HOOFDSTUK 13

Bijzondere toepassingen

13.1 Bijzondere observaties

Behalve in de psychologie worden meerdimensionale schaaltechnieken ook in tal van andere wetenschappen gebruikt. Voorbeelden zijn te vinden in de sociologie, de politicologie, de antropologie, de fysiologie, de aardrijkskunde, de literatuurwetenschappen, de wetenschapsstudies en de *marketing*. Daaruit blijkt dat MDS-methoden niet aan één bepaald wetenschapsgebied gebonden zijn, maar op allerlei soorten inhoudelijke data kunnen worden toegepast.
In dit hoofdstuk worden enkele bijzondere toepassingen van MDS-analyse besproken. Deze toepassingen zijn niet zozeer bijzonder vanwege het inhoudelijke terrein waarop ze plaatsvinden, maar omdat het om bijzondere observaties gaat of om observaties die doorgaans op andere manieren geanalyseerd worden. Met name zullen we kijken naar MDS van paarsgewijze preferentiedata, sorteergegevens, correlatiecoëfficiënten en beoordelingsschalen.

13.2 MDS van paarsgewijze preferentiedata

In Hoofdstuk 1 hebben we een vrij uitgebreide uiteenzetting gegeven van de analyse van paarsgewijze vergelijkingsdata (PGV-data) met behulp van Thurstones *Law of comparative judgment*. In de voorbeelden die daarbij behandeld werden, ging het om zogenaamde voorkeurs- of preferentiedata die verzameld worden doordat men proefpersonen alle mogelijke paren stimuli aanbiedt. Voor ieder paar stimuli wordt de proefpersonen gevraagd aan te geven welke stimulus zij in een bepaald opzicht prefereren. Door te tellen hoe vaak (dat wil zeggen door hoeveel proefpersonen) stimulus *i* boven stimulus *j* verkozen wordt, kunnen proporties p_{ij} verkregen worden die volgens Thurstones *law* een

functie zijn van de schaalwaarden u_i en u_j van deze stimuli. Voor deze functie geldt in ieder geval dat

1 $p_{ij} = .50$ als $u_i - u_j = 0$
2 $p_{ij} > .50$ als $u_i - u_j > 0$
3 $p_{ij} < .50$ als $u_i - u_j < 0$

waarbij de relatie tussen p_{ij} en $(u_i - u_j)$ *monotoon stijgend* moet zijn, met $p_{ij} = 1.0$ voor extreem positieve waarden van $(u_i - u_j)$ en $p_{ij} = 0.0$ voor extreem negatieve. Uit het bovenstaande kunnen we afleiden dat $|p_{ij} - .50|$ een monotone functie is van $|u_i - u_j|$ en dus ook dat $|u_i - u_j| \approx g(|p_{ji} - .50|)$.
Het absolute verschil $|u_i - u_j|$ is niets anders dan de afstand d_{ij} tussen de stimuli *i* en *j* op de ene dimensie waarop de stimuli met elkaar vergeleken moeten worden. De observaties $|p_{ij} - .50|$ zijn dus equivalent met afstanden die op *ordinaal niveau* gemeten zijn tussen de punten *i* en *j*. Het ligt dus voor de hand om deze observaties met behulp van niet-metrische MDS te analyseren. Het voordeel van deze door Davison en Wood (1983) voorgestelde aanpak is dat er veel minder strenge eisen aan de data gesteld worden: in plaats van de exact gedefinieerde cumulatieve normaalverdeling, die in Thurstones model met de monotone functie *g* correspondeert, voldoet in dit geval elke willekeurige monotoon stijgende functie aan de eisen van het model.
Passen we deze aanpak toe op de auteurschapsdata uit Tabel 1.2, dan berekenen we in eerste instantie de in Tabel 13.1 weergegeven matrix met waarden $|p_{ij} - .50|$.

Tabel 13.1 De matrix met de uit Tabel 1.2 afgeleide waarden |pij – .50|

	BE	LE	DV	DA	SC	AG
bedenken (BE)	–	.025	.288	.238	.375	.275
leiding geven (LEI)	.025	–	.275	.113	.413	.188
dataverzameling (DV)	.288	.275	–	.125	.438	.213
data-analyse (DA)	.238	.113	.125	–	.463	.138
schrijven (SC)	.375	.413	.438	.463	–	.488
auteurschap (AG)	.275	.188	.213	.138	.488	–

De data uit Tabel 13.1 kunnen we met behulp van ALSCAL analyseren. Mede gezien de resultaten uit Hoofdstuk 1 verwachten we dat deze gegevens op een eendimensionale schaal kunnen worden weergegeven. Toch geven we ALSCAL de opdracht om naast een eendimensionale ook een tweedimensionale oplossing te zoeken. Dat doen we om twee redenen. In de eerste plaats omdat het een bekend feit is dat eendimensionale MDS-analyses vaak in een lokaal minimum terechtkomen. Dat kan gebeuren als de stimuli in de initiële configuratie in de verkeerde volgorde op de dimensie geplaatst zijn. Als bijvoorbeeld de beginconfiguratie A-C-B-D is in plaats van de correcte volgorde A-B-C-D,

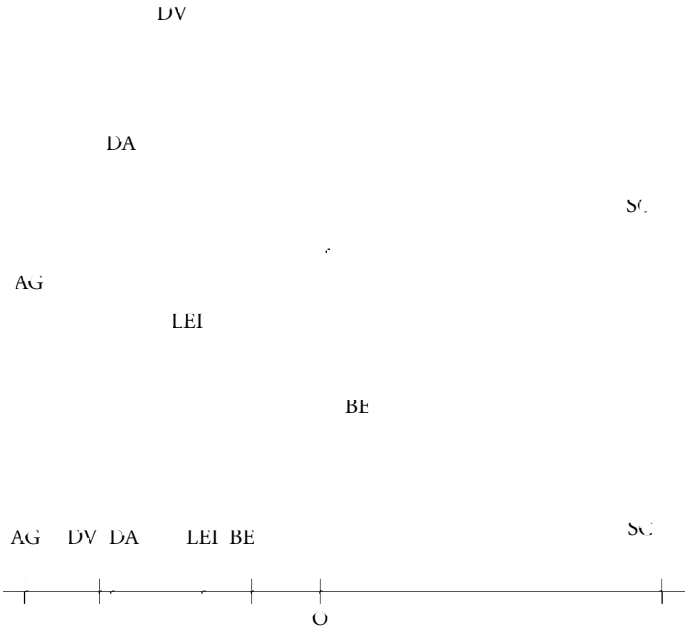
zal het MDS-programma in de opeenvolgende iteraties B meer in de richting van A, en C meer in de richting van D willen verplaatsen. Echter, daardoor komen C en B steeds dichter bij elkaar te liggen, waardoor het kan gebeuren dat de overeenstemming tussen de rangorde van de afstanden en de rangorde van de observaties steeds kleiner wordt. Grotere verplaatsingen in genoemde richting leveren dan geen stressverlaging op, zodat het algoritme stopt. Punt B kan als het ware niet voorbij punt C komen, en omgekeerd. De stressfunctie zit nu in een lokaal minimum. Pas als B en C elkaar kunnen passeren, bereikt het algoritme het globale minimum. En dat gaat gemakkelijker als ALSCAL met een tweedimensionale oplossing kan beginnen.

De tweede reden om een tweedimensionale oplossing zoeken, is dat we ervan uit moeten gaan dat de data *error*, dat wil zeggen, inconsistenties zullen bevatten. Data met inconsistenties kunnen niet goed in een laag-dimensionale ruimte worden weergegeven, en zeker niet in één dimensie. Door een meerdimensionale oplossing te zoeken, verwachten we dat de error zich verspreidt over de tweede en eventueel derde en vierde dimensies. De coördinaten op de eerste dimensie van de meerdimensionale oplossingen zouden dan betere schattingen zijn van de gezochte schaalwaarden.

De opdracht `ALSCAL VARIABLES BE LEI DV DA SC AU /LEVEL=ORDINAL /CRITERIA DIMENSION(1,2) CONVERGE(.0001) ITER(100) STRESSMIN(.001)` geeft een oplossing in twee en één dimensies. De bijbehorende $Stress_1$-waarden zijn .0045 en .053, zodat zelfs de eendimensionale oplossing acceptabel is[1]. De tweedimensionale configuratie is in Figuur 13.1 weergegeven, samen met de afbeelding van de eendimensionale oplossing en de schaalwaarden van de Thurstone-oplossing. Wat we aan deze figuur kunnen zien is dat de schaalwaarden op de eerste dimensie wel veel lijken op de schaalwaarden uit de Thurstone-analyse (de correlatie is .972) maar dat ze er niet perfect mee overeenkomen. Er zijn twee belangrijke verschillen: zowel de schaalwaarde van de auteurschapsgrens als die van data-analyse liggen in deze oplossing onder die van dataverzameling, terwijl ze daar in de Thurstone-oplossing boven lagen. De ALSCAL-oplossing suggereert dus dat alle bijdragen aan een wetenschappelijk onderzoek voor

1 Een oplossing in drie dimensies heeft een stress van .0008. Bij een oplossing voor zes stimuli in twee dimensies moet ALSCAL twaalf parameters berekenen uit $6(5 \times 1)/2 = 15$ gegevens. Dat gaat nog net, al waarschuwt ALSCAL wel dat er gezien het aantal observaties eigenlijk te veel parameters berekend worden. Bij een driedimensionale oplossing moeten er achttien parameters berekend worden. Dat zijn er meer dan het aantal observaties, waardoor ALSCAL deze opdracht niet uitvoert. Via een trucje kan men wel een driedimensionale oplossing krijgen, namelijk door de *complete* matrix in te voeren (dat wil zeggen: alle waarden zowel boven, onder als op de diagonaal van de matrix) en vervolgens de opdracht te geven `ALSCAL VARIABLES BE TO AU /LEVEL=ORDINAL /SHAPE=ASYMMETRIC /MODEL=EUCLID /CRITERIA DIMENSION (1,3)`. Daardoor 'denkt' ALSCAL dat er dertig observaties buiten de diagonaal beschikbaar zijn. NB: een vierdimensionale oplossing heeft een stress van nul, omdat we *m* punten ordinaal altijd perfect in een ruimte met $m - 2$ dimensies kunnen afbeelden.

auteurschap in aanmerking komen en dat dataverzameling (nog net iets) belangrijker is dan data-analyse.



Figuur 13.1 Twee- en eendimensionale ALSCAL-oplossingen van de auteurschapsdata uit Tabel 13.1

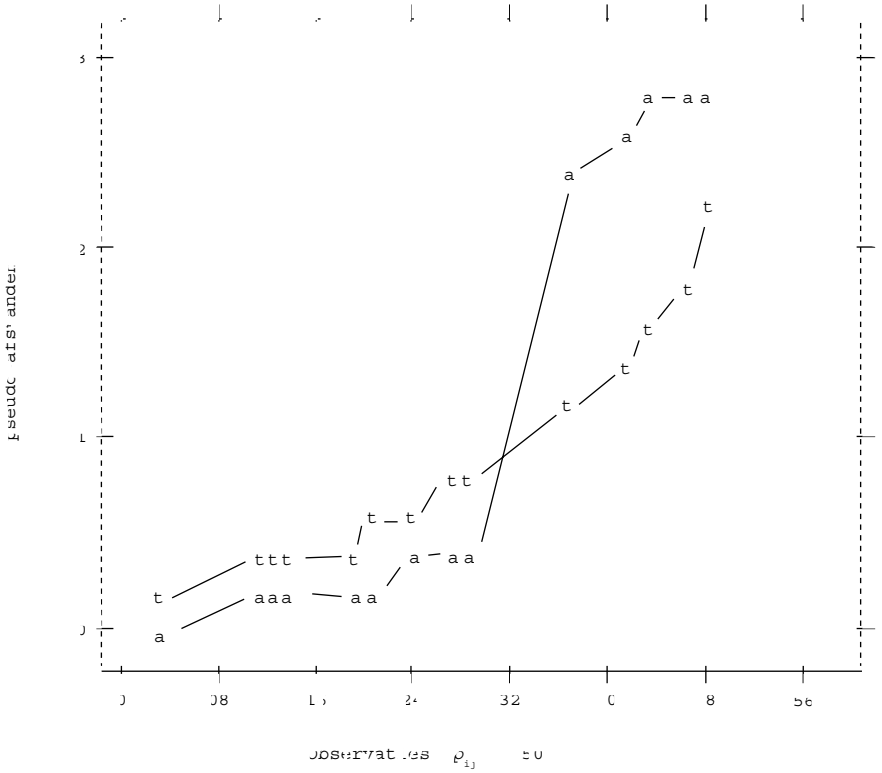
Of bovenstaande conclusie juist is, hangt onder andere af van de vraag of de ALSCAL-oplossing wel optimaal is, dat wil zeggen, niet in een lokaal minimum terecht is gekomen terwijl er misschien een globaal minimum bestaat met coördinaten die meer op de Thurstone-oplossing lijken. Om dat na te gaan moeten we de ALSCAL-analyse laten beginnen met de Thurstone-schaalwaarden als startconfiguratie. De aansturing ziet er dan als volgt uit:

```
data list table/ dim1 1-5 dim2 8 type_ 10-15 (A).
begin data.
  .18  0 config
  .02  0 config
 -.66  0 config
 -.34  0 config
 1.34  0 config
 -.55  0 config
end data.
save outfile='thurston.sys'.
data list free /label (a) BE LEI DV DA SC AU .
```

```
begin data.
[hier de nabijheidsdata uit Tabel 13.1]
end data.
alscal variables be to au
 /file='thurston.sys' config (initial)
/level=ordinal
 /shape=symmetric
 /model=euclid
 /print data
 /plot
 /criteria dimension(1,2) iter(100) converge(.0001) stressmin(.001).
```

Deze analyse levert $Stress_1$-waarden op van .106 voor de beginconfiguratie (deze waarde wordt in ALSCAL berekend door `CONFIG (FIXED)` op te geven), gevolgd door .016 voor de twee- en .062 voor de eendimensionale oplossing. Het ziet er dus juist naar uit dat deze oplossing in een lokaal minimum is terechtgekomen.

