has all valid objects (8)  in the same category (1).
```

HOMALS geeft allereerst aan hoe vaak elke antwoordcategorie gekozen is en hoeveel scores er bij elke variabele ontbreken. In dit geval geeft HOMALS bij elke variabele de waarschuwing dat alle betekenisvolle observaties in één en dezelfde antwoordcategorie zitten. Hier is dat inderdaad het geval en ook de bedoeling; in de meeste andere toepassingen zou het echter kunnen betekenen dat de data verkeerd zijn ingelezen. Daarna wordt het aantal iteraties afgedrukt dat nodig was om een gestabiliseerde oplossing te krijgen, gevolgd door een aantal indicaties voor de *goodness-of-fit*: de eigenwaarden en de discriminatiewaarden op beide dimensies.

```
THE ITERATIVE PROCESS STOPS BECAUSE THE CONVERGENCE CRITERION IS REACHED,
THE NUMBER OF ITERATIONS IS:   23

DIMENSION   EIGENVALUE
-----       -----
   1         .3075
   2         .1993

DISCRIMINATION MEASURES PER VARIABLE PER DIMENSION
==================================================
VARIABLE  DIMENSION
          1     2
REVE       .104    .001
HERM       .213    .000
MULI       .038    .121
WOLK       .042    .030
GIPH      1.051    .116
ZWAG       .497    .054
PALM       .078    .343
MOOR       .219    .374
NOOR      1.179    .018
HEIJ       .000    .319
WINT       .133    .725
VDIS       .135    .292
```

De eigenwaarden zijn niet erg hoog en laten dus zien dat de studenten die een bepaald boek gelezen hebben, lang niet perfect van de niet-lezers onderscheiden worden. Ook zien we dat de eerste dimensie belangrijker is dan de tweede om lezers en niet-lezers van elkaar te onderscheiden. De coördinaten van de studenten worden weergegeven in de tabel met *object scores*. Daarna volgen van elke variabele de antwoordfrequenties (*marginal frequencies*) en de categoriekwantificaties van de antwoordcategorieën.

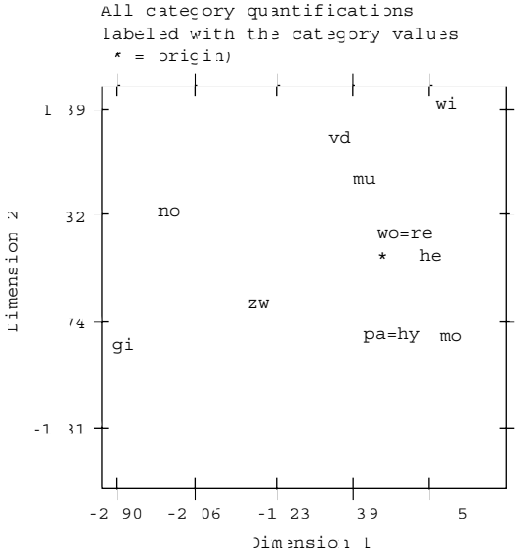
```
THE OBJECT SCORES ARE:
 ======================
OBJECT *     DIMENSION
          1      2
   1 *     .54     .59
   2 *   -3.97     .40
   3 *    1.25     .42
   4 *    1.08   -1.14
   5 *    1.25   -1.75
   6 *   -1.44    1.57
   7 *     .83   -1.79
   8 *     .10   -1.09
   9 *    1.49   -1.56
  10 *     .34    -.23
  11 *    1.41     .78
  12 *     .21    3.05
  13 *   -1.54    1.73
  14 *     .21    3.05
  15 *     .71   -1.42
  16 *    1.12    -.56
  17 *   -3.09   -1.04
  18 *   -1.64   -2.24
  19 *     .58    1.32
  20 *    1.72    -.38
  21 *    1.41     .78
  22 *     .91    2.58
  23 *   -1.85   -1.20
  24 *     .51    -.41

        MARGINAL FREQUENCIES AND CATEGORY QUANTIFICATIONS
        ================================================

 VARIABLE: REVE
 - - - - - - -
 CATEGORY                    MARGINAL FREQUENCY
 - - - - - -                       - - - - - - - - - - - - -
    1   re                    18
              MISSING:         6

 CATEGORY QUANTIFICATIONS
 - - - - - - - - - - - - - - - - - -
 CATEGORY   DIMENSIONS
 - - - - - -
          1      2
     1    .37    .04
==============================================================================
=======
```

De HOMALS-uitvoer met betrekking tot de variabelen is hierboven alleen

voor de eerste variabele REVE afgedrukt. In verband met de ruimte zijn de categoriekwantificaties hier in onderstaande tabel samengenomen.

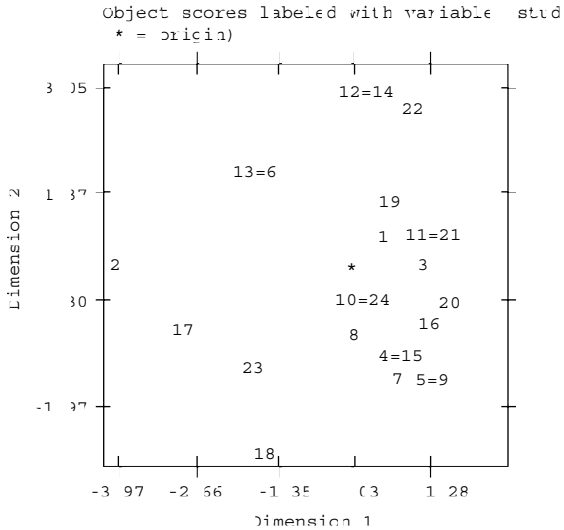
VARIABLE	CATEGORY	MARGINAL FREQUENCIES	CATEGORY QUANTIFICATIONS DIM1	DIM2
REVE	1 re	18	.37	.04
HERM	1 he	18	.53	.00
MULI	1 mu	14	-.25	.46
WOLK	1 wo	16	.25	.21
GIPH	1 gi	3	-2.90	-.96
ZWAG	1 zw	6	-1.41	-.47
PALM	1 pa	11	.41	-.86
MOOR	1 mo	10	.72	-.95
NOOR	1 no	5	-2.38	.29
HELJ	1 hy	7	.03	-1.05
WINT	1 wi	9	.60	1.39
VDIS	1 vd	8	-.64	.94

Hieronder de door HOMALS afgedrukte grafiek van de categoriekwantificaties.



