HOOFDSTUK 7

Interpretatie van MDS-oplossingen

7.1 Aspecten van oplossingen

Het uiteindelijke resultaat van een MDS-analyse bestaat onder andere uit een verzameling *coördinaten* op een *aantal dimensies* en een *stresswaarde* die de goodness-of-*fit* van de oplossing aangeeft. De stress en het aantal dimensies zijn twee grootheden die samenhangen, omdat in het algemeen geldt: hoe meer dimensies, hoe lager de stress. Bij de interpretatie van een MDS-oplossing gaat het dus primair om de keuze van het aantal dimensies waarin we de objecten willen weergeven. Hoewel deze keuze in eerste instantie gemaakt wordt aan de hand van een aantal formele kenmerken (zoals de grootte van de stress, het stressverloop, het aantal punten in de configuratie en het aantal observaties) zijn er ook inhoudelijke aspecten die bij deze keuze een rol spelen. Immers: we willen een configuratie inhoudelijk kunnen begrijpen en daarbij speelt de dimensionaliteit van die configuratie een belangrijke rol.
Wat betreft de inhoudelijke interpretatie van een MDS-configuratie kan men vier manieren van aanpak onderscheiden. De eerste aanpak komt erop neer dat men op min of meer *intuïtieve manier* inzicht in de configuratie tracht te verkrijgen: men maakt een plaatje en hoopt dat de 'echte' relaties tussen de objecten daarin zichtbaar worden. In de tweede aanpak gebruikt men de resultaten van *additionele analyses* van dezelfde data waarop de MDS-oplossing is gebaseerd. Bijvoorbeeld: men voert op een matrix met nabijheidsgegevens zowel een MDS-analyse als een clusteranalyse uit en gebruikt de gevonden clusters om de MDS-configuratie te interpreteren.[1] In de derde aanpak gebruikt

1 Clusteranalyse is een familie van analystechnieken die op basis van nabijheidsdata een indeling van de onderzochte objecten in groepen (clusters) tot stand brengt.

men *additionele gegevens* (met name over eigenschappen van de objecten) om de oplossing inzichtelijk te maken. In dit derde geval spreekt men van *externe* analyse (zie Carroll, 1972). Ten slotte is het mogelijk de configuratie in zijn geheel te vergelijken met een andere, bijvoorbeeld met een MDS-oplossing die in eerder onderzoek was gevonden, of met een configuratie die op grond van *a priori*-hypothesen was opgesteld. In dit geval spreken we van *Procrustes-analyse*. In dit hoofdstuk worden deze vier verschillende interpretatiemethoden behandeld.

7.2 DIMENSIONALITEIT EN STRESS

Aangezien dimensionaliteit en stress onverbrekelijk aan elkaar gekoppeld zijn, is de *stress* van een oplossing een van de belangrijkste indicatoren van de correcte dimensionaliteit. Behalve de stress zijn er ook enkele *a priori*-overwegingen, op grond waarvan men iets kan zeggen over het aantal dimensies, met name over het *maximum* aantal dimensies dat voor een oplossing acceptabel is. Deze overwegingen betreffen het aantal punten en het aantal observaties waarover men beschikt.

Dimensionaliteit en het aantal punten

Zoals de meeste andere multivariate methoden is MDS een methode waarbij men streeft naar *data-reductie*. Het doel van MDS is om het complex van relaties tussen een (groot) aantal objecten op een relatief simpele manier weer te geven, dat wil dus zeggen, in een ruimte die (veel) kleiner is dan de $(m-1)$-dimensionale ruimte die m punten in principe op kunnen leveren.[2] Nu kan men natuurlijk twisten over de vraag hoeveel dimensies nog relatief simpel zijn en welk aantal veel kleiner is dan m. Over het algemeen kunnen we echter stellen dat een oplossing met meer dan $\frac{1}{2}m$ dimensies niet *veel* minder dimensies heeft dan het maximale aantal $m-1$. Ook kunnen we volhouden dat een oplossing met vier of vijf dimensies niet echt simpel is. Behoudens uitzonderingen waarin men expliciet oplossingen met grotere dimensionaliteit verwacht, zullen over het algemeen slechts MDS-oplossingen met één tot hooguit vijf dimensies acceptabel zijn.