Uit bovengenoemde resultaten moeten we constateren dat er voor de niet-metrische MDS van de paarsgewijze vergelijkingen van Tabel 13.1 een 'betere' oplossing (dat wil zeggen een oplossing met minder stress) bestaat dan de Thurstone-oplossing. Dergelijke resultaten zijn ook gevonden door Davison en Wood (1983). Hun verklaring is als volgt: de rangorde van (sommige) waarden uit de datamatrix zijn inconsistent met de rangorde die ze zouden moeten hebben als de punten perfect op een eendimensionale schaal zouden liggen (dat is na te gaan aan de ruwe data, maar blijkt natuurlijk ook al uit het feit dat de stress groter dan nul is). ALSCAL lost deze inconsistenties op een andere manier op dan de Thurstonemethode, waardoor er niet alleen andere schaalwaarden gevonden worden, maar ook een andere transformatiefunctie. Figuur 13.2 toont de door ALSCAL gevonden transformatiefunctie, naast de functie die in de Thurstonemethode gebruikt wordt. We zien dat ALSCAL de oorspronkelijke observaties zodanig transformeert dat er min of meer twee groepen ontstaan. Zo'n soort transformatie is uiteraard erg steekproefgevoelig. Zouden we dit onderzoek herhalen, dan zou ALSCAL waarschijnlijk niet alleen andere schaalwaarden opleveren, maar ook een andere transformatiefunctie gebruiken. De Thurstone-transformatiefunctie is per definitie steeds dezelfde. Daarom geven we bij deze specifieke toepassing de voorkeur aan de Thurstone-oplossing. Bovendien hebben we bij de singuliere-waardendecompositie van de preferentierangordeningen die aan deze geaggregeerde data ten grondslag liggen (zie Hoofdstuk 11) dezelfde resultaten gevonden als met de Thurstonemethode.



Figuur 13.2 Afbeelding van de transformatiefuncties van de eendimensionale ALSCAL-oplossing (a) en van het Thurstonemodel (t). Verticaal staan de pseudo-afstanden, horizontaal de geobserveerde dissimilarities

13.3 DE ANALYSE VAN SORTEERGEGEVENS

Een van de handigste methoden om gelijkenisgegevens over een groot aantal stimuli te verzamelen is de zogenaamde sorteermethode. Voor deze methode, die in het Engels *method of free sorting* (ook: *method of unconstrained sorting*) genoemd wordt, heeft men slechts twee dingen nodig: een pakje systeemkaartjes en een grote tafel. Op de systeemkaartjes worden de namen of de afbeeldingen van de stimuli weergegeven. De kaartjes worden geschud als een pak speelkaarten en aan een proefpersoon overhandigd, met het verzoek ze te sorteren door de kaartjes op stapeltjes te leggen. Kaartjes met stimuli die bij elkaar horen of op elkaar lijken moeten op één stapeltje gelegd worden, kaartjes met stimuli die verschillend zijn moeten in verschillende stapeltjes terechtkomen. Hoe vaker (dat wil zeggen: hoe groter het aantal proefpersonen door wie) twee kaartjes bij elkaar op één stapeltje worden gelegd, des te groter de gelijkenis tussen de stimuli die op die kaartjes staan. Deze gelijkenissen kunnen vervol-

gens met behulp van een MDS-programma en/of door middel van clusteranalyse geanalyseerd worden[2].

Instructie

Een voorbeeld van de manier waarop men een proefpersoon deze taak kan aanbieden is de op Coxon en Jones (1979, p. 172) gebaseerde instructie die door Verkes, Van der Kloot en Van der Meij (1989) gebruikt is in een sorteeronderzoek van 176 woorden voor het beschrijven van pijn.

> *Voor u ligt een pakje kaartjes met woorden die gebruikt kunnen worden om allerlei soorten pijn te beschrijven. Op ieder kaartje staat één woord. Het is de bedoeling dat u aangeeft welke soorten pijn volgens u op elkaar lijken. Dat doet u door de kaartjes op stapeltjes te leggen. U doet dat zodanig dat de woorden voor soorten pijn die volgens u erg op elkaar lijken in hetzelfde stapeltje terechtkomen. Woorden voor soorten pijn die niet op elkaar lijken, legt u in verschillende stapeltjes neer.*
>
> *U mag net zoveel stapeltjes maken als u zelf wilt en in elk stapeltje mag u net zoveel kaartjes leggen als u nodig vindt. Als u dat wilt, mag u kaartjes van het ene stapeltje naar het andere stapeltje verplaatsen.*
> *Als er woorden bij zitten die volgens u absoluut niet gebruikt kunnen worden om pijn aan te duiden, houdt u die dan apart. Houdt u ook de woorden apart waarvan u niet precies weet wat ze betekenen.*
>
> *Wilt u nog iets vragen? Anders kunt u nu beginnen met stapeltjes maken.*

In deze instructie wordt de proefpersoon gevraagd stapeltjes te maken van woorden die qua betekenis *op elkaar lijken*. In andere gevallen kan men vragen de stimuli te sorteren in groepjes die *bij elkaar horen* of *met elkaar samengaan* (bijvoorbeeld: persoonlijkheidskenmerken van mensen).

Registratie van sorteergegevens

Als een proefpersoon klaar is met de sorteertaak, liggen er twee of (veel) meer groepjes kaartjes op tafel. De beste manier om deze sorteringen te registreren is door elk groepje een nummer te geven en op te schrijven welke stimuli in welk groepje zitten. Stimuli die alleen zijn gebleven (dus in groepjes van één zitten) krijgen ook een apart groepsnummer; het is handig om daar de hoogste nummers voor te gebruiken. Vervolgens is het handig om de groepsnummers te noteren in een matrix van stimuli × proefpersonen. Bijvoorbeeld: stel dat vijf

2 Voorzover kon worden nagegaan, is Miller (1967, 1969) de eerste onderzoeker geweest die sorteerdata in gelijkenissen heeft omgezet om ze met clusteranalyse te analyseren. Rosenberg, Nelson en Vivekananthan (1968) waren waarschijnlijk de eersten die de uit sorteerdata afgeleide gelijkenissen met MDS geanalyseerd hebben.

proefpersonen de vijf pijnwoorden *gloeiend* (G), *flitsend* (F), *bonkend* (B), *klemmend* (K) en *stekend* (S) als volgt hebben gesorteerd:
Persoon 1: {(G), (F, B), (K, S)}
Persoon 2: {(G, F, B), (K, S)}
Persoon 3: {(G, S), (F, B, K)}
Persoon 4: {(G, K), (F, B), (S)}
Persoon 5: {(G, F), (B, K), (S)}.

Bovenstaande sorteringen kunnen genoteerd worden in de matrix die in Tabel 13.2 is weergegeven.

Tabel 13.2 Matrix van stimuli × personen met groepsnummers van de sorteringen

	Persoon				
Stimulus	1	2	3	4	5
gloeiend	1	1	1	1	1
flitsend	2	1	2	2	1
bonkend	2	1	2	2	2
klemmend	3	2	2	1	2
stekend	3	2	1	3	3

Op basis van de datamatrix van Tabel 13.1 kunnen verschillende andere matrices geconstrueerd worden. In de eerste plaats kunnen we voor elke persoon een stimulus × stimulus matrix met enen en nullen maken. Een één in een bepaalde cel van die matrix wil zeggen dat de desbetreffende rijstimulus samen met de kolomstimulus in één groepje gesorteerd is. Een nul betekent dat de rij- en kolomstimulus niet op hetzelfde stapeltje gelegd zijn. Zo'n matrix is dus een primitieve gelijkenis- of nabijheidsmatrix. Primitief, omdat de nabijheidsscores slechts de waarden één (nabij) of nul (niet nabij) aan kunnen nemen. De vijf matrices van de personen uit ons voorbeeld staan in Tabel 13.3.

Gelijkenismaten

Als we de waarde in cel (i,j) van Persoon k met het symbool o_{ijk} aanduiden dan kunnen we de *dissimilarity* s_{ijk} definiëren als $s_{ijk} = 1 - o_{ijk}$. De variabele S is een afstandsfunctie, een metriek (zie Hoofdstuk 4), omdat

- $s_{ijk} \geq 0$ (eis van niet-negativiteit),
- $s_{ijk} = 0$ als en slechts dan als i en j bij Persoon k samenvallen, dat wil zeggen, in één groepje gesorteerd zijn (eis van niet-gedegenereerdheid),
- $s_{ijk} = s_{jik}$ (eis van symmetrie), en
- $s_{ijk} \leq s_{ihk} + s_{hjk}$ (eis van driehoeksongelijkheid).

Tabel 13.3 Vijf gelijkenismatrices met enen en nullen voor de sorteringen van vijf stimuli door vijf personen

	Persoon 1: {(G), (F, B), (K, S)}				
	G	F	B	K	S
G	1	0	0	0	0
F	0	1	1	0	0
B	0	1	1	0	0
K	0	0	0	1	1
S	0	0	0	1	1
	Persoon 2: {(G, F, B), (K, S)}				
	G	F	B	K	S
G	1	1	1	0	0
F	1	1	1	0	0
B	1	1	1	0	0
K	0	0	0	1	1
S	0	0	0	1	1
	Persoon 3 = {(G, S), (F, B, K)}				
	G	F	B	K	S
G	1	0	0	0	1
F	0	1	1	1	0
B	0	1	1	1	0
K	0	1	1	1	0
S	1	0	0	0	1
	Persoon 4 = {(G, K), (F, B), (S)}				
	G	F	B	K	S
G	1	0	0	1	0
F	0	1	1	0	0
B	0	1	1	0	0
K	1	0	0	1	0
S	0	0	0	0	1
	Persoon 5 = {(G, F), (B, K), (S)}:				
	G	F	B	K	S
G	1	1	0	0	0
F	1	1	0	0	0
B	0	0	1	1	0
K	0	0	1	1	0
S	0	0	0	0	1

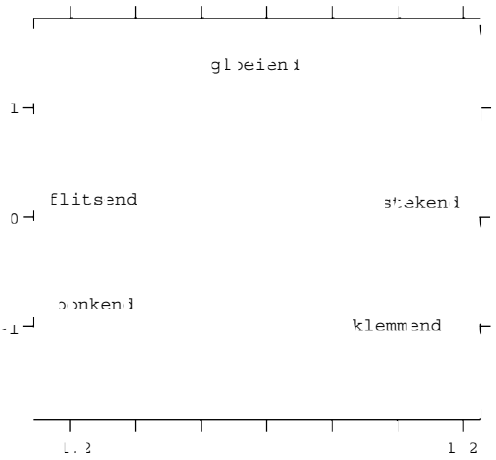
Het zal duidelijk zijn dat we met de individuele matrix met *S*-waarden van één persoon weinig kunnen doen, ook al zijn het in principe afstandachtige getallen. Om deze data te kunnen analyseren worden ze meestal geaggregeerd (opgeteld) over de personen volgens

$$\delta_{ij} = \sum_{k=1}^{n} s_{ijk} \qquad [13.1]$$
$$= \sum_{k=1}^{n} (1 - o_{ijk})$$
$$= n - \sum_{k=1}^{n} o_{ijk}.$$

De grootheid δ is ook weer een metriek omdat het de som is van *n* afzonderlijke metrieken. Het resultaat van deze bewerking staat in Tabel 13.4. De met ALSCAL verkregen configuratie van de pijnwoorden is afgebeeld in Figuur 13.3. Deze oplossing heeft met de discrete behandeling van *ties* een $Stress_1 = .004$.

Tabel 13.4 Matrix met geaggregeerde ongelijkheidsmaten δij

	G	F	B	K	S
Gloeiend	0	3	4	4	4
Flitsend	3	0	1	4	5
Bonkend	4	1	0	3	5
Klemmend	4	4	3	0	3
Stekend	4	5	5	3	0



Figuur 13.3 MDS-oplossing van de sorteringen van vijf pijnwoorden

BLOK 13.1 SPSS PROXIMITY

In een klein voorbeeld als het bovenstaande kunnen de matrices met enen en nullen gemakkelijk met de hand gemaakt worden. Zijn er meer stimuli en/of meer personen, dan is het efficiënter om de computer te gebruiken. Lezen we in SPSS de datamatrix van Tabel 13.2 in, dan heeft de SPSS-file vijf *cases* (de stimuli) en vijf variabelen (de personen), die we P1, P2, P3, P4 en P5 zullen noemen. Om uiteindelijk de matrix met δ_{ij}-waarden te kunnen berekenen moeten we een SPSS-programma maken met de volgende stappen:

```
COMPUTE C11=0.
COMPUTE C12=0.
COMPUTE C13=0.
COMPUTE C21=0.
COMPUTE C22=0.
COMPUTE C31=0.
COMPUTE C32=0.
COMPUTE C41=0.
COMPUTE C42=0.
COMPUTE C43=0.
COMPUTE C51=0.
COMPUTE C52=0.
COMPUTE C53=0.
IF (P1=1) C11=1.
IF (P1=2) C12=1.
IF (P1=3) C13=1.
IF (P2=1) C21=1.
IF (P2=2) C22=1.
IF (P3=1) C31=1.
IF (P3=2) C32=1.
IF (P4=1) C41=1.
IF (P4=2) C42=1.
IF (P4=3) C43=1.
IF (P5=1) C51=1.
IF (P5=2) C52=1.
IF (P5=3) C53=1.
```

Met bovenstaande opdrachten creëren we een nieuwe matrix die we **G** (met elementen g_{ic}) zullen noemen. Deze matrix heeft vijf rijen (m, het totale aantal stimuli) en 3 + 2 + 2 + 3 + 3 = 13 kolommen, evenveel

kolommen als het totale aantal stapeltjes dat de proefpersonen gevormd hebben toen ze de stimuli moesten sorteren. Als proefpersoon k in totaal t_k stapeltjes gemaakt heeft, dan is het totaal aantal kolommen van **G** dus gelijk aan $T = \sum_{k=1}^{n} t_k$. Deze nieuwe matrix is weergegeven in Tabel 13.5. De matrix **G** wordt een *indicatormatrix* genoemd. Een 1 in cel g_{ic} van deze matrix geeft aan dat de stimulus van rij i in het stapeltje van kolom c terecht is gekomen.