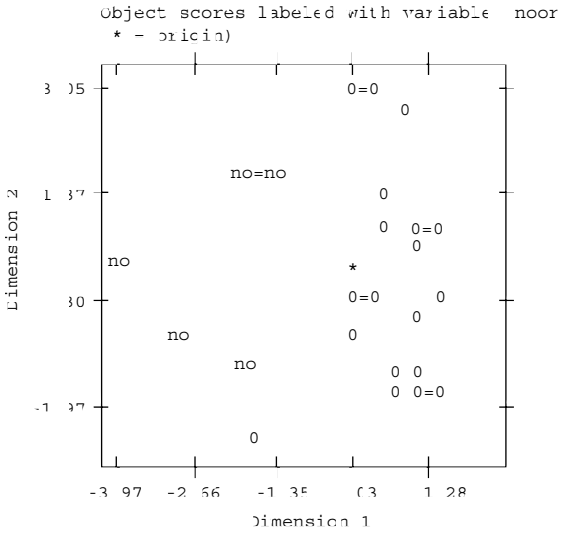
Deze configuratie kan als volgt geïnterpreteerd worden. In het midden

zien we een cluster boeken, geschreven door de 'grote vier' van de naoorlogse Nederlandse literatuur: Mulisch, Hermans, Van het Reve en Wolkers. Deze boeken zijn door veel studenten gelezen; op één na alle studenten hebben twee, drie of alle vier de boeken gelezen. Om deze 'klassieke kern' ligt een schil met boeken die door jongere auteurs geschreven zijn. Deze boeken zijn door minder studenten gelezen en studenten die één van deze boeken gelezen hebben, hebben niet ook steeds de meeste overige gelezen. Er is één cluster van recente boeken te onderscheiden (Palmen, Van der Heijden en De Moor) en verder een aantal min of meer geïsoleerde (Van Dis, De Winter, Noordervliet, Giphart en Zwagerman). De onderlinge afstanden van alle boeken hebben te maken met het aantal gemeenschappelijke lezers. Hoe vaker ze door dezelfde studenten gelezen zijn, hoe dichter ze bij elkaar liggen. Onderstaande grafiek van de objectscores toont een afbeelding van de studenten. Punten die in de oorspronkelijke grafiek samenvallen, zijn afzonderlijk afgebeeld en door middel van een '=' met elkaar verbonden.

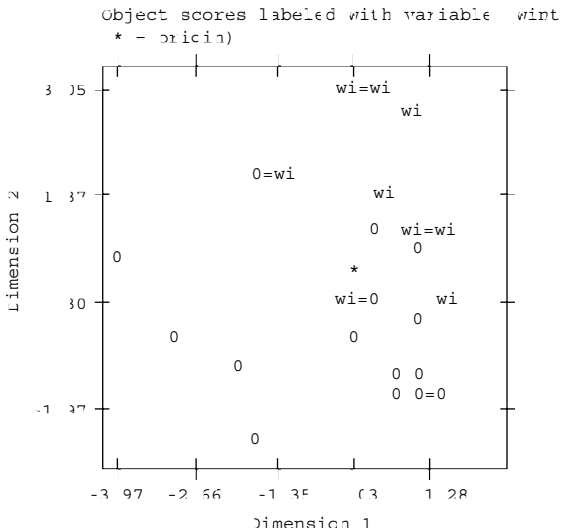


Bovenstaande figuur is het resultaat van de opdracht `/PLOT OBJECT (STUD)`. De punten voor de objecten worden gelabeld met de *value labels* van de variabele STUD. Door de objectpunten steeds met de value labels van andere variabelen te labelen krijgt men meer inzicht in de betekenis van de ruimte. Hieronder volgt de grafiek van objectscores, gelabeld met de variabele NOOR, dat wil zeggen, dat bij elk objectpunt is aangegeven

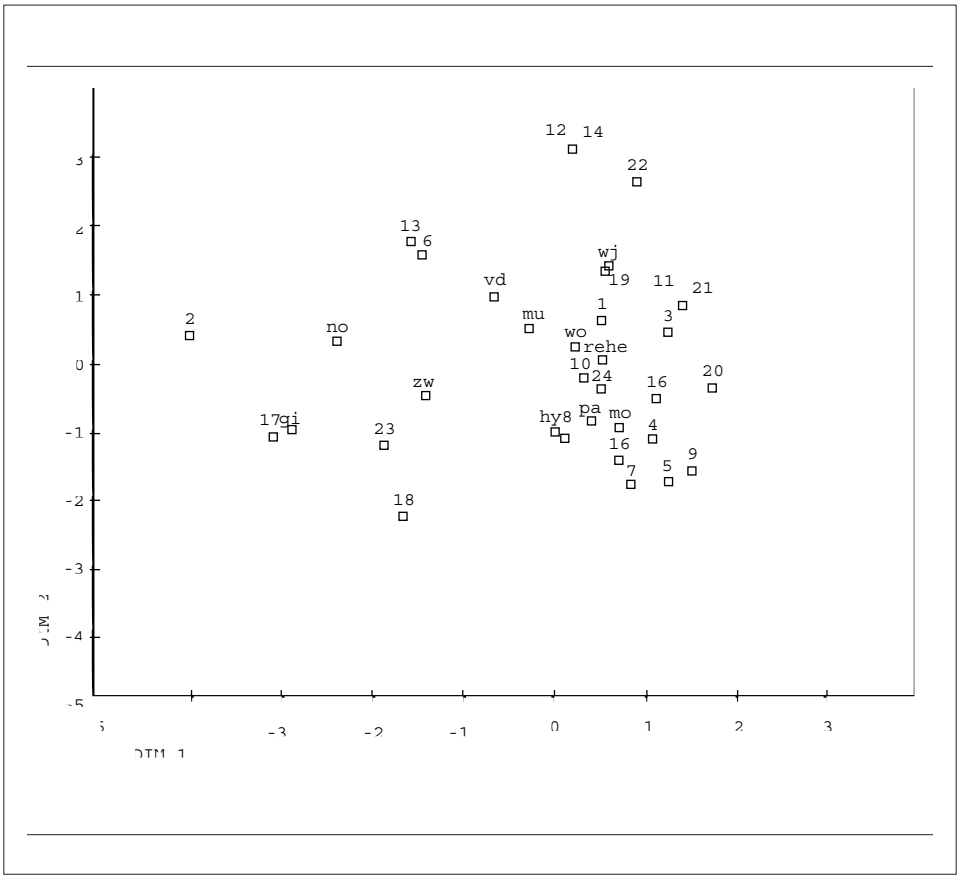
of de desbetreffende student het boek van Nelleke Noordervliet al of niet gelezen heeft. Deze figuur laat zien dat alle studenten die het boek van Nelleke Noordervliet gelezen hebben aan de linkerkant van de configuratie liggen. Zij worden perfect onderscheiden van de



studenten die dit boek niet gelezen hebben. Daardoor is de discriminatiewaarde van deze variabele hoog.