Dimensionaliteit en het aantal observaties

Als men van m punten een oplossing in r dimensies zoekt, dan laat men het MDS-algoritme $m \times r$ onbekende parameters schatten. In het CMDS-geval beschikt men slechts over één datamatrix met $m(m-1)/2$ nabijheidsgegevens. Nu is het een algemeen aanvaard uitgangspunt dat men, om betrouwbare

2 In het algemeen geldt dat m punten altijd in een $(m-1)$-dimensionale ruimte kunnen worden afgebeeld. Bij niet-metrische MDS gaat daar nog een dimensie van af: m punten passen perfect in een $(m-2)$-dimensionale ruimte.

schattingen te krijgen, altijd (veel) meer data moet verzamelen dan men parameters wil schatten. Anders gezegd: men dient (veel) minder parameters te schatten dan men observaties heeft. Ook hier kan men weer van mening verschillen over het woordje 'veel'. De meeste onderzoekers zijn het er echter over eens dat men minstens twee keer zoveel observaties moet hebben als het aantal parameters dat men wil schatten. In dat geval is $2(m \times r) \leq m(m-1)/2$ zodat Formule [7.1] een goede vuistregel voor het aantal dimensies is:

$$r \leq (m-1)/4. \qquad [7.1]$$

ALSCAL houdt als criterium $r \leq (m-1)/5$ aan. Wanneer de opgegeven dimensionaliteit groter is, reageert ALSCAL met de waarschuwing die in Blok 6.2 is afgedrukt.

Stress

In de vorige paragrafen zijn enkele overwegingen naar voren gebracht die helpen een bovengrens aan het aantal dimensies te bepalen. Daarmee is het probleem van de dimensionaliteit echter niet opgelost. Een zeer belangrijke aanwijzing voor het bepalen van de dimensionaliteit is de stresswaarde. Het ligt immers voor de hand om die oplossing te kiezen die een acceptabele stress heeft bij een zo laag mogelijke dimensionaliteit. Kruskal (1964a) heeft voor het CMDS-geval de volgende richtlijnen gegeven. Een stress van .20 noemde hij *poor*, .10 *fair*, .05 *good*, .025 *excellent* en .000 *perfect*. Toch is het riskant om deze kwalificaties altijd blindelings toe te passen. In de eerste plaats gelden zij alleen voor Kruskals $Stress_1$ en in de tweede plaats hebben zij uitsluitend betrekking op matrixconditionele CMDS-problemen waarin Kruskals continue monotone transformatie gebruikt wordt. Kruskals kwalificaties gelden niet zonder meer voor andere analyse-opties[3] en voor andere typen schaalproblemen.

Er zijn verschillende pogingen ondernomen om via *Monte Carlo*-studies statistische normen te bepalen waaraan een stresswaarde moet voldoen om acceptabel te zijn. Een van deze studies is het onderzoek van Wagenaar en Padmos (1971) dat in Blok 7.1 besproken wordt.

Stress en het aantal punten

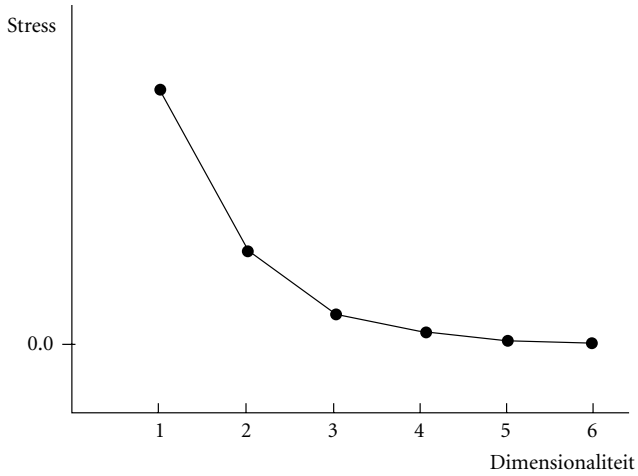
Behalve in die gevallen dat er een perfecte MDS-afbeelding van m punten in r dimensies te krijgen is, is de stress over het algemeen groter naarmate het aantal punten groter is. Dat is als volgt in te zien: voegen we een extra punt toe aan de verzameling objecten, dan voegt dat m extra nabijheidsgegevens toe aan de $m(m-1)/2$ observaties. Als er geen perfecte oplossing is, dan betekent dat,