Tabel 13.5 De indicatormatrix **G** *voor de sorteringen van vijf pijnwoorden door vijf proefpersonen*

stimulus	C11	C12	C13	C21	C22	C31	C32	C41	C42	C43	C51	C52	C53
Gloeiend	1	0	0	1	0	1	0	1	0	0	1	0	0
Flitsend	0	1	0	1	0	0	1	0	1	0	1	0	0
Bonkend	0	1	0	1	0	0	1	0	1	0	0	1	0
Klemmend	0	0	1	0	1	0	1	1	0	0	0	1	0
Stekend	0	0	1	0	1	1	0	0	0	1	0	0	1

Met behulp van de SPSS-opdracht `PROXIMITIES C11 TO C53 /VIEW=CASE /MEASURE=BLOCK /MATRIX OUT(DELTA).` kunnen we de matrix met δ -waarden berekenen. Door middel van dit commando wordt de city-block afstand ($d^{(1)}$)van elk tweetal stimuli berekend. Dus: voor gloeiend (G) en flitsend (F) is $d^{(1)}_{GF} = 1 + |-1| + 0 + 0 + 0 + 1 + |-1| + 1 + |-1| + 0 + 0 + 0 + 0 = 6 = 2\delta_{GF}$. De afstanden die op deze manier berekend worden zijn allemaal twee keer zo groot als de δ-waarden uit Tabel 13.4. In plaats van `/MEASURE=BLOCK` kunnen we ook `MEASURE=SEUCLID of MEASURE=BSEUCLID (1,0)` gebruiken. Ook dan is de uitkomst een matrix met dissimilarities die twee keer zo groot zijn als de δ's. Dat maakt voor de analyse natuurlijk niet uit!
De δ-waarden worden weggeschreven in de matrix DELTA met vijf rijen en vijf kolommen (VAR1 tot en met VAR5) die in ALSCAL kan worden aangeroepen via `ALSCAL VARIABLES=VAR1 VAR2 VAR3 VAR4 VAR5 /MATRIX IN(DELTA).` ALSCAL geeft dan een niet-metrische oplossing in twee dimensies[3]. In dit geval is dat een oplossing met $Stress_1 = .004$ (met discrete behandeling van *ties*).

De coördinaten van de vijf pijnwoorden zijn: gloeiend (08, 1.16), flitsend (-1.34, .04), bonkend (-1.25, -.51), klemmend (.54, -.91), en stekend (1.97,

3 In de WINDOWS-versie van SPSS kan het berekenen van Euclidische afstanden direct vanuit het ALSCAL-venster worden aangestuurd.

.21); deze zijn al eerder afgebeeld in Figuur 13.3.

Sorteringen van machtsstrategieën

Een ander voorbeeld van een MDS-analyse van gelijkenissen die op sorteerdata gebaseerd zijn, is een onderzoek van Van der Kloot en Van Herk (1991) waarin 25 proefpersonen sorteringen van 16 stimuli gaven. De stimuli waren strategieën die mensen ten opzichte van anderen kunnen toepassen om hun zin te krijgen. Voorbeelden zijn: iemand manipuleren, onderhandelen, bedreigen. Deze stimuli zijn ontleend aan Falbo (1977), die gevonden had dat ze in twee onafhankelijke dimensies varieerden (direct-indirect en rationeel-irrationeel) en op de omtrek van een cirkel lagen. De stimuli en de sorteringen zijn weergegeven in de stimuli × personenmatrix van Tabel 13.6. De gelijkenismatrix staat in Tabel 13.7 en de daarvan afgeleide MDS-configuratie is weergegeven in Figuur 13.4[4]. De *S-stress* van deze oplossing is .147, de $Stress_1$ is .131 en de *RSQ* is gelijk aan .897.

Tabel 13.6 Sorteringen van 16 machtsstrategieën door 25 personen[a]

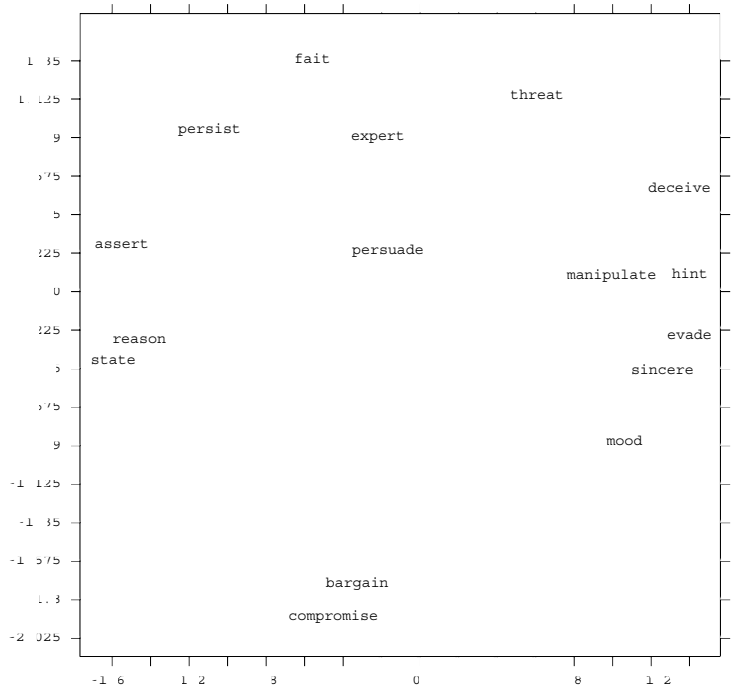
Stimuli	Personen
to manipulate	53 411 54 3 422 42 235 412 44 214 1
to hint	83 451 34 4 262 86 325 452 14 324 1
to put in a good mood	53444454413542252514622 42
to deceive	11152543122432355321232 11
to look sincere	13142344212542252524862 42
to evade	92152763222663262355252 16
to threaten	71356512115315311334241 45
to pose a fait accompli	45316512144311336341741 14
to assert	24533122354274231642313 34
to persist	64363222346171231432311 34
to state simply	24233131354254176242513 24
to claim expertise	11516241181632426434311 31
to reason	44515121331121126242513 24
to compromise	3 2 2 27634573 7 4614 3 2 1 3 1 2 3 2 3
to bargain	3 2 2 24433513 7 4714 3 2 1 3 1 2 3 2 3
to persuade	53415241431121215136413 41

[a] met de vetgedrukte getallen is na te gaan dat manipuleren en hints geven door elf personen in dezelfde categorie gesorteerd zijn en dat compromis en onderhandelen twintig keer op één stapeltje terecht kwamen.

4 Aan deze figuur is te zien dat de stimuli in dit onderzoek bepaald niet op een cirkel liggen. Toch is er maar weinig toename van de stress als we een oplossing zoeken waarin de stimuli op de omtrek van een cirkel *gedwongen* worden (Busing, Groenen & Van der Kloot, 1996).

Tabel 13.7 Uit de sorteerdata van Tabel 13.6 afgeleide ongelijkheidsmaten van 16 machtsstrategieën

manipulate	0														
hint	14	0													
mood	16	17	0												
deceive	14	16	22	0											
sincere	17	13	10	16	0										
evade	22	18	22	15	18	0									
threat	20	23	22	18	22	22	0								
fait accompli	21	24	25	19	25	23	13	0							
assert	25	24	24	25	24	24	23	20	0						
persist	24	24	24	25	24	24	20	18	11	0					
state	25	25	25	25	25	25	25	21	10	19	0				
expert	19	23	23	18	21	24	19	20	21	18	22	0			
reason	23	25	25	25	25	25	25	19	15	17	11	18	0		
compromise	24	24	20	25	23	24	25	25	24	25	19	25	21	0	
bargain	23	25	18	24	23	23	24	25	24	25	19	25	21	5	0
persuade	14	18	19	22	21	24	22	23	22	19	22	17	15	24	24	0



Figuur 13.4 MDS-configuratie van 16 machtsstrategieën in twee dimensies

Problemen met de δ-gelijkenismaten

Een probleem met de ongelijkheidsmaten s_{ijk} en δ_{ij} is dat zij geen rekening houden met de grootte van de groepen waarin de stimuli *i* en *j* gesorteerd zijn. In principe kan men op twee manieren redeneren:

1 De afstanden tussen stimuli in een groep waarin veel objecten gesorteerd zijn, zouden (gemiddeld) groter moeten zijn dan de afstanden tussen stimuli in een groep met weinig objecten. Dit is de situatie die men bij clusteranalyse aantreft. Kleine clusters bestaan meestal uit een paar punten die relatief dicht bij elkaar liggen. Grote clusters bevatten daarnaast ook een aantal punten die verder van elkaar af liggen.

2 De afstanden tussen de objecten in een kleine groep zouden groter moeten zijn dan de afstanden tussen objecten in een grote groep. Zoals Coxon en Jones (1979, p. 172) opmerken is dit '... unlikely in sorting behavior, but provides a useful limiting case...' voor de klasse van afstandsfuncties die men op sorteerdata zou kunnen toepassen.

Bovenstaande mogelijkheden zijn onderzocht door Burton (1975) en Coxon en Jones (1979). Hun onderzoek toonde aan dat het eerste alternatief (dus: afstanden binnen groepen zijn proportioneel met groepsgrootte) betere resultaten oplevert. De afstandsfunctie die zij voor dit geval gebruikten was ontleend aan eerder werk van Boorman en Arabie (1972).

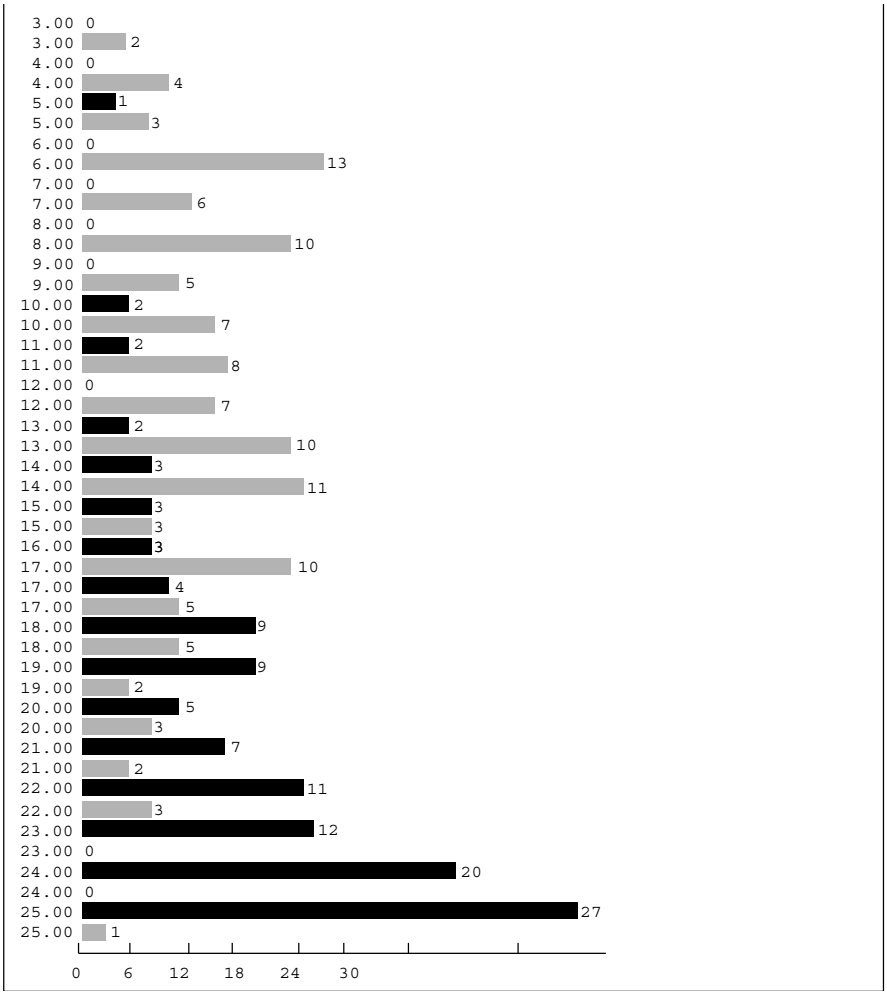
Een tweede probleem van de oorspronkelijke ongelijkheidsmaten $\{\delta_{ij}\}$ is dat de verdeling ervan zeer *scheef* is en heel veel *ties* bevat. Figuur 13.5 toont de verdeling van de δ-waarden van de machtsstrategieën, samen met de afstanden uit de uiteindelijke MDS-configuratie. Aan deze figuur is te zien dat de δ's een heel andere verdeling volgen dan Euclidische afstanden. Met name de kleinere δ-waarden komen heel weinig voor. Bijna veertig procent van de δ's hebben de waarde 24 of 25; er is dus een gebrek aan differentiatie tussen de grotere afstanden.