Uit de discriminatiewaarden blijkt dat vooral de boeken van Noordervliet en Giphart bijdragen aan de variatie tussen de studenten in horizontale richting. Deze boeken zijn uitsluitend gelezen door studenten die links in de configuratie liggen. De verticale dimensie onderscheidt de studenten die wél het boek van De Winter hebben gelezen van hen die dat niet hebben gedaan. Hieronder staat de grafiek van objectscores, gelabeld met de variabele WINT. In dit voorbeeld is het niet echt mogelijk een inhoudelijke interpretatie van de dimensies te geven. De afbeelding is niet meer dan een grafische weergave van de data.



Ten slotte geven we hieronder een zogenaamde *joint plot*: de studenten en boeken in één grafiek. We zien nu duidelijk dat de boeken zijn afgebeeld te midden van de studenten die ze gelezen hebben. De boeken nemen minder extreme posities in dan de studenten. Boeken die door veel studenten gelezen zijn liggen meer in het midden, weinig gelezen boeken liggen meer naar buiten toe. Ook studenten die weinig gelezen hebben, liggen meer aan de buitenkant van de configuratie; studenten die meer boeken hebben gelezen, liggen wat meer in het centrum.



10.5 KEUZEDATA ALS DOMINANTIERELATIES

Het latente-trekmodel voor keuzedata

Wanneer we twee personen met elkaar vergelijken, waarvan de ene (Persoon A) een bepaald object (O) wél gekozen heeft en de ander (Persoon B) dat niet heeft gedaan, dan kunnen we dit gegeven soms als volgt interpreteren. Persoon A overtreft Persoon B met betrekking tot één of andere eigenschap die maakt dat A object O wel kiest en B object O niet. Die eigenschap zou men honger, lust of hebzucht kunnen noemen, al naargelang de aard van het object. In een andere context zou men die eigenschap als begaafdheid kunnen interpreteren. Bijvoorbeeld: Persoon A is begaafd genoeg om het juiste antwoord op vraag O te kiezen en Persoon B is dat niet. In dat geval hebben we te maken met *dominantiedata*: personen en objecten kunnen elkaar al dan niet overtreffen of domineren.

Om van bovengenoemd type data een ruimtelijke afbeelding te maken, kan men elke vraag (ieder object) voorstellen door middel van *twee punten in de ruimte*: een punt voor het correcte antwoord en een punt voor het foute antwoord. Personen die het juiste antwoord geven (het object kiezen) liggen in de buurt van het eerste punt, personen die het foute antwoord geven (het object niet kiezen) liggen dicht bij het tweede punt. Dit is het soort afbeelding dat men met HOMALS kan verkrijgen. Een andere manier waarop men elke vraag of elk object ruimtelijk kan afbeelden, is door middel van een lijn die de ruimte van personen in tweeën deelt: aan de ene kant liggen de mensen met het juiste antwoord (de kiezers) aan de andere kant de personen met het foute antwoord (de niet-kiezers). We zijn dan terechtgekomen in een familie van modellen die men zou kunnen aanduiden als *latente-trekmodellen*. Iedere vraag (elk object) deelt de denkbeeldige (latente) ruimte van personen in tweeën door middel van een scheidingslijn. De dimensies van die ruimte zijn *latente trekken*, waarop zowel de personen als de vragen posities (scores, coördinaten, schaalwaarden) innemen. De term latente-trekmodel wordt echter voornamelijk gebruikt voor die gevallen dat de vragen en personen op één dimensie kunnen worden afgebeeld.
In dit hoofdstuk zullen we laten zien hoe keuzedata die in het latente-trekmodel passen, ook met behulp van HOMALS geanalyseerd kunnen worden. Aangezien het meerdimensionale geval al eerder besproken is, zullen we vooral aandacht besteden aan twee eendimensionale varianten, de Guttman- en de Mokkenschaal.

HOMALS en het latente-trekmodel voor keuzedata

Een afbeelding van keuzedata volgens het latente-trekmodel impliceert dat niet alleen de kiezers van een object bij elkaar in hetzelfde deel van de ruimte moeten liggen, maar ook dat de niet-kiezers van een object bij elkaar in de buurt moeten liggen. Ook de niet-kiezers hebben volgens het model dus iets gemeenschappelijks. Dit alles betekent dat we empirische keuzedata ook volgens het vectormodel kunnen analyseren met behulp van HOMALS, als we er maar voor zorgen dat ook de nullen als valide observaties meedoen. Daartoe moeten we de observaties hercoderen (bijvoorbeeld: 0 wordt 1 = niet-gekozen; 1 wordt 2 = wel gekozen) en moeten we aangeven dat elke variabele twee categorieën heeft. We krijgen dan een oplossing van het type dat al eerder in Tabel 10.2 en Figuur 10.1 geïllustreerd is. Aan dat voorbeeld hoeven we op dit moment niet veel meer toe te voegen.