3 In principe geldt dat de stress van een matrixconditionele oplossing groter is dan die van een rijconditionele analyse, dat een discrete transformatie meer stress oplevert dan een continue transformatie en dat een metrische analyse grotere stress geeft dan een niet-metrische analyse.

dat er geen perfecte monotone relatie tussen de observaties en de afstanden-in-de-oplossing bestaat: sommige observaties staan dan niet in dezelfde volgorde als de afstanden. Hoe meer observaties er zijn, hoe groter de kans is dat er zich verschillen tussen de rangordes van de afstanden en de observaties zullen voordoen, en dus hoe groter de stress in principe zal zijn.

Stressverloop

Het belangrijkste criterium om de dimensionaliteit te bepalen is het stressverloop, de manier waarop de stress afneemt als men oplossingen met meer dimensies kiest. Het is gebruikelijk om het stressverloop grafisch weer te geven zoals in Figuur 7.1 is gedaan. Op de verticale as is de stress van een oplossing uitgezet tegen de dimensionaliteit van die oplossing. In het ideale geval zien we een monotoon dalende curve die bij een bepaalde dimensionaliteit (nagenoeg) gelijk aan nul wordt. Ook als de stress nergens gelijk aan nul wordt, zien we vaak een typische elleboog of knik in de curve. Vóór de knik daalt de stress sterk, als men nieuwe dimensies aan de oplossing toevoegt. Na de knik leveren nieuwe dimensies weinig stressvermindering op. De knik geeft dus de optimale dimensionaliteit van de oplossing aan.

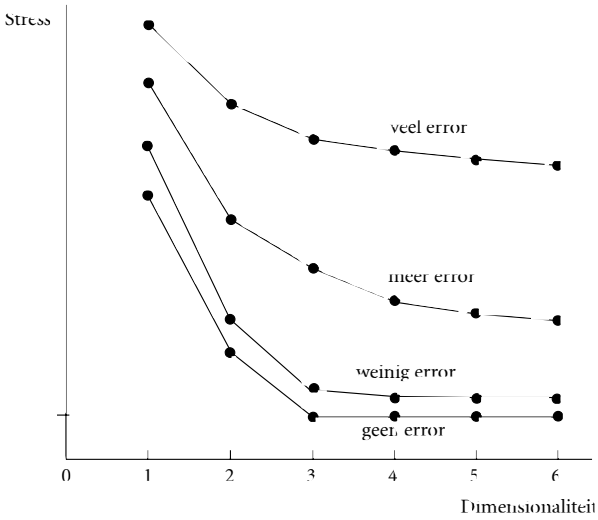


Figuur 7.1 Stressverloop als functie van de dimensionaliteit van de oplossing

Meetfouten en stress

In het ideale geval laat de curve van het stressverloop een scherpe knik zien, met lage stresswaarden voor de dimensionaliteiten na de knik. In de praktijk ziet zo'n curve er vaak anders uit. Soms is er slechts een heel flauwe knik (of helemaal geen) en blijven de stresswaarden ook bij vier, vijf en zes dimensies nog op een relatief hoog niveau. Er is dan geen 'goede' oplossing te vinden in een (betrekkelijk) lage dimensionaliteit. De oorzaak hiervan kan zijn dat er

toevallige of systematische meetfouten (*error)* in de data zitten. Om in zo'n geval te beslissen wat de 'correcte' dimensionaliteit van een oplossing is, zouden we moeten weten hoe het stressverloop eruitziet bij verschillende combinaties van hoeveelheid error en 'ware' dimensionaliteit. Dat kunnen we te weten komen door de volgende procedure toe te passen.
Stel, we construeren een matrix met echte Euclidische afstanden tussen *m* punten in een driedimensionale ruimte. Stel vervolgens dat we deze afstanden op twee manieren veranderen (verstoren) door er random error aan toe te voegen: éénmaal weinig error en éénmaal veel error. Als we de oorspronkelijke en de veranderde afstandsmatrices afzonderlijk met MDS analyseren in één, twee, drie, vier en vijf dimensies, dan vinden we voor het stressverloop de drie curven die in Figuur 7.2 getekend zijn. De onderste curve hoort bij de analyse van de oorspronkelijke, foutloze matrix. Deze heeft een scherpe knik bij drie dimensies en heeft vanaf dat punt een stress van nul. De tweede curve van onderen (voor de matrix met weinig error) begint hoger, heeft een flauwe knik bij drie dimensies en houdt daarna stresswaarden die iets groter dan nul zijn. De curves voor de nabijheidsmatrices met meer en veel error beginnen en eindigen veel hoger dan de eerste twee curven en laten nauwelijks een knik zien.