Het ligt nu voor de hand om de δ's zodanig te transformeren dat ze qua verdeling meer op echte afstanden gaan lijken. Een dergelijke transformatie is de door Rosenberg (zie bijvoorbeeld Rosenberg, Nelson & Vivekananthan, 1968) voorgestelde *indirecte afstandsmaat* die we hier met δ^* zullen aanduiden. Deze getransformeerde afstandsmaat is

$$\delta_{ij}^{*} = \sqrt{\sum_{h=1}^{m} (\delta_{ih} - \delta_{jh})^2} \qquad [13.2]$$

Deze transformatie berust op de volgende gedachtegang (zie Van der Kloot & Van Herk, 1991). Als proefpersonen sorteringen produceren, dan gaat er waardevolle informatie verloren met betrekking tot de subtielere verschillen tussen de onderlinge afstanden van stimuli. Dit geldt in het bijzonder voor stimulusparen die grote afstanden tot elkaar hebben. Bij MDS kan dit tot problemen leiden, omdat MDS-oplossingen voornamelijk gericht zijn op het goed representeren van grote afstanden. Volgens Rosenberg c.s. kan een deel van de verdwenen informatie herwonnen worden door bovengenoemde indirecte

afstanden te berekenen. De indirecte afstand tussen twee stimuli *i* en *j* bevat extra informatie over hun gelijkenis omdat δ^*_{ij} niet alleen op de directe afstand tussen deze twee stimuli gebaseerd is, maar ook op de overeenkomst tussen de afstand van *i* tot *h* en die van *j* tot *h* (en die van *i* tot *l* en *j* tot *l*, enzovoort). De verdeling van de δ^*-waarden is meer continu (dat wil zeggen δ^* neemt meer verschillende waarden aan) en is minder scheef. De δ^*-waarden zouden zich meer als echte afstanden gedragen en MDS-oplossingen met lagere stress opleveren. Onderzoek van Drasgow en Jones (1979) en van Van der Kloot en Van Herk (1991) toonde echter aan dat de δ^*-maten niet superieur waren aan de gewone δ's. Een recent onderzoek naar andere mogelijkheden om sorteerafstanden te transformeren is dat van Simmen (1996).



Figuur 13.5 Verdelingen van δ-waarden (■) en MDS-afstanden (■) die behoren bij de sorteringen van machtsstrategieën

BLOK 13.2 CORRESPONDENTIEANALYSE VAN G

Een van de bezwaren tegen Millers δ's is dat er bij deze ongelijkheidsmaat geen rekening wordt gehouden met de grootte van de categorieën waarin de stimuli gesorteerd zijn. Naast de afstandsmaten van Burton (1972) bestaan er ook andere manieren om categoriegrootte in de afstanden te betrekken. Dat kan door uit te gaan van de indicatormatrix **G**, met de stimuli als rijen en de categorieën als kolommen. Als voorbeeld nemen we de 5×13 indicatormatrix met betrekking tot de sorteringen van pijnwoorden, die in Tabel 13.4 is weergegeven. We zien dat elk rijtotaal gelijk is aan *n*, het aantal proefpersonen dat de stimuli gesorteerd heeft. De kolomtotalen zijn echter niet allemaal gelijk; elk kolomtotaal ($\sum_i g_{ic}$) geeft aan hoeveel stimuli in de bijbehorende categorie *c* gesorteerd zijn. Voor iedere cel (*i,c*) van de indicatormatrix kunnen we nu de *verwachte waarde* e_{ic} uitrekenen, onder aanname dat de rijen en kolommen onafhankelijk van elkaar zijn. In dat geval is

$$e_{ic} = \frac{\sum_{i=1}^{m} g_{ic}}{mxn} \times n = \frac{\sum_i g_{ic}}{m} \qquad [13.3]$$

Daarna kunnen we een matrix van *residuen* **R** berekenen, waarvan de elementen gelijk zijn aan

$$r_{ic} = \left(\frac{g_{ic} - e_{ic}}{\sqrt{e_{ic}}} \right) \left(\frac{1}{\sqrt{m \times n}} \right) \qquad [13.4)]$$

Deze matrix **R** is weergegeven in Tabel 13.8.

Tabel 13.8 De matrix **R***, behorend bij de sorteringen van vijf pijnwoorden door vijf personen*

	1.7885	-.6325	-.6325	.5164	-.6325	.9487	-.7746	.9487	-.6325	-.4472	.9487	-.6325	-.4472
	-.4472	.9487	-.6325	.5164	-.6325	-.6325	.5164	-.6325	.9487	-.4472	.9487	-.6325	-.4472
	-.4472	.9487	-.6325	.5164	-.6325	-.6325	.5164	-.6325	.9487	-.4472	-.6325	.9487	-.4472
(1/ (5x5)}x	-.4472	-.6325	.9487	-.7746	.9487	-.6325	.5164	.9487	-.6325	-.4472	-.6325	.9487	-.4472
	-.4472	-.6325	.9487	-.7746	.9487	.9487	-.7746	-.6325	-.6325	1.7889	-.6325	-.6325	1.7885

Bijvoorbeeld: $r_{12} = \{1/\sqrt{(5 \times 5)}\}\{(1 - .2)/\sqrt{.2}\} = \{1/\sqrt{(5 \times 5)}\}\{1.7885\}$.

Vervolgens gebruiken we de getallen in **R** om Euclidische afstanden tussen de stimuli te berekenen volgens onderstaande formule.

$$d(\chi^2)_{ij} = \sqrt{\sum_{c=1}^{T}(r_{ic} - r_{jc})^2} = \left(\frac{1}{\sqrt{m \times n}}\right) \sqrt{\sum_{c=1}^{T}\left(\frac{g_{ic} - e_{ic}}{\sqrt{e_{ic}}} - \frac{g_{jc} - e_{jc}}{\sqrt{e_{jc}}}\right)^2} \qquad [13.5]$$

Deze afstanden worden χ^2-afstanden genoemd (zie ook Hoofdstuk 11). Op deze χ^2-afstanden kunnen we een MDS-analyse toepassen. Gezien de aard van de bewerkingen die we met de data hebben uitgevoerd, ligt het voor de hand om een *metrische* analyse uit te voeren (in ALSCAL: /LEVEL=RATIO). De χ^2-afstanden tussen de vijf pijnwoorden en de coördinaten na een metrische analyse staan in Tabel 13.9.

Tabel 13.9 χ2-afstanden en coördinaten in de metrische mds-oplossing van vijf pijnwoorden

	Stimulus					Dimensie	
Stimulus	G	F	B	K	S	1	2
Gloeiend	0.0					-.08	1.58
Flitsend	4.082/5	0.0				-1.17	.10
Bonkend	4.655/5	2.236/5	0.0			-1.04	-.62
Klemmend	4.564/5	4.378/5	3.764/5	0.0		.48	-1.08
Stekend	5.164/5	5.323/5	5.323/5	4.378/5	0.0	1.81	.03

Het aardige is nu dat de hierboven beschreven stappen (het berekenen van **E**, **R**, de χ^2-afstanden en de metrische analyse) samen neerkomen op een *correspondentieanalyse* van de personen-bij-categorieën-matrix **G**. Het SPSS-programma ANACOR voert als het ware al deze stappen uit en geeft (op een normalisatie na) identieke coördinaten van de stimuli. Hieronder zullen we echter zien dat dezelfde resultaten nog eenvoudiger verkregen kunnen worden, namelijk door het toepassen van HOMALS op de tabel van stimuli bij personen.

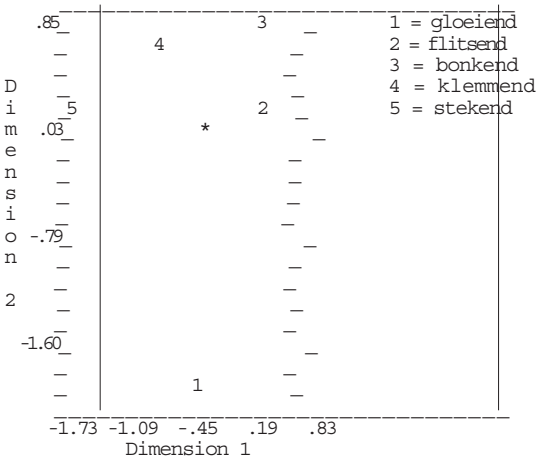
De analyse van sorteerdata door middel van HOMALS

Zoals we gezien hebben, is de eenvoudigste manier om sorteergegevens te registreren een tabel met rijen en kolommen die gevormd worden door respectievelijk de stimuli of objecten en de personen. In deze tabel staan getallen die aangeven in welke categorie een persoon een stimulus gesorteerd heeft. De *n* kolommen van zo'n tabel zijn dus op te vatten als evenzovele *nominale variabelen*. Immers: iedere kolom heeft een aantal waarden of categorieën die alleen maar aangeven dat objecten binnen de ene categorie niet hetzelfde zijn als objecten binnen een andere categorie. Verder is er over de categorieën van de variabelen niets bekend. Dergelijke data zijn exact het soort data waar HOMALS voor bedoeld is. Hieronder volgen de SPSS-opdrachten voor een HOMALS-analyse van de data van Tabel 13.2.

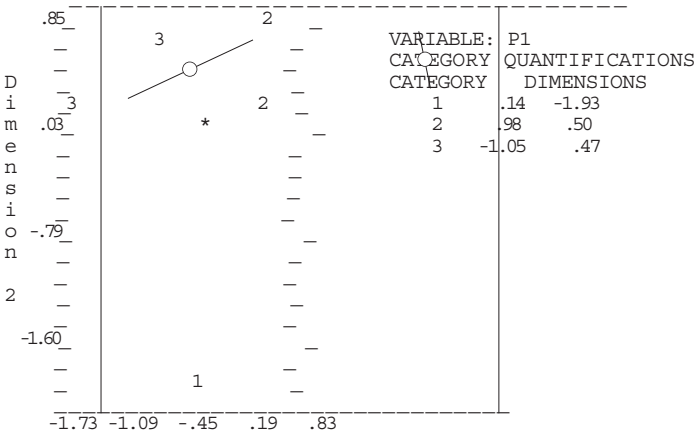
```
DATA LIST TABLE/OBJECT 1-10 (A) P1 TO P5 11-20.
BEGIN DATA.
gloeiend   1 1 1 1 1
flitsend   2 1 2 2 1
bonkend    2 1 2 2 2
klemmend   3 2 2 1 2
stekend    3 2 1 3 3
END DATA.
HOMALS VARIABLES P1 (3) P2 P3 (2) P4 P5 (3)
 /ANALYSIS P1 TO P5
 /NOBSERVATIONS=5
 /DIMENSION=2
 /PRINT DEFAULT OBJECT
 /PLOT DEFAULT OBJECT (P1 TO P5) DISCRIM
    QUANT (P1  TO P5).
```

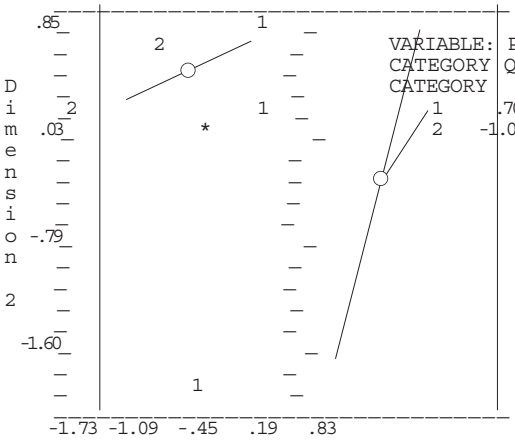
De belangrijkste uitvoer van dit programma bestaat natuurlijk uit de waarden en het plaatje van de *objectscores*. Deze volgen hieronder.