Hieronder zullen we echter een bijzonder geval van keuzedata analyseren. In dit voorbeeld gaat het om het speciale geval dat de objecten en personen een zogenaamde *Guttmanschaal* vormen. De objecten en personen zijn dan op een bijzondere manier geordend op een eendimensionale schaal. We zullen de desbetreffende data op twee manieren analyseren: in de eerste plaats met HOMALS in twee dimensies (met andere woorden: we doen alsof we niet van tevoren weten dat de desbetreffende data eendimensionaal zijn) en in de tweede plaats

met een methode die er *a priori* van uitgaat dat de data een eendimensionale structuur hebben: de *Mokkenschaal*-analyse.

De Guttmanschaal

In Hoofdstuk 1 is een onderzoek van Van der Kloot en Willemsen (1991) beschreven. In die studie ging het om de vraag hoe belangrijk bepaalde bijdragen aan wetenschappelijk onderzoek zijn voor het vaststellen van de auteursvolgorde. Een van de vragen die aan de proefpersonen gesteld is, luidde:

Wie van onderstaande personen moeten volgens u in ieder geval als (mede)auteur van een publicatie vermeld worden?

- *iemand die uitsluitend het onderzoek **bedacht** heeft?* *ja/nee*
- *iemand die uitsluitend **de leiding** over het onderzoek heeft gehad?* *ja/nee*
- *iemand die in het onderzoek uitsluitend de **dataverzameling** gedaan heeft?* *ja/nee*
- *iemand die in het onderzoek uitsluitend de **data-analyse** verricht heeft?* *ja/nee*
- *iemand die uitsluitend de tekst van de publicatie **geschreven** heeft?* *ja/nee*

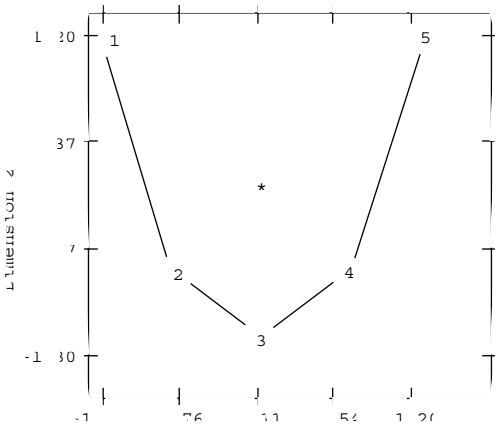
De uitkomsten van de in Hoofdstuk 1 besproken Thurstone-analyse, lieten zien dat de belangrijkheidsvolgorde van deze bijdragen schrijven, bedenken, leidinggeven, data-analyse en dataverzameling was.
Op grond van die resultaten mogen we nu het volgende verwachten. Als een respondent vindt dat ook de minst belangrijke bijdrage (dataverzameling) al genoeg is om voor auteurschap in aanmerking te komen, zal deze respondent ook vinden dat alle andere (belangrijker) bijdragen daarvoor in aanmerking komen. Vindt iemand dat data-analyse voor auteurschap in aanmerking komt, dan verwachten we dat deze persoon ook vindt dat schrijven, bedenken en leidinggeven tot auteurschap moet leiden. Wat zo iemand van dataverzameling vindt, is daarentegen onzeker. Sommige personen die data-analyse voor auteurschap in aanmerking vinden komen, zullen ook dataverzameling voldoende vinden, anderen echter niet.
Als de meningen van de proefpersonen *perfect* met bovengenoemde verwachtingen overeenkomen, dan zullen alleen maar de antwoordpatronen uit Tabel 10.4 voor mogen komen. In die tabel zijn de personen (de rijen) geordend van respondenten met weinig ja-antwoorden (enen) naar respondenten die vaak ja zeggen. De bijdragen (de kolommen) lopen van bijdragen met veel ja-antwoorden naar bijdragen met weinig enen. We zien nu dat de enen en nullen (de nee-antwoorden) in de tabel in twee perfecte driehoeken uiteenvallen. In dat geval vormen de objecten en de proefpersonen een perfect *scalogram* of, naar degene die deze ideeën ontwikkelde, een *Guttmanschaal.*

Tabel 10.3 Antwoordpatronen volgens een perfecte Guttmanschaal

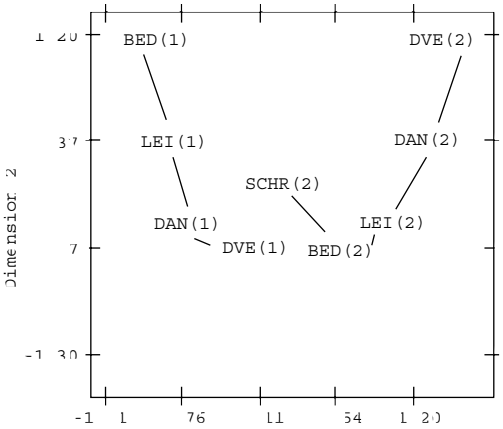
patroon	schrijven	bedenken	leiding	analyse	verz
1	1	0	0	0	0
2	1	1	0	0	0
3	1	1	1	0	0
4	1	1	1	1	0
5	1	1	1	1	1

HOMALS op een perfecte Guttmanschaal

Om de perfecte Guttmanschaal-data uit Tabel 10.4 met HOMALS te analyseren moeten we de enen en nullen eerst hercoderen, bijvoorbeeld tot tweeën voor de ja-antwoorden en enen voor de nee-antwoorden. Een analyse van deze data in twee dimensies levert de resultaten op die in Figuur 10.2 en 10.3 zijn weergegeven.



Figuur 10.2 Tweedimensionale afbeelding van de objectscores uit de HOMALS-analyse van de keuzedata van Tabel 10.3



Figuur 10.3 Tweedimensionale afbeelding van de categoriekwantificaties uit de homals-analyse van de keuzedata van Tabel 10.3

Het bijzondere van deze HOMALS-oplossing (met eigenwaarden .50 en .1667) is dat de objecten een hoefijzervormige configuratie vormen. Dit is een signaal dat de geanalyseerde data in wezen eendimensionaal zijn. Omdat HOMALS de opdracht krijgt een tweedimensionale afbeelding te maken, moet HOMALS twee stel coördinaten voor de objecten zoeken die maximaal tussen de objecten discrimineren en onderling onafhankelijk zijn. Als de data eendimensionaal zijn, bestaat er eigenlijk maar één stel coördinaten die maximaal discrimineren. In zo'n geval kiest HOMALS voor de tweede dimensie coördinaten die een functie zijn van de coördinaten op de eerste dimensie, maar er niet mee gecorreleerd zijn. Een van de mogelijke oplossingen is om voor de coördinaten op de tweede as een tweede-machtsfunctie van die van de eerste dimensie te nemen. De grafiek van de objecten heeft dan de vorm van een parabool, een hoefijzer.

Als de objectscores de vorm van een hoefijzer hebben, hebben de categoriekwantificaties eveneens zo'n vorm. In Figuur 10.3 zien we twee van zulke hoefijzers: één van de ja-antwoorden (2) en één van de nee-antwoorden (1). De coördinaten van de ja-antwoorden op de eerste dimensie kunnen we nu beschouwen als de schaalwaarden van de bijdragen. Deze schaalwaarden zijn 0 voor schrijven, .35 voor bedenken, .71 voor leidinggeven, 1.06 voor data-analyse en 1.41 voor dataverzameling. De belangrijkste bijdrage, dat wil zeggen, de bijdrage die het meest in aanmerking komt voor auteurschap, heeft hier de laagste schaalwaarde; de minst belangrijke bijdrage, die het minst voor auteurschap in aanmerking komt, heeft de grootste schaalwaarde. Omdat elk antwoordpatroon even vaak voorkomt (namelijk slechts één keer) zijn de verschillen tussen de opeenvolgende schaalwaarden in dit voorbeeld aan elkaar gelijk.

Als we in de praktijk oplossingen vinden die er precies of nagenoeg hetzelfde als boven uitzien, dan hebben we waarschijnlijk te maken met gegevens die zonder verlies van informatie in één dimensie weergegeven kunnen worden. Als alle nee-antwoorden bovendien aan één kant van de ruimte liggen en alle ja-antwoorden aan de andere kant, dan vormen de data kennelijk een Guttmanschaal.

Tabel 10.5 Antwoorden (2 = ja; 1 = nee) op de vraag of een wetenschappelijke bijdrage voor auteurschap in aanmerking komt

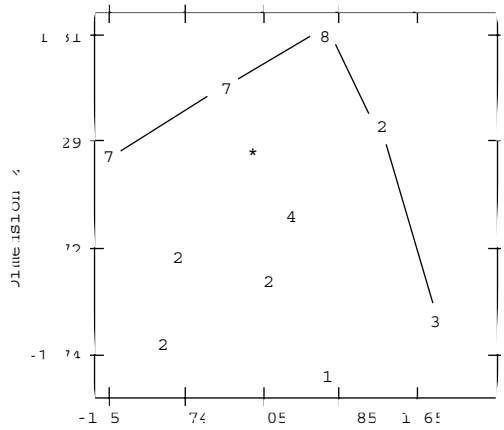
pp[a]	s	b	l	a	v	pp	s	b	l	a	v
1	2	1	1	1	1	20	2	2	2	1	1
2	2	1	1	1	1	21	2	2	2	1	1
3	2	1	1	1	1	22	2	2	2	1	1
4	2	1	1	2	1	23	2	2	2	1	1
5	2	1	2	2	1	24	2	2	2	1	1
6	2	1	2	2	1	25	2	2	2	2	1
7	2	1	2	2	2	26	2	2	2	2	1
8	2	1	2	2	2	27	2	2	2	2	1
9	2	2	1	1	1	28	2	2	2	2	1
10	2	2	1	1	1	29	2	2	2	2	1
11	2	2	1	2	1	30	2	2	2	2	1
12	2	2	1	2	1	31	2	2	2	2	1
13	2	2	1	2	1	32	2	2	2	2	2
14	2	2	1	2	1	33	2	2	2	2	2
15	2	2	1	2	2	34	2	2	2	2	2
16	2	2	1	2	2	35	2	2	2	2	2
17	2	2	2	1	1	36	2	2	2	2	2
18	2	2	2	1	1	37	2	2	2	2	2
19	2	2	2	1	1	38	2	2	2	2	2

[a] Twee van de 40 proefpersonen ontbreken vanwege incomplete antwoorden

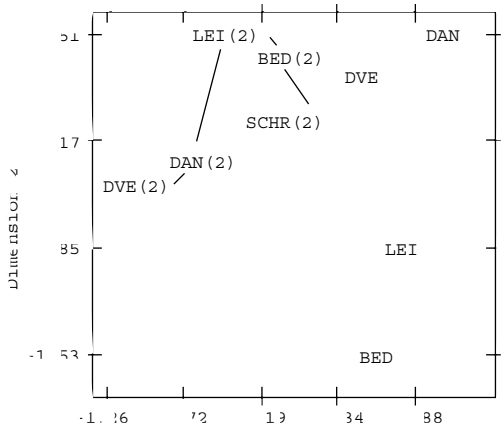
Een HOMALS-analyse van empirische data

In Tabel 10.5 staan de echte gegevens die in het onderzoek van Van der Kloot en Willemsen (1991) geobserveerd zijn. Deze gegevens van 38 personen zijn zodanig gesorteerd dat de respondenten met het kleinste aantal tweeën (ja-antwoorden) bovenaan staan en dat de objecten (de bijdragen) van links naar rechts geordend zijn van de bijdrage met de meeste tweeën (schrijven) naar de bijdrage met de minste tweeën (dataverzameling). Ook nu valt het typische driehoekspatroon van een Guttmanschaal te onderkennen, zij het dat het

patroon niet perfect is. Elf personen (29 procent) hebben antwoordpatronen die niet volledig met de onderliggende schaal in overeenstemming zijn (zie de vetgedrukte antwoordcodes in Tabel 10.5).
De HOMALS-oplossing van deze data (met eigenwaarden .3139 en .2232) is weergegeven in de Figuren 10.4 en 10.5. In de eerste figuur staan de objectscores, in de tweede de kwantificaties van de antwoordcategorieën.