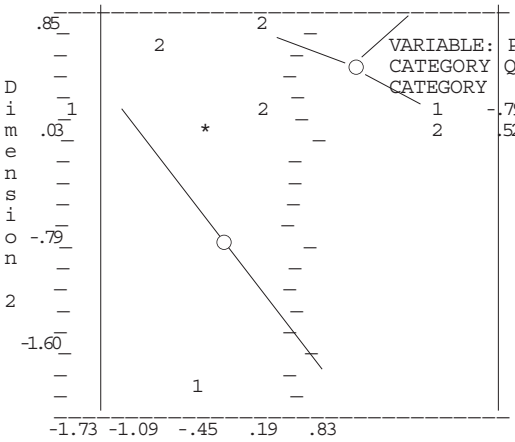
```
OBJECT *     DIMENSION
          1      2
   1 *     .14   -1.93
   2 *    1.02     .15
   3 *     .94     .85
   4 *    -.38     .68
   5 *   -1.71     .25
```

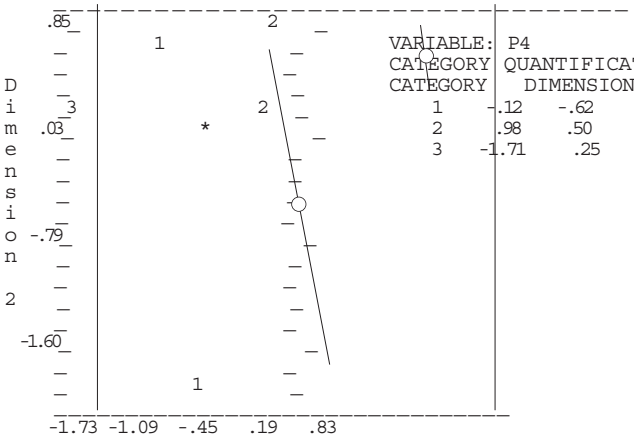


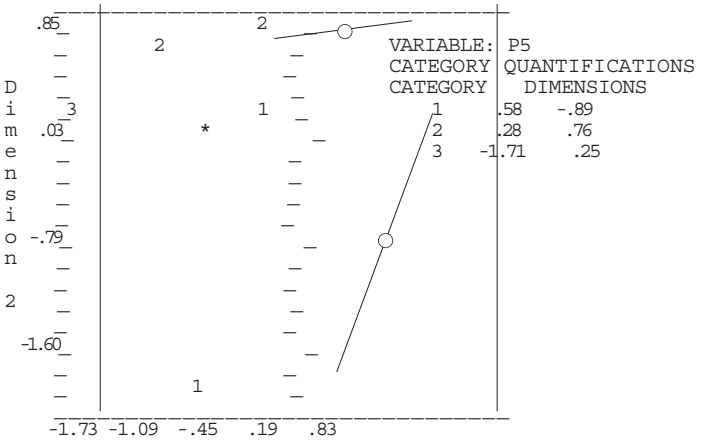
Bovenstaande figuur lijkt als twee druppels water op de ALSCAL-oplossing van de χ^2-afstanden uit Blok 13.2 en is identiek aan de ANACOR-oplossing (met genormaliseerde rijcoördinaten) van de personen × categorieën-matrix **G**. Hieronder volgen vijf figuren – voor elke proefpersoon één – waarin de objecten gecodeerd zijn met het nummer van de categorie waarin ze door de desbetreffende proefpersoon gesorteerd waren. Ook zijn de objecten van dezelfde categorie met elkaar verbonden. In dezelfde figuren zijn door middel van rondjes de zogenaamde *categoriekwantificaties* aangegeven, dat wil zeggen, de *centroïden* van de stimuli die samen in één categorie terecht waren gekomen.




VARIABLE: P2
CATEGORY QUANTIFICATIONS
CATEGORY DIMENSIONS
1 .70 -.31
2 -1.05 .47
Dimension 2
.85
.03
-.79
-1.60
-1.73 -1.09 -.45 .19 .83

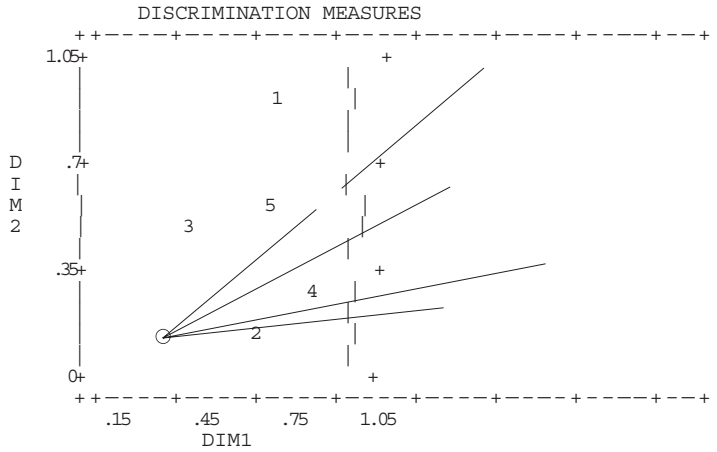

VARIABLE: P3
CATEGORY QUANTIFICATIONS
CATEGORY DIMENSIONS
1 -.79 -.84
2 .52 .56
Dimension 2
.85
.03
-.79
-1.60
-1.73 -1.09 -.45 .19 .83


VARIABLE: P4
CATEGORY QUANTIFICATIONS
CATEGORY DIMENSIONS
1 -.12 -.62
2 .98 .50
3 -1.71 .25
Dimension 2
.85
.03
-.79
-1.60
-1.73 -1.09 -.45 .19 .83



Andere interessante uitvoer van HOMALS bestaat uit de tabel en bijbehorende grafiek van de *discriminatiematen* van de proefpersonen. Deze maten zijn de tussen-categorieën-kwadratensommen per dimensie. Als zo'n discriminatiemaat groot is, dan betekent dat dat de categoriegemiddelden (de categoriekwantificaties) op de betreffende dimensie een grote spreiding hebben, in vergelijking met de spreiding van de objecten binnen de categorieën. In dat geval heeft een proefpersoon de dimensie dus 'gebruikt' om de stimuli te sorteren, dat wil zeggen: stimuli die op deze dimensie dicht bij elkaar liggen zijn op één stapeltje gelegd, stimuli die verder uiteen liggen zijn op verschillende groepjes gesorteerd.

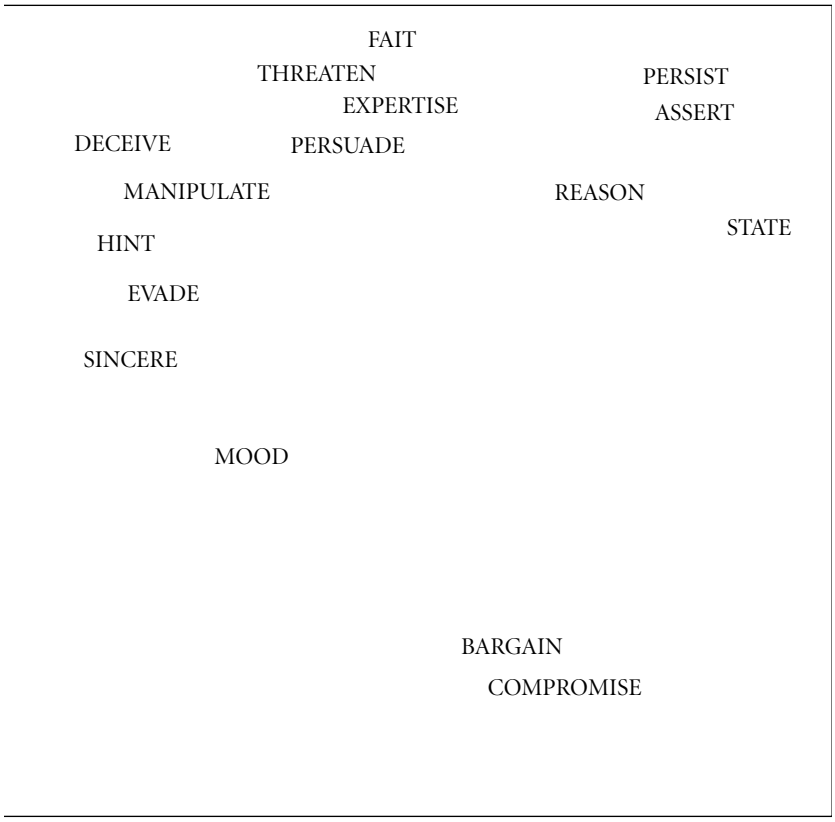
```
DISCRIMINATION MEASURES PER VARIABLE PER DIMENSION
==================================
VARIABLE     DIMENSION
          1     2
P1         .824   .933
P2         .729   .145
P3         .412   .472
P4         .972   .268
P5         .747   .563
- - - - - - - - - - - - - - - - - -
Eigenwaarde .737    .476
```



Aan de discriminatiematen kunnen we dus zien dat Persoon 1 beide dimensies gebruikt heeft bij het sorteren. Persoon 2 en Persoon 4 hebben voornamelijk de eerste dimensie gebruikt. De sorteringen van Persoon 3 zijn minder goed te voorspellen uit de posities van de stimuli op deze twee HOMALS-dimensies.

HOMALS-analyse van de machtsstrategieën

De in Tabel 13.6 weergegeven sorteringen zijn volgens de hierboven beschreven methode met HOMALS geanalyseerd. Het resultaat is de configuratie die in Figuur 13.6 is weergegeven. We zien dat onderhandelen en compromissen sluiten wat verder van de andere stimuli zijn komen af te liggen. Voor het overige wijkt deze figuur niet erg veel af van de eerder in Figuur 13.4 gepresenteerde ALSCAL-oplossing.



Figuur 13.6 HOMALS-oplossing van de sorteringen van machtsstrategieën

Voordelen van HOMALS voor de analyse van sorteergegevens

Aan de twee voorbeelden die hierboven behandeld zijn, hebben we kunnen zien dat HOMALS niet een totaal andere configuratie oplevert dan MDS op Millers δ-maten. Voorzover er verschillen optreden, zijn die toe te schrijven aan het feit dat HOMALS op een ingewikkelde manier rekening houdt met de grootte van de categorieën waarin de stimuli gesorteerd zijn. Het grote voordeel van HOMALS als methode voor de analyse van sorteergegevens is dus niet dat er een geheel andere configuratie ontstaat (dat zou immers juist een nadeel zijn!) maar is gelegen in drie andere eigenschappen:

1 HOMALS geeft met de discriminatiematen ook informatie over het sorteergedrag van de afzonderlijke proefpersonen (bij MDS van de δ's wordt er immers over proefpersonen geaggregeerd, waardoor de informatie over individuele verschillen verloren gaat).

2 HOMALS ondervindt geen onoverkomelijke problemen van zogenaamde *missing data* (hier: stimuli waar een proefpersoon geen raad mee wist en daardoor in afzonderlijke, uit één object bestaande categorieën gesorteerd heeft.
3 HOMALS kent vrijwel geen beperkingen wat betreft het aantal rijen van de datamatrix, wat in dit geval betekent dat men een vrijwel onbeperkt aantal stimuli kan laten sorteren.

Terwijl de sorteermethode zich bij uitstek leent voor het laten beoordelen van (zeer) grote aantallen stimuli, zijn er beperkingen met betrekking tot het aantal stimuli dat de gangbare MDS-programma's aan kunnen. Deze beperkingen hebben vooral te maken met het vereiste computergeheugen en de hoeveelheid rekentijd. Bij HOMALS zijn er nauwelijks zulke beperkingen. Drie voorbeelden van onderzoek waarin met grote aantallen stimuli gewerkt is, zijn de studies van Willemsen en Van der Kloot (1987), Verkes, Van der Kloot en Van der Meij (1989) en Van der Kloot en Slooff (1989). In deze onderzoekingen moesten proefpersonen respectievelijk 100, 176 en 119 stimuli sorteren. In het onderzoek van Van der Kloot en Slooff was er sprake van 281 stimuli die in een aantal overlappende deelverzamelingen van 119 stimuli waren verdeeld. Deze deelverzamelingen werden voorgelegd aan verschillende groepen proefpersonen. De sorteringen werden opgeslagen in een stimulus-bij-personen-matrix met 281 rijen. De stimuli die door een bepaalde persoon *niet* beoordeeld waren, kregen in de datamatrix codes die door HOMALS als *missing data* geïnterpreteerd werden. Als iedere stimulus maar door voldoende personen beoordeeld wordt en met een groot aantal andere, in iedere deelverzameling wisselende, stimuli wordt gecombineerd, is HOMALS in principe in staat een goede representatie van alle stimuli te geven.
De toepassing van HOMALS op sorteerdata is al eerder beschreven in Gifi (1990). Eveneens zijn andere programma's voor multipele correspondentieanalyse (dual scaling; Nishisato, 1980, 1994) al eerder toegepast op sorteergegevens. Ook Takane's programma MDSORT (1980, 1981), dat speciaal voor de analyse van sorteerdata geschreven was, is in feite identiek aan multipele correspondentieanalyse en levert ook identieke resultaten op.
Een aantal nuttige suggesties voor de analyse en interpretatie van sorteergegevens met multipele correspondentieanalyse zijn te vinden bij Nishisato (1994) en bij Van der Kloot (1996). Nieuwe algoritmen voor MDS van nabijheidsdata die op sorteerdata gebaseerd zijn, zijn gepresenteerd door Hojo (1993) en Bimler (1996).

13.4 MDS-ANALYSE VAN CORRELATIECOËFFICIËNTEN

Een klassieke vorm van psychologisch onderzoek bestaat uit het afnemen van tests, vragenlijsten of andere meetinstrumenten bij een (groot) aantal proefpersonen. Na de dataverzameling wordt dan meestal een tabel met Pearson

product-moment correlatiecoëfficiënten tussen de variabelen berekend, waarna de onderzoekers via allerlei statistische en data-analytische methoden proberen zinvolle patronen in de correlatiematrix te ontdekken. De methoden die daarbij het vaakst gebruikt worden zijn principale-componentenanalyse, factoranalyse, padanalyse en structurele-vergelijkingsmodellen. Ook MDS kan echter op dit soort gegevens worden toegepast.

Correlatiecoëfficiënten als gelijkheidsmaten

Het basisidee hiervoor is heel eenvoudig: van twee variabelen die hoog met elkaar correleren kunnen we zeggen dat ze meer op elkaar lijken en dus een kleinere afstand tot elkaar hebben dan twee variabelen die minder hoog met elkaar correleren. In deze redenering heeft de hoogte van een correlatiecoëfficiënt een (op zijn minst monotoon) dalende relatie met de afstand tussen de variabelen. Correlatiecoëfficiënten kunnen we dus opvatten als (op zijn minst ordinale) gelijkenismaten, zodat we elke willekeurige matrix van correlatiecoëfficiënten zonder meer met een niet-metrisch MDS-programma zouden kunnen analyseren. Op deze manier kunnen we de onderlinge relaties tussen m variabelen altijd perfect afbeelden als afstanden in een ruimte met $m - 2$ dimensies. Het enige waar we op moeten letten is dat het bij correlaties om *similarities* gaat in plaats van dissimilarities (dus: `/LEVEL=ORDINAL(SIMILARITIES)`) en dat correlaties negatief kunnen zijn en afstanden niet. Bij een programma als ALSCAL worden negatieve getallen automatisch als ontbrekende observaties beschouwd, tenzij we door middel van `/CRITERIA CUTOFF (-1.00)` aangeven dat alle getallen groter dan of gelijk aan -1.0 mee mogen doen.