Figuur 10.4 Tweedimensionale afbeelding van de objectscores uit de HOMALS-analyse van de keuzedata van Tabel 10.5



Figuur 10.5 Afbeelding van categoriekwantificaties uit de HOMALS-analyse van de data van Tabel 10.5

De objectscores van de personen met antwoordpatronen die perfect in een Guttmanschaal passen, zijn in Figuur 10.4 met elkaar verbonden. Zij vormen weer een hoefijzerachtige structuur. Elf personen liggen nogal ver naast het hoefijzer. Dit zijn de respondenten met antwoorden die niet volledig in de

Guttmanschaal passen (de vetgedrukte antwoorden in Tabel 10.5). Het plaatje laat dus duidelijk zien dat de data geen perfecte Guttmanschaal vormen. In Figuur 10.5 zijn de categoriekwantificaties van de ja-antwoorden met elkaar verbonden. Ook zij vormen een hoefijzerachtige figuur. De bijbehorende coördinaten op de eerste dimensie zijn 0 voor schrijven, -.15 voor bedenken, -.34 voor leidinggeven, -.55 voor data-analyse en -1.26 voor dataverzameling[4].

Wat is nu de 'winst' van deze HOMALS-analyse boven die van de Thurstonemethode, nu we gezien hebben dat beide methoden in dit geval vergelijkbare schaalwaarden opleveren? Die winst zit hem in het feit dat in de HOMALS-analyse ook informatie over de afzonderlijke proefpersonen en hun relaties tot de objecten bewaard blijft. We zien aan de oplossing dat er proefpersonen zijn die vinden dat elke bijdrage belangrijk genoeg is voor auteurschap, en we zien dat er proefpersonen zijn die alleen schrijven belangrijk genoeg vinden. Bovendien zien we aan het patroon van én objectscores én categoriekwantificaties dat de data mogelijk een Guttmanschaal vormen, zij het geen perfecte.

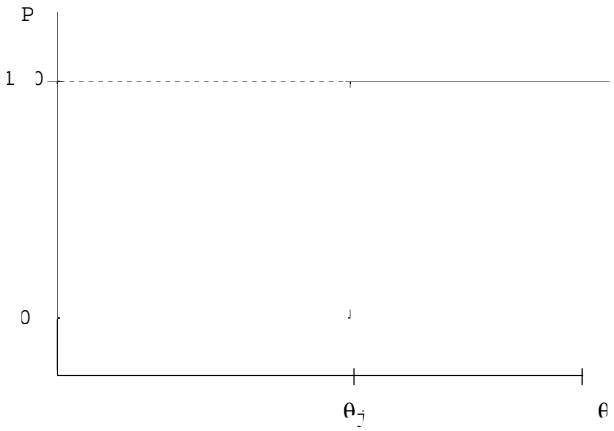
De item-karakteristieke curve

De essentie van het door Guttman (1944, 1950) ontwikkelde scalogrammodel is de gedachte dat personen te ordenen zijn langs één onderliggende *latente dimensie* en dat elk object (of: elke vraag van een psychologische test) deze dimensie in twee stukken verdeelt: een gebied waarin de personen liggen die het object niet kiezen (of: de vraag fout beantwoorden) en een gebied met de proefpersonen die het object wél kiezen (de vraag wél goed beantwoorden). De personen verschillen dus ten opzichte van elkaar met betrekking tot een eigenschap die door de latente dimensie vertegenwoordigd wordt. Dat kan van alles zijn, bijvoorbeeld een attitude, een persoonlijkheidseigenschap of een vaardigheid.

Verschillende objecten delen de latente dimensie op verschillende plaatsen in tweeën. De populaire objecten (in dit geval: de gemakkelijkere vragen) doen dit bij een lagere waarde op de dimensie dan de objecten die minder populair zijn (in dit geval de vragen die moeilijker zijn). Het scalogrammodel impliceert dat respondenten die een minder populair object kiezen ook alle populairdere items zullen kiezen en dat personen die een moeilijker item goed

4 Deze schaalwaarden correleren .822 met de schaalwaarden die we in Hoofdstuk 1 met de Thurstonemethode gevonden hebben. Die correlatie is hoog, maar niet perfect. Dat hoeft ook niet, omdat de gegevens waarop de HOMALS-schaalwaarden gebaseerd zijn, maar een deel vormen van de data die voor de Thurstone-analyse gebruikt zijn. De HOMALS-analyse betreft alleen de gegevens die in de onderste rijen van de Tabellen 1.1, 1.2 en 1.3 zijn samengevat. Stellen we de schaalwaarde van schrijven op nul, dan leveren de *z*-waarden uit de onderste rij van Tabel 1.3 de schaalwaarden -1.495, -1.771, -1.898 en -2.811 voor bedenken, leidinggeven, data-analyse en dataverzameling op. Deze waarden correleren .869 met de HOMALS-schaalwaarden: iets hoger maar nog steeds niet perfect. Ze blijken overigens wél nagenoeg perfect (r = .975) te correleren met de wortel uit de HOMALS-schaalwaarden. Dit is het gevolg van de verschillende manieren waarop Thurstone en HOMALS frequenties in schaalwaarden omzetten.

kunnen beantwoorden ook alle makkelijkere items goed beantwoorden. De kenmerken van het scalogrammodel kunnen we als volgt formuleren. Elk object *j* heeft een schaalwaarde θ_j op het latente continuüm Θ en elke proefpersoon *i* heeft een schaalwaarde θ_i. Voor de kans $P_{ij} = P_{j/\theta i}$ dat Object *j* door Persoon *i* (met θ_i) gekozen wordt (of goed wordt beantwoord) geldt dat $P_{ij} = 1$ als $\theta_i - \theta_j > 0$ en dat $P_{ij} = 0$ als $\theta_i - \theta_j \leq 0$. Deze relatie is weergegeven in Figuur 10.6, waarin P_{ij} is afgebeeld als functie van θ_i voor een bepaald object met schaalwaarde θ_j. Deze functie wordt in de test-theorie een *item-karakteristieke curve* of een *trace-line* genoemd. In het scalogrammodel bestaat deze curve uit twee horizontale lijnen, een lijn voor $P_{ij} = 0$ en een lijn voor $P_{ij} = 1$. De item-karakteristieke curves van verschillende objecten verschillen van elkaar met betrekking tot de locatie op het latente continuüm, waar de sprong van $P_{ij} = 0$ naar $P_{ij} = 1$ plaatsvindt.