Correlatie en afstand

In deze paragraaf zullen we zien dat er een precies te omschrijven relatie bestaat tussen correlatiecoëfficiënten en afstanden in een ruimte. Dit is het gemakkelijkst aan te tonen met behulp van een voorbeeld. Stel, we hebben de metingen van n objecten op drie variabelen X, Y en Z. Om deze gegevens ruimtelijk weer te geven, zijn er de volgende twee alternatieven:

1. We beelden de *objecten* af als n punten in een ruimte, met de variabelen X, Y en Z als eerste, tweede en derde as. De metingen op die variabelen fungeren dan als de coördinaten van de objecten. Deze manier van afbeelden heet het *puntenwolkmodel* (zie Van de Geer, 1967).
2. We beelden de *variabelen* af als drie punten in een ruimte waarvan de assen gevormd worden door de n objecten. De meting van een object op een variabele fungeert dan als coördinaat van die variabele op de as die correspondeert met het betreffende object. We hebben dan een configuratie van drie punten in n dimensies. Deze manier van afbeelden wordt het *vectormodel* genoemd, met name omdat het gebruikelijk is de punten die de variabelen voorstellen met de oorsprong te verbinden. Op die manier ontstaat er een afbeelding waarin de variabelen door pijlen (vectoren) worden voorgesteld.

Wat is nu de relatie tussen zo'n vectormodel en de correlaties tussen de variabelen? Zoals elk elementair statistiekboek laat zien, is de formule voor de correlatie tussen twee variabelen X en Y

$$r_{XY} = \frac{\left(\Sigma(x_i - \bar{x})\,(y_i - \bar{y})\right)/(n-1)}{\sqrt{\left(\frac{\Sigma_i(x_i - \bar{x})^2}{n-1}\right)\left(\frac{\Sigma(y_i - \bar{y})^2}{n-1}\right)}} \qquad [13.6]$$

zodat

$$r_{XY} = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X\sqrt{n-1}}\right)\left(\frac{y_i - \bar{y}}{s_Y\sqrt{n-1}}\right) = \sum_{i=1}^{n}\left(\frac{{}^{Z}x_i}{\sqrt{n-1}}\right)\left(\frac{{}^{Z}y_i}{\sqrt{n-1}}\right). \qquad [13.7]$$

Gebruikmakend van de theorie uit Hoofdstuk 2, zien we dat volgens bovenstaande formule een correlatiecoëfficiënt is op te vatten als het *scalaire product* van een stel *getransformeerde coördinaten* op n dimensies. In dit geval zijn dat de coördinaten $\{x_i\}$ en $\{y_i\}$ van de variabelen X en Y op dimensies die met de objecten corresponderen. De metingen van de objecten op elke variabele worden eerst verminderd met het gemiddelde van de desbetreffende variabele, vervolgens gedeeld door de standaarddeviatie van die variabele en ten slotte gedeeld door $\sqrt{(n-1)}$. Gebruiken we deze getransformeerde metingen als coördinaten van de variabelen op assen door de objecten, dan krijgen we een verzameling van punten die allemaal even ver van de oorsprong af liggen (want $\sum_i\{z^2_{Xi}/(n-1)\} = 1.00$). De vectoren vanuit de oorsprong naar de punten van de variabelen hebben dus allemaal de lengte 1. Nu weten we op grond van de *cosinusregel* (zie Formule [2.5]) dat de afstand tussen twee punten p en q gelijk is aan

$$d_{pq} = \sqrt{d^2_{Op} + d^2_{Oq} - 2\sum_{s=1}^{r} x_{ps}x_{qs}} \qquad [13.8]$$

waarbij d_{Op} en d_{Oq} de afstanden zijn van de punten p en q tot de oorsprong O (x_{ps} en x_{qs} zijn de coördinaten van p en q op dimensie s). De correlatie tussen twee variabelen X en Y is dus te herleiden tot de afstand d_{XY} volgens

$$d_{XY} = \sqrt{1+1 - 2\sum_{i=1}^{n}\frac{{}^{Z}x_i\,{}^{Z}y_i}{n-1}} = \sqrt{2 - 2r_{XY}} \qquad [13.9]$$

Met andere woorden: correlatiecoëfficiënten zijn een perfecte, monotoon dalende functie van de Euclidische afstanden tussen punten in een ruimte met *standaardscores* als coördinaten. Niet-metrische MDS op een matrix met corre-

latiecoëfficiënten moet *in principe*[5] hetzelfde opleveren als MDS op een matrix met Euclidische profielafstanden, berekend over de per variabele gestandaardiseerde scores van de objecten.

MDS en exploratieve factoranalyse

Exploratieve factoranalyse, met name *principale-componenten (factor)analyse* is een verzameling technieken waarmee men wil nagaan hoeveel en welke dimensies er ten grondslag liggen aan (de correlaties tussen) een verzameling variabelen. In feite maken alle factoranalysetechnieken gebruik van de ontbinding in eigenwaarden en eigenvectoren van de correlatiematrix **R** volgens **R** = **GΛG'**. Aangezien – zoals we hierboven hebben gezien – **R** een matrix van scalaire producten is, is de ontbinding van **R** in eigenwaarden en eigenvectoren niets anders dan metrische MDS volgens de Young-Householdermethode (zie Hoofdstuk 2). Het belangrijkste verschil zit in het feit dat correlatiecoëfficiënten scalaire producten zijn in een ruimte met een vaste oorsprong. De oorsprong ligt namelijk in het nulpunt van de standaardscores, dat wil zeggen in het gemiddelde van de scores van de objecten op de variabelen. Bij toepassing van de Young-Householdermethode, zoals die in Hoofdstuk 2 geschetst is, mogen we de oorsprong van de ruimte vrij kiezen, waarbij het gebruikelijk is om de oorsprong te laten samenvallen met de centroïde van de punten. In sommige gevallen kan dit een (op het eerste gezicht) andere oplossing geven: zie Blok 13.3.

MDS-analyse van de correlaties tussen vragenlijstitems wordt vaak toegepast door onderzoekers die in de traditie van de *facet-theorie* werken. Zie bijvoorbeeld Guttman (1960) voor een klassiek artikel op dit gebied en Borg en Shye (1995) voor een recente inleiding in deze aanpak. Deze onderzoekers vinden dat MDS-configuraties een duidelijker beeld geven van de overeenkomsten en verschillen tussen de vragenlijstitems dan bijvoorbeeld factoranalyse dat doet.

5 Exact dezelfde oplossingen krijgen we als we een metrische MDS (`/LEVEL=RATIO`) uitvoeren op de Euclidische profielafstanden en op correlatiecoëfficiënten die getransformeerd zijn volgens $f(r) = \sqrt{(2 - 2r)}$. Doen we niet-metrische analyses op zowel profielafstanden als correlaties, dan hoeven we niet precies dezelfde oplossingen te krijgen omdat het MDS-algoritme in beide analyses verschillende transformatiefuncties zal vinden. Toch verwachten we dat beide analyses inhoudelijk vergelijkbare oplossingen zullen opleveren.

BLOK 13.3 PRINCIPALE-COMPONENTENANALYSE EN MDS

In Tabel 13.10 staan de scores van tien objecten op zes variabelen. Deze scores zijn zodanig geconstrueerd dat de eerste (X) en vijfde (V) variabele onderling niet gecorreleerd zijn. De tweede (Y), derde (Z) en vierde (U) variabele zijn uit X en V geconstrueerd volgens $Y = (2X + V)/3$, $Z = (X + V)/2$ en $U = (X + 2V)/3$. Daardoor zijn alle correlatiecoëfficiënten nul of positief. Omdat Y, Z en U lineaire combinaties van U, X en V zijn, is de matrix met scores van rang twee. Bij principale-(componenten)factoranalyse blijkt daarom dat deze correlatiematrix perfect verklaard kan worden met behulp van twee factoren. De correlatiecoëfficiënten, factorladingen en eigenwaarden zijn weergegeven in Tabel 13.11; een ruimtelijke afbeelding van de factoroplossing staat in Figuur 13.7a.

Tabel 13.10 Scores van tien objecten op vijf variabelen

Object	X	Y	Z	U	V
1	1.000	1.667	2.000	2.333	3.000
2	2.000	2.000	2.000	2.000	2.000
3	2.000	2.667	3.000	3.333	4.000
4	3.000	2.333	2.000	1.667	1.000
5	3.000	2.667	2.500	2.333	2.000
6	3.000	3.333	3.500	3.667	4.000
7	3.000	3.667	4.000	4.333	5.000
8	4.000	3.333	3.000	2.667	2.000
9	4.000	4.000	4.000	4.000	4.000
10	5.000	4.333	4.333	3.667	3.000

Niet-metrische MDS van de correlatiematrix levert in twee dimensies een oplossing met $Stress_1 = .0006$ en in één dimensie een oplossing met $Stress_1 = 0$. In de tweedimensionale oplossing blijft het algoritme steken in een lokaal minimum. De coördinaten van de een- en tweedimensionale oplossingen staan in Tabel 13.12; de ruimtelijke afbeeldingen in Figuur 13.7b en 13.7c.

Tabel 13.11 Correlaties, factorladingen en eigenwaarden van vijf variabelen

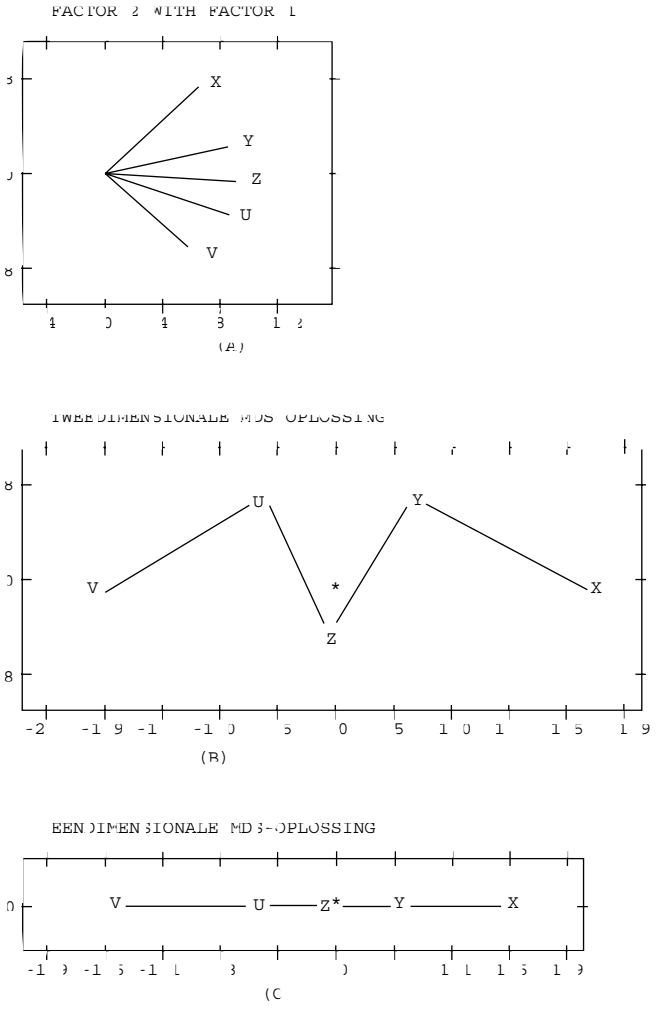
	X	Y	Z	U	V	F1	F2
X	1.000	.880	.679	.420	.000	.683	.731
Y	.880	1.00	.946	.801	.475	.948	.318
Z	.679	.946	1.00	.951	.734	1.000	-.005
U	.420	.801	.951	1.00	.907	.950	-.313
V	.000	.475	.734	.907	1.00	.731	-.683
					Eigenwaarden	3.801	1.199

Tabel 13.12 Coördinaten uit de MDS-oplossing van correlatiecoëfficiënten

	Twee dimensies:		Eén dimensie
	Dim. 1	Dim. 2	Dim. 1
X	2.075	-.081	1.522
Y	.698	.384	.512
Z	-.039	-.612	-.027
U	-.751	-.408	-.551
V	-1.984	.098	-1.455
	$Stress_1$	.0006	.0000

Merk op dat de coördinaten op de tweede dimensie van de MDS-oplossing een sinusachtige functie van de coördinaten op de eerste dimensie zijn. Ook dit kan erop wijzen dat de 'echte' configuratie eendimensionaal is. Vergelijken we de tweedimensionale factor- en de eendimensionale MDS-oplossingen, dan lijken deze nogal van elkaar te verschillen. Maar als we goed kijken, zien we dat de MDS-oplossing als het ware uit de factoroplossing is ontstaan, namelijk door de oorsprong naar rechts te verschuiven (en vrijwel met *Z* te laten samenvallen) en daarna het cirkelsegment tussen *X* en *V* recht te buigen. Voorzover er een probleem is, betreft dat de inhoudelijke interpretatie van de twee oplossingen. Uitgaande van de factoranalyse zal men ofwel concluderen dat de variabelen *X* en *V* twee onafhankelijke eigenschappen zijn (met name na Varimaxrotatie), ofwel dat zij voor een deel hetzelfde en voor een deel het tegenovergestelde meten. De gemeenschappelijke componenten vallen samen met Factor 1, de tegenovergestelde componenten met Factor 2. In de MDS-oplossing van dit voorbeeld komt alleen deze tweede factor terug. Daardoor zou men ten onrechte geneigd kunnen zijn te concluderen dat de variabelen *X* en *V*

alleen maar elkaars tegengestelde zijn en in het geheel niets met elkaar gemeen hebben. Dat hoeft natuurlijk niet zo te zijn. Of en in hoeverre er inderdaad sprake is van een tegenstelling, is een kwestie van nadere inspectie. Waar de MDS-oplossing duidelijk de aandacht op richt (meer dan de factoranalyse) is de ordening van de variabelen langs de dimensie die van *X* naar *V* loopt. Bij de interpretatie zullen we vooral proberen na te gaan in welke opzichten de opeenvolgende variabelen van elkaar verschillen.