Figuur 10.6 Item-karakteristieke curve voor de keuze van een object volgens het Guttmanschaalmodel

Schaalbaarheid

In essentie is het scalogrammodel een *deterministisch model*: het schrijft dwingend voor welke observaties wél en welke observaties niet voor mogen komen. Zo gauw er één observatie voorkomt die niet in het model past, zouden we in principe het model moeten verwerpen. Dat is natuurlijk zonde als er maar weinig observaties voorkomen die niet met het model in overeenstemming zijn. Het ligt dus voor de hand een of andere *goodness-of-fit*-maat te gebruiken die rekening houdt met de verhouding van het aantal foute observaties ten opzichte van het totale aantal observaties. Zo'n maat is Loevingers (1948) *H*, die gedefinieerd is als

$$H = 1 - \frac{\sum_{j=1}^{m-1} \sum_{k=j+1}^{m} f_{jk}}{\sum_{j=1}^{m-1} \sum_{k=j+1}^{m} e_{jk}} \qquad [10.3]$$

In deze formule is f_{jk} het aantal *geobserveerde fouten* voor de keuzen van Object j en Object k, waarbij j voor k komt op de Guttmanschaal. De waarde e_{jk} is het aantal *verwachte fouten* als we ervan uitgaan dat de keuzen voor beide objecten onafhankelijk van elkaar zijn. Als Object j gemiddeld populairder is dan Object k, dan is f_{jk} gelijk aan het aantal personen dat Object j niet kiest, maar Object k wel. De verwachte frequentie e_{jk} is gelijk aan

$$e_{jk} = \frac{n_k (n - n_j)}{n} \qquad [10.4]$$

In deze formule is n_j het aantal personen dat Object j gekozen heeft; n_k is het aantal personen dat Object k koos en n is het totaal aantal proefpersonen. Bijvoorbeeld: uit de data van Tabel 10.5 kunnen we tien 2×2-tabelletjes construeren die aangeven hoe vaak een bepaald antwoord met betrekking tot de ene bijdrage samengaat met de antwoorden voor de andere bijdragen. Deze tabelletjes zijn weergegeven in Tabel 10.6.

Tabel 10.6 Twee-bij-twee kruistabellen van de antwoorden met betrekking tot schrijven (S), bedenken (B), leidinggeven (L), data-analyseren (A) en dataverzamelen (V)

	B		L		A		V	
	ja	nee	ja	nee	ja	nee	ja	nee
S								
ja	30	8	26	12	25	13	11	27
nee	0	0	0	0	0	0	0	0
B								
ja			22	8	20	10	9	21
nee			4	4	5	3	2	6
L								
ja					18	8	9	17
nee					7	5	2	10
A								
ja							11	14
nee							0	13

Als de onderliggende Guttmanschaal schrijven-bedenken-leiden-analyseren-verzamelen is, dan geven de (vetgedrukte) getallen die linksonder in de tabelletjes vermeld zijn, de aantallen foute antwoorden ten opzichte van de Guttmanschaal weer. We zien dat er geen fouten geobserveerd worden voor

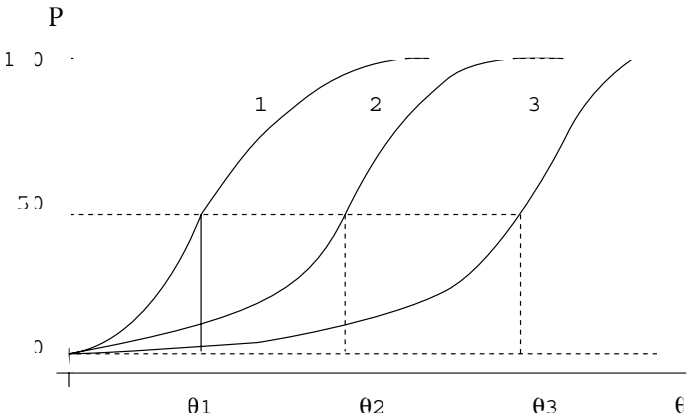
alle objectparen die het object 'schrijven' bevatten. Evenmin zijn er fouten bij het paar 'data-analyse' en 'dataverzameling'. Wel zien we fouten in de paren van bedenken en leidinggeven (4), bedenken en data-analyse (5), bedenken en dataverzameling (2), leidinggeven en data-analyse (7), en leidinggeven en dataverzameling (2). De verwachte aantallen fouten zijn voor de opeenvolgende objectparen respectievelijk 0, 0, 0, 0, $(8 \times 26)/38 = 5.47$, $(8 \times 25)/38 = 5.26$, $(8 \times 11)/38 = 2.32$, $(12 \times 25)/38 = 7.89$, $(12 \times 11)/38 = 3.47$, en $(13 \times 11)/38 = 3.76$. Voor de Guttmanschaal schrijven-bedenken-leiden-analyseren-verzamelen is de *schaalbaarheidscoëfficiënt* dus gelijk aan

$$H = 1 - \frac{4 + 5 + 2 + 7 + 2}{5.47 + 5.26 + 2.32 + 7.89 + 3.47 + 3.76} = 1 - \frac{20}{28.17} = 1 - .71 = .29,$$

wat tamelijk laag is. Meestal wordt .30 als grens genomen voor wat nog als (zwakke) schaal acceptabel is.
Door Formule 10.1 te gebruiken, waarin rekening gehouden wordt met het aantal *verwachte* fouten, gaan we er impliciet van uit dat er ook bij een goed passend model foute antwoorden voor kunnen komen en dat we de waarschijnlijkheid dat zulke fouten optreden, kunnen schatten. We hebben daarmee een stap gemaakt naar een *probabilistische* versie van de Guttmanschaal. Zo'n schaal wordt een Mokkenschaal genoemd.