Figuur 13.7 Configuraties van factor en MDS-oplossingen

13.5 DRIEWEG/DRIEMODALE DATA

In Hoofdstuk 3 van dit boek is de term drieweg/driemodale data geïntroduceerd. Het gaat hierbij om data die de nabijheidsrelaties aangeven tussen elementen die uit drie verschillende verzamelingen komen. Daarbij kunnen we denken aan de voorkeursrangordeningen van objecten, door een aantal personen op verschillende tijdstippen gegeven. Bijvoorbeeld: bij drie opeenvolgende verkiezingen (voor respectievelijk Tweede Kamer, Gemeenteraad en Provinciale Staten) vragen wij dezelfde groep van dertig kiezers hun voorkeursrangordeningen aan te geven voor de vijf grootste politieke partijen. We krijgen dan een drieweg datamatrix met dertig rijen, vijf kolommen en drie 'plakjes'. Een ander soort driewegdata, dat in de sociale wetenschappen en in marktonderzoek zeer veelvuldig en in veel verschillende variaties voorkomt, is een matrix met beoordelingen van een aantal objecten (stimuli, producten) door een groep personen op een verzameling schalen. Een klassiek geval is het onderzoek met de semantische differentiaal van Osgood, Suci en Tannenbaum (1957).
In dit hoofdstuk wordt kort ingegaan op de verschillende mogelijkheden die er bestaan om drieweg/driemodale gegevens te analyseren. In principe kan men daarbij drie richtingen inslaan: aggregatie, concatenatie en echte drieweganalyse. Deze methoden zullen we hieronder bespreken.

Aggregatie van drieweg/driemodale data

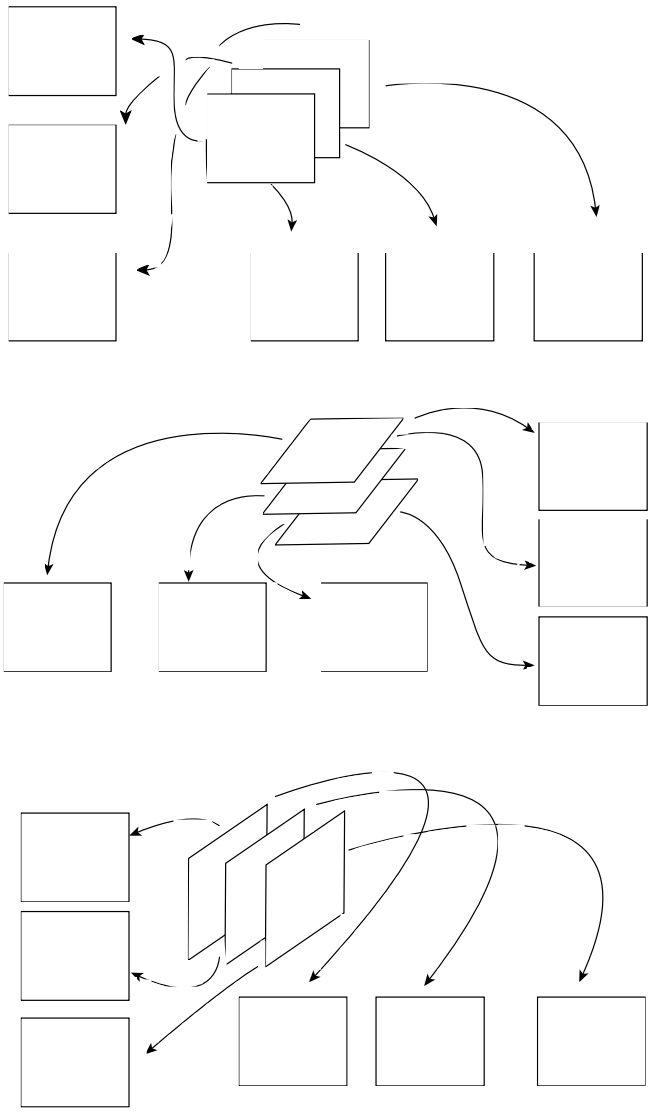
Waar het bij aggregatie op neerkomt, is dat de elementen van een van de wegen worden samengevoegd, meestal door optelling of door gemiddelden te berekenen. Bijvoorbeeld: in een onderzoek waarin n psychotherapiepatiënten op t achtereenvolgende tijdstippen hun eigen welbevinden beoordelen op s symptoomschalen, kunnen we de effecten-in-de-tijd van de therapie bekijken door op elk tijdstip de schalen over de patiënten te middelen. Ook kunnen we per patiënt per tijdstip de scores op de symptoomschalen bij elkaar optellen. Daardoor wordt het mogelijk de vooruitgang in de tijd van elke patiënt te volgen. De derde mogelijkheid, per patiënt en schaal aggregeren over tijdstippen, ligt in dit voorbeeld niet erg voor de hand. Daardoor verdwijnt namelijk alle informatie over veranderingen in de toestand van de patiënten, en dat is toch waar het in de therapie meestal om gaat. Kortom, bij aggregatie wordt sommige informatie (over verschillen; variatie!) behouden en andere 'weggegooid' en hangt het uiteindelijk van de aard en het doel van een onderzoek af welke elementen van welke modus men wel en welke men beter niet moet samennemen. Een ander aspect van aggregatie is dat het niet altijd zinvol is om observaties zomaar bij elkaar op te tellen of te middelen. Als de data op ordinaal niveau gemeten zijn, moeten ze misschien eerst getransformeerd worden voordat ze kunnen worden samengenomen (zie Hoofdstuk 9). Bij dit laatste probleem speelt de conditionaliteit van de driewegmatrix uiteraard een grote rol.

Een speciale vorm van aggregeren bestaat eruit dat men gaat tellen *hoe vaak* een bepaalde waarde voorkomt in de modus die men laat vallen. Bijvoorbeeld: we beschikken over de voorkeursrangordeningen van een aantal kiezers met betrekking tot een aantal politieke partijen op een aantal verkiezingsdata. We kunnen nu tellen (a) hoe vaak elke partij over de verkiezingen heen door een persoon als eerste is gekozen (of bij de eerste drie zat) , en (b) door hoeveel personen een partij per verkiezing op de eerste (tweede, derde) plaats gekozen is. Beide vormen van aggregatie, respectievelijk over verkiezingen en over personen, leveren een kruistabel met frequenties op die op verschillende manieren (waaronder correspondentieanalyse) geanalyseerd zou kunnen worden.
Merk op dat bij de hierboven genoemde vormen van aggregeren de oorspronkelijke drieweg/driemodale matrix wordt omgezet in een tweeweg/tweemodale tabel. In weer een andere manier van aggregatie wordt de drieweg/driemodale matrix getransformeerd in een drieweg/tweemodale matrix. Hoe dat gaat en hoe zo'n matrix geanalyseerd kan worden, wordt aan het eind van dit hoofdstuk in Blok 13.4 uit de doeken gedaan.
Bij alle vormen van aggregatie worden steeds de elementen van een van de modi genegeerd; deze elementen worden als replicaties van elkaar opgevat, waarbij wordt aangenomen dat de manier waarop ze ten opzichte van elkaar variëren inhoudelijk oninteressant is. Het is ook mogelijk driewegdata zonder aggregatie zodanig te analyseren dat een van de modi niet in de uiteindelijke oplossing wordt afgebeeld. Alle data worden dan wel gezamenlijk geanalyseerd, maar volgens een model dat slechts parameters voor twee van de drie modi bevat. Een drieweg/tweemodale variant, *replicated multidimensional scaling* (RMDS), is behandeld in Hoofdstuk 9. Een soortgelijke aanpak is ook mogelijk met drieweg/driemodale data.

Concatenatie

Een drieweg/driemodale datamatrix met respectievelijk r, k en p elementen in de drie wegen c.q. modi, kan men zich op drie manieren voorstellen. Ten eerste kan men hem zien als een blok dat bestaat uit p achter elkaar geplakte matrices met r rijen en k kolommen; de analogie is een cake die al wel in plakjes is gesneden maar nog een geheel vormt. De tweede voorstelling van de datamatrix is als een stapeltje papieren: het datablok is gevormd uit r op elkaar liggende, horizontale vlakken die elk uit een $k \times p$-matrix bestaan. Ten derde kan men de datamatrix vergelijken met boeken in een boekenkast: er zijn k verticaal staande matrices met afmetingen $r \times p$ die van links naar rechts tegen elkaar zijn aangezet. Met elk van deze voorstellingen corresponderen weer twee manieren waarop men de datamatrix zou kunnen concateneren of 'platmaken' door de verschillende plakjes van de matrix naast elkaar te plaatsen.
De plakjes van een *cake* kunnen naast of onder elkaar worden neergelegd. In het eerste geval ontstaat er een (tweeweg)matrix met r rijen en $(p \times k)$ kolommen. In het tweede geval krijgen we een matrix met $(p \times r)$ rijen en k kolommen. Ook de papieren van een *stapeltje* kunnen op twee manieren naast elkaar worden neergelegd: zodanig dat er een matrix met p rijen en

($r \times k$) kolommen of een matrix met ($r \times p$) rijen en k kolommen ontstaat. Ten slotte kunnen de boeken uit een *boekenkast* naast of onder elkaar op een tafel worden uitgespreid. Daardoor ontstaat een matrix met r rijen en ($k \times p$) kolommen, of een matrix met ($k \times r$) rijen en p kolommen. Al deze mogelijkheden zijn weergegeven in Figuur 13.8. Welke mogelijkheid men in een bepaald geval kiest, hangt uiteraard van de aard en het doel van het onderzoek af.



Figuur 13.8 De verschillende concatenatiemogelijkheden van een drieweg datamatrix

Heeft men eenmaal een driewegmatrix tot een van bovengenoemde tweewegmatrices geconcateneerd, dan hangt het van het type observaties (met name hun inhoud, meetniveau en conditionaliteit) af wat voor analysemethode men op die data zou kunnen toepassen. Als we ons realiseren dat we in het algemeen weer een vectormodel en een ideaalpuntmodel kunnen gebruiken, dan wordt het aantal mogelijkheden (twee modellen, zes concatenatievormen, twee meetniveaus en twee of drie conditionaliteiten) veel te groot om allemaal in dit boek te behandelen. We beperken ons daarom tot een paar interessante gevallen.

Ontvouwing van drieweg voorkeursrangordeningen

We nemen als voorbeeld een datamatrix van r kiezers bij k politieke partijen op p verkiezingen, waarin de observaties bestaan uit de voorkeursrangordeningen van de kiezers voor de partijen. We gaan ervan uit dat de data ordinaal zijn en rijconditioneel. Dat wil zeggen dat de rangnummers van de partijen alleen met elkaar vergeleken mogen worden binnen de rangordening van één persoon op één verkiezing. Gegeven deze conditionaliteit moeten de data als een cake beschouwd worden waarvan de plakjes onder elkaar gelegd moeten worden. Dat is hier de enig zinvolle manier van platmaken. We krijgen dan een matrix met $(p \times r)$ rijen en k kolommen.

Analyse met behulp van het ideaalpuntmodel levert in principe dan een meerdimensionaal plaatje op waarin de k politieke partijen als punten zijn afgebeeld, met daartussenin $(p \times r)$ ideaalpunten van de personen op de verschillende verkiezingen. Elke proefpersoon wordt dus afgebeeld door middel van p punten in een ruimte van politieke partijen. Dit plaatje maakt het mogelijk om van alle personen de veranderingen in de tijd te bestuderen. Zoals in Hoofdstuk 11 vermeld is, kan niet-metrische analyse volgens het ideaalpuntmodel tot problemen leiden. Dat kan zich vooral voordoen als $(p \times r)$ veel groter is dan k: zie Hoofdstuk 11 voor eventuele oplossingen.

Stel dat de voorkeursrangordeningen niet rijconditioneel, maar persoonsconditioneel zijn. Dat wil zeggen, alle rangnummers die door een en dezelfde persoon gegeven zijn, mogen met elkaar vergeleken worden. In dat geval kan de matrix ook geconcateneerd worden tot een r bij $(p \times k)$ matrix. Ontvouwing geeft dan een afbeelding met r ideaalpunten voor de personen te midden van $(p \times k)$ punten voor de politieke partijen. Op deze manier zouden we dus na kunnen gaan of en hoe de posities van de politieke partijen veranderen in de loop der tijd.