De Mokkenschaal

Het deterministische model dat aan de Guttmanschaal ten grondslag ligt, is niet erg realistisch omdat het impliceert dat er geen fouten gemaakt mogen worden. Het lijkt verstandiger om een model te postuleren waarin er wel fouten voor mogen komen. Zo'n model is in 1971 ontwikkeld door Mokken (zie ook: Mokken & Lewis, 1982). Ook daarin is er sprake van een latent continuüm θ waarop zowel personen als objecten posities innemen. De itemkarakteristieke curve (IKC) van een object heeft nu echter niet de vorm van een sprongfunctie, maar is een monotoon stijgende functie die alle waarden tussen 0 en 1 in kan nemen. De schaalwaarde van een object is nu gedefinieerd als de waarde van θ waarvoor geldt dat $P_{j/\theta j} = .50$. Een voorbeeld met de IKC's van drie objecten is afgebeeld in Figuur 10.7. In dit voorbeeld snijden de IKC's van de drie objecten elkaar niet, ook al zou dit in principe wel kunnen. Een schaal met objecten waarvan de IKC's elkaar niet snijden heeft de eigenschap van *dubbele monotonie*. In het model van Mokken wordt ervan uitgegaan dat de gegevens aan deze eigenschap voldoen. Zo'n schaal wordt een Mokkenschaal genoemd.



Figuur 10.7 Item-karakteristieke curve voor de keuze van een object volgens het Mokkenschaal-model

In het Mokkenmodel wordt verder uitgegaan van *lokale stochastische onafhankelijkheid*, dat wil zeggen dat de kans op een keuze van zowel Object *j* als Object *k* door iemand met een bepaalde waarde θ_i gelijk is aan $P_{ij} \times P_{ik} = P_{j/\theta i} \times P_{k/\theta i}$. Op analoge wijze zijn de kansen uit te drukken voor de keuze van wel Object *k* maar niet Object *j* en van wel Object *j* maar niet Object *k*. Aangezien deze kansen doorgaans niet gelijk aan nul zijn, laat het model dus toe dat er in de empirie 'foute' antwoordpatronen gevonden zullen worden. Het model van Mokken, dat de probabilistische versie van de Guttmanschaal is, kunnen we ook beschouwen als het niet-parametrische equivalent van het Rasch-model uit de moderne test-theorie (zie Hambleton e.a., 1991; Van der Linden & Hambleton, 1997).
Om gegevens te analyseren volgens het Mokken-model, kunnen we gebruikmaken van het computerprogramma MSP van Debets en Brouwer (1989). Met dit programma zijn de data van Tabel 10.5 geanalyseerd nadat ze weer teruggecodeerd waren in enen (ja-antwoorden) en nullen (nee-antwoorden). Het belangrijkste resultaat is dat de schaal schrijven-bedenken-leiden-analyseren-verzamelen een onacceptabel lage *fit* heeft ($H = .29$). MSP zoekt naar schalen die hogere *H*-waarden hebben door achtereenvolgens een of meer items weg te laten. Ook de schaal schrijven-bedenken-analyse-verzamelen heeft een lage *fit* ($H = .38$). De beste schaal heeft slechts drie items: schrijven-analyse-verzamelen. Deze schaal is perfect ($H = 1.00$); in de data van Tabel 10.5 komen geen foute antwoordpatronen op deze drie items voor.

Wat nu?
In het onderzoek van Van der Kloot en Willemsen (1991) ging het om het vinden van schaalwaarden voor vijf verschillende bijdragen aan een wetenschappelijk onderzoek. Als we die schaalwaarden eenmaal gevonden hebben, dan is het mogelijk om alle onderzoekers, op grond van de door hen geleverde bijdragen, een totaalscore te geven teneinde hun positie in de auteursvolgorde te

bepalen. Daarvoor is het nodig om te weten wat de schaalwaarde van 'helemaal geen bijdrage' is, zodat we dit als absoluut nulpunt kunnen gebruiken. Door middel van enkele aanvullende analyses (zie Van der Kloot & Willemsen, 1991) kon worden afgeleid dat dit absolute nulpunt de waarde -4.14 op de Thurstoneschaal inneemt. Tellen we dus bij alle schaalwaarden het getal 4.14 op, dan krijgen we nieuwe schaalwaarden die uitdrukken wat de waarde van een bijdrage is, vergeleken met helemaal geen bijdrage. De nieuwe schaalwaarden worden dan 5.48, 4.32, 4.16, 3.80, 3.48 voor schrijven, bedenken, leidinggeven, data-analyse en dataverzameling. De auteurschapsgrens komt nu op 3.59 te liggen.

Stel nu dat een promovenda een onderzoek uitvoert onder leiding van de hoogleraar die het onderzoek bedacht heeft. De promovenda verzamelt de data, verricht de analyse en schrijft het artikel. Volgens de getransformeerde Thurstone-schaalwaarden krijgt de promovenda dus 5.48 + 3.80 + 3.48 = 12.76 punten. De hoogleraar krijgt dan 4.32 + 4.16 = 8.48 punten. Beiden komen ruimschoots boven de auteurschapsgrens uit, zij zullen dus beiden als auteurs van het artikel vermeld moeten worden. Daarbij zal de promovenda volgens deze telling de eerste auteur van het gezamenlijke artikel zijn.

Voor bovenstaande toepassing is het voldoende om over een stel schaalwaarden te beschikken die overeenkomen met de 'gemiddelde' ideeën over de belangrijkheid van de bijdragen. De Thurstone-schaalwaarden zijn zulke 'gemiddelde' waarden, evenals de overeenkomstige categoriekwantificaties op de eerste dimensie uit de HOMALS-analyse. In Hoofdstuk 1 hebben we nagegaan dat de Thurstone-schaalwaarden een goede weergave zijn van de oorspronkelijke data. In HOMALS kunnen we dat zien aan de eigenwaarde op de eerste dimensie (.314, en dus niet erg veel minder dan de .50 van het perfecte geval). Om deze schaalwaarden te kunnen gebruiken is het niet nodig dat de onderliggende oordelen een perfecte Guttmanschaal vormen.