Behalve de analyses die hierboven beschreven zijn, bestaat er in ALSCAL nog de mogelijkheid om drieweg/driemodale ideaalpuntdata te analyseren met *replicated multidimensional unfolding* (RMDU) en met het *gewogen-ontvouwingsmodel* (*weighted multidimensional unfolding*; WMDU). RMDU leent zich per definitie niet voor data met grote (systematische) variatie tussen de replicaties, en ook bij kleine, toevallige verschillen zou het wel eens moeilijk kunnen zijn om een structuur van objecten en ideaalpunten te vinden die goed bij alle replicaties past. WMDU gaat uit van een model waarin de afstanden tussen ideaalpunten en

objecten per replicatie mogen verschillen, namelijk omdat de dimensies van de afbeelding binnen iedere replicatie andere gewichten kunnen krijgen. In principe is dit een interessant model om systematische variatie tussen replicaties te bestuderen. Voorzover bekend zijn er echter nog geen overtuigende toepassingen van dit model beschreven.

Analyse van drieweg voorkeursrangordeningen met het vectormodel

Aangezien politieke voorkeur bij uitstek een onderwerp is dat met het ideaalpuntmodel verklaard kan worden (immers, kiezers prefereren veelal partijen die in het midden van de politieke ruimte liggen; dat past niet goed bij een vectormodel) ligt het niet voor de hand om rangordeningen van politieke partijen met een vectormodel te analyseren. Daarom geven we als voorbeeld r patiënten die gedurende de p weken van een psychotherapie elke week een rangordening geven van de k symptomen die op hen van toepassing zijn. De rangordening loopt van meest van toepassing tot minst van toepassing. Gezien de (rij)conditionaliteit van de data, ligt het ook nu weer voor de hand de data als een cake te beschouwen en de plakjes onder elkaar te zetten, waardoor een matrix van $(p \times r)$ rijen bij k kolommen ontstaat.
Zoals in Hoofdstuk 11 getoond is, is PRINCALS een handige manier om preferentierangordeningen niet-metrisch volgens het vectormodel te analyseren. Daartoe moet de datamatrix echter eerst gekanteld worden, zodat er een nieuwe matrix van k objecten en $(p \times r)$ variabelen ontstaat. De symptomen worden door middel van de objectscores weergegeven als punten in een ruimte en iedere patiënt wordt daarin afgebeeld als een bundeltje vectoren. Veranderingen in de richting van die vectoren kunnen dus inzicht geven in de effecten van de therapie.
Zouden de data niet rijconditioneel maar persoonsconditioneel zijn, dan zou men de PRINCALS-analyse ook kunnen doen op een matrix van $(p \times k)$ objecten bij r variabelen. Elke patiënt wordt dan voorgesteld door één vector, terwijl elk van de symptomen door middel van p punten wordt afgebeeld. Op deze manier zouden we veranderingen in de prominentie van symptomen kunnen bestuderen.

Beoordelingen van objecten door personen op schalen

Een type data dat veel voorkomt in allerlei soorten onderzoek binnen allerlei disciplines bestaat uit beoordelingen van objecten (stimuli, producten) op een aantal beoordelingsschalen (rating scales) door een aantal personen. Zo kan men een groep kiezers vragen de tien grootste politieke partijen te beoordelen op de kenmerken 'links - rechts', 'confessioneel - onconfessioneel', 'progressief - behoudend', 'verkwistend - zuinig', 'stabiel - chaotisch', 'principieel - opportunistisch', enzovoort. Meestal moeten de respondenten hun beoordelingen in een cijfer uitdrukken op een beoordelingsschaal die doorgaans de getallen 1 - 5, 1 - 7, 1 - 9 of (net als schoolcijfers!) 1 - 10 bevat. Als we deze getallen als ordinaal opvatten en als persoonsconditioneel, kunnen ze na concatenatie geanalyseerd worden op een van de manieren die hierboven besproken zijn.

Daarbij ligt het vectormodel waarschijnlijk het meest voor de hand. Vatten we de observaties op als rij-, dat wil zeggen, persoonsconditionele intervaldata dan kunnen diezelfde analysemethoden metrisch worden toegepast (bijvoorbeeld SVD in plaats van PRINCALS). Als we echter de observaties als onconditionele, metrische data beschouwen, dan komen er een aantal andere analysemogelijkheden bij. Voorbeelden zijn multivariate variantie-analyse (MANOVA) met herhaalde metingen, structurele-vergelijkingsmodellen (zoals LISREL en EQS) en exploratieve of confirmatieve factoranalyse. Aangezien die methoden niet onder de meerdimensionale schaaltechnieken vallen, laten we ze hier buiten beschouwing. Wel bespreken we tot slot nog twee technieken die wel als schaalmethoden beschouwd kunnen worden: drieweg principale-componentenanalyse en een bijzondere toepassing van analyse met het INDSCAL-model.

Drieweg principale-componentenanalyse

Het doel van drieweg principale-componentenanalyse is een matrix met drieweg observaties o_{ijk} te herleiden tot het product van een aantal tweewegmatrices die de coördinaten bevatten van de rij-, kolom- en plakjeselementen van de driewegmatrix, op een aantal dimensies die bij de betreffende modi horen. Het achterliggende idee is dat de verschillende objecten, schalen en personen gekwantificeerd kunnen worden op een aantal dimensies die respectievelijk de variatie tussen de objecten, schalen en personen beschrijven. Het aantal dimensies van de verschillende modi kan daarbij gelijk zijn, maar dat hoeft niet. Het is mogelijk dat de objecten in twee, de schalen in drie en de personen in één dimensie variëren. In formulevorm is het achterliggende model (zie Kroonenberg, 1983a, 1996b)

$$o_{ijk} = \sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{t=1}^{T} x_{ir} y_{js} z_{kr} g_{rst}. \qquad [13.10]$$

In dit model is o_{ijk} een observatie in rij i, kolom j en plakje k van de driewegdatamatrix met n rijen, m kolommen en p plakjes. Het symbool x_{ir} duidt de coördinaat van rij i op de r-e dimensie van de rijen (bijvoorbeeld de objecten) aan; y_{js} is de coördinaat van de j-e kolom op dimensie s van de kolommen (bijvoorbeeld de schalen) en z_{kt} is de coördinaat van plakje k op dimensie t van de plakjes (bijvoorbeeld de personen). Het symbool g_{rst} is een element van de zogenaamde *kernmatrix* die de relatie aangeeft tussen de dimensies van de drie modi. Deze kernmatrix heeft R rijen, S kolommen en T plakjes, evenveel als de aantallen dimensies van de drie modi (NB: in het algemeen willen we natuurlijk dat het model eenvoudiger is dan de observaties, dus dat $R < n$, $S < m$ en $T < p$). De kernmatrix is zelf dus ook een driewegmatrix, wat de interpretatie niet eenvoudig maakt. Het model gaat er iets simpeler uitzien als we het anders opschrijven:

$$o_{ijk} = \sum_{r=1}^{R} \sum_{s=1}^{S} x_{ir} y_{js} \left(\sum_{t=1}^{T} z_{kr} g_{rst} \right) = \sum_{r=1}^{R} \sum_{s=1}^{S} x_{ir} y_{js} h_{rs}^{(k)}. \qquad [13.11]$$

Hierin is de notatie $h_{rs}^{(k)}$ gekozen om de elementen aan te geven in het *k*-e plakje van de driewegmatrix **H** met rijen *r*, kolommen *s* en plakjes *k*. In matrixnotatie kunnen we dit plakje aanduiden als $\mathbf{H}^{(k)}$. Noteren we de plakjes van de oorspronkelijke driewegmatrix **O** als $\mathbf{O}^{(k)}$ dan wordt het eenvoudiger om het model in matrixnotatie te schrijven:

$$\mathbf{O}^{(k)} = \mathbf{X}\mathbf{H}^{(k)}\mathbf{Y}' . \qquad [13.12]$$

Bovenstaande formules geven het drieweg principale-componentenmodel weer dat in 1966 door Tucker bedacht is; daarom wordt dit het Tucker-model genoemd. Formule [13.12] beschrijft het zogenaamde Tucker3-model (Kroonenberg, 1983a). In dit model wordt aangenomen dat het aantal dimensies (*T*) van de plakjes kleiner is dan *p*, het aantal plakjes van de datamatrix. Het Tucker3-model streeft dus naar datareductie in alle drie de modi. Een variant van dit model probeert slechts voor twee van de drie modi onderliggende componenten te schatten. Dit is het Tucker2-model dat er net zo uitziet als in Formule [13.12], met dien verstande dat de matrix $H^{(k)}$ geen speciale structuur heeft, dat wil zeggen, niet verder herleid kan worden tot het product van een kernmatrix en coördinaten op componenten van de plakjes-modus. Een bijzonder geval van het Tucker3-model is het zogenaamde parallelle-factorenmodel (Parafac) dat we hier verder niet bespreken.
In de afgelopen decennia heeft Kroonenberg (1996a,b) een uitgebreid softwarepakket, 3WAYPACK, voor drieweg principale-componentenanalyse ontwikkeld. In dit pakket wordt een *alternating least squares* algoritme toegepast; vandaar de namen TUCKALS2 en TUCKALS3. Voordat deze programma's gebruikt kunnen worden, moeten de observaties soms nog worden voorbewerkt. Met name kunnen driewegdata op verschillende manieren in afwijkingen van gemiddelden gezet en eventueel verder gestandaardiseerd worden.
Er bestaat een uitgebreide literatuur met toepassingen van de TUCKALS-programma's (zie Kroonenberg, 1983b, 1997). In drie van die toepassingen (Van der Kloot & Kroonenberg 1982, 1985; Van der Kloot, Kroonenberg, & Bakker, 1985) ging het om onderzoek waarin proefpersonen van een aantal stimuluspersonen (hypothetische personen die met behulp van adjectieven beschreven zijn) moesten beoordelen in welke mate deze stimuluspersonen bepaalde persoonlijkheidseigenschappen bezaten. De analyses lieten zien dat de persoonlijkheidseigenschappen en de stimuluspersonen goed konden worden weergegeven in een tweedimensionale ruimte, terwijl de proefpersonen eigenlijk maar op één dimensie varieerden. Dat betekent dat alle proefpersonen min of meer hetzelfde plaatje van stimuluspersonen en eigenschappen 'in hun hoofd hadden', maar dat zij verschilden met betrekking tot de grootte van de getallen die zij bij hun beoordelingen gebruikten. Er bleek dus sprake te zijn van verschillen in *response style* en niet van individuele verschillen met betrekking tot de *inhoud* van hun cognitieve structuren.

BLOK 13.4 INDSCAL OP PROFIELAFSTANDEN

In bovenstaande paragrafen over drieweganalyse zijn verschillende manieren van aanpak besproken die van drieweg-datamatrices tweeweg-matrices maken. Deze manieren liggen op een continuüm, met als ene uiterste het elimineren van de derde weg (via aggregatie) en als andere uiterste het gebruiken en afbeelden van de informatie van alle wegen en modi.
In dit blok wordt een aanpak behandeld waarin de derde modus geëlimineerd wordt zonder daarbij de derde weg kwijt te raken; de observaties worden herleid tot drieweg/*tweemodale* data. Deze aanpak berust op het berekenen van *profielafstanden* (zie Hoofdstuk 4) tussen de rij-, kolom- of plakjeselementen van de driewegmatrix. Bijvoorbeeld in het klassieke objecten × schalen × personen geval, is er voor elke persoon een tweewegmatrix van objecten × schalen geobserveerd. Binnen ieder van die matrices kunnen we de profielafstanden uitrekenen tussen de objecten (dus: afstanden, gedefinieerd over de schalen). Er ontstaat dan een drieweg/tweemodale matrix van objecten × objecten × personen. Een tweede mogelijkheid is om voor elke persoon de profielafstanden tussen de schalen (gedefinieerd over de objecten) te berekenen. In dat geval krijgen we een drieweg/tweemodale matrix van schalen × schalen × personen. Zulke matrices kunnen met RMDS, INDSCAL of IDIOSCAL geanalyseerd worden. INDSCAL levert dan een groepsconfiguratie van de objecten (c.q. de schalen) op en een tabel met de gewichten die door de personen aan de dimensies worden toegekend.
Een bijzondere, inventieve, toepassing van deze aanpak is beschreven door Wish, Deutsch en Kaplan (1976). De 'objecten' in hun onderzoek waren omschrijvingen van 44 interpersoonlijke relaties (onder andere ouder - kind, echtgenoot – echtgenote, docent - student) die door 87 proefpersonen beoordeeld werden op 25 schalen zoals coöperatief - competitief, intens - oppervakkig, vriendelijk - vijandig. Uit deze gegevens werden *voor iedere schaal afzonderlijk* profielafstanden berekend tussen de relaties, dus over de proefpersonen heen. INDSCAL-analyse van de driewegmatrix met profielafstanden leverde een vierdimensionale configuratie van relaties op en een verzameling gewichten voor de schalen. Deze gewichten laten dus zien voor welke schaal en welke dimensie de afstanden tussen de objecten groot zijn en voor welke schaal en welke dimensie die afstanden klein of misschien zelfs nul zijn. De gewichten geven dus de associaties tussen schalen en dimensies aan, waardoor de stimulusconfiguratie gemakkelijk(er) geïnterpreteerd kan worden.
Overigens stelt het afleiden van profielafstanden wel enige eisen aan de data. Immers, profielafstanden berusten op verschilscores en dus moeten meetniveau en conditionaliteit van de gegevens het toelaten dat men zulke verschilscores berekent. Daarbij komt nog dat profielafstanden op ruwe

observaties in feite drie verschillende componenten bevatten. Profielafstanden zijn gevoelig voor verschillen in gemiddelde beoordeling (*elevation*), voor verschillen in spreiding (*scatter*) en voor verschillen in profielvorm (*shape*; zie Skinner, 1978). Daarom moet men soms de observaties in afwijkingen van het gemiddelde zetten en eventueel ook standaardiseren voordat het zinvol is om afstanden te berekenen.