Figuur 7.2 Het stressverloop van MDS van een nabijheidsmatrix zonder error en drie matrices met toegevoegde error. De nabijheidsmatrices hebben betrekking op een driedimensionale configuratie

Stel, ten slotte, dat we een matrix hebben met geobserveerde nabijheidsgegevens voor eenzelfde *m* aantal punten als de geconstrueerde matrices. Voor die matrix weten we dus niet wat de correcte dimensionaliteit van de onderliggende configuratie is en evenmin weten we hoeveel error er in de data zit. We kunnen het plaatje van Figuur 7.2 nu als volgt gebruiken. We analyseren de nieuwe

matrix eveneens in één, twee, drie, vier en vijf dimensies en tekenen de curve van het stressverloop tussen de andere curven in Figuur 7.2. Als de observaties meer error bevatten dan die van de bovenste curve, dan komt de nieuwe curve daarboven te liggen en zal deze in nog mindere mate een knik vertonen. Bevatten de geobserveerde data minder error dan de bovenste, dan komt de nieuwe curve ergens tussen de andere drie uit. Heeft de nieuwe curve een knik op min of meer dezelfde plaats als de andere, dan kan men aannemen dat de onderliggende configuratie driedimensionaal is. Ligt de knik meer naar links, dan heeft de geobserveerde dataset een kleinere dimensionaliteit. Ligt de knik meer naar rechts, dan is er sprake van een grotere dimensionaliteit.
Als we nu ook afstandsmatrices construeren van bekende configuraties die respectievelijk één-, twee- en vierdimensionaal zijn en daar vervolgens wisselende hoeveelheden error aan toevoegen, dan kunnen we er drie figuren bijmaken die analoog zijn aan Figuur 7.2. Het idee is nu om het stressverloop van een geobserveerde afstandsmatrix met al deze vier figuren te vergelijken. De figuur waar de nieuwe stresscurve het beste in past, heeft dan waarschijnlijk dezelfde dimensionaliteit als de nieuwe, onbekende configuratie. Dit principe ligt ten grondslag aan twee methoden om de dimensionaliteit te bepalen die door Wagenaar en Padmos (1971) en Spence en Graef (1974) ontwikkeld zijn. Deze methoden worden in Blok 7.1 behandeld.

Oorzaken van hoge stress

Zoals hierboven is uiteengezet, zijn de twee voornaamste oorzaken van hoge stress een te klein aantal dimensies van de oplossing en meetfouten in de observaties. Ook is betoogd dat de stress in het algemeen hoger wordt naarmate er meer punten moeten worden afgebeeld. Naast deze factoren zijn er nog enkele andere die van invloed zijn op de hoogte van de stress.
In de eerste plaats kan de stress aan de hoge kant zijn doordat het MDS-algoritme in een *lokaal minimum* is terechtgekomen. Die situatie wordt in de volgende paragraaf nader besproken.
In de tweede plaats kan een hoge stress het gevolg zijn van een aantal opties die bij de analyse gekozen zijn. Bijvoorbeeld, wanneer de observaties een groot aantal *ties* bevatten en de analyse de *discrete aanpak van ties* toepast, kan de stress (veel) hoger uitvallen dan wanneer de continue aanpak gebruikt wordt en de *ties* dus mogen worden losgelaten. Dit doet zich vaak voor wanneer de data uit gelijkenisoordelen bestaan die proefpersonen op zogenaamde *rating scales* hebben uitgebracht. Zelfs bij een klein aantal objecten (bijvoorbeeld tien) zijn er toch al enkele tientallen observaties (45 bij $m = 10$), die maar een beperkt aantal waarden kunnen krijgen omdat de meeste rating scales niet meer dan vijf, zeven of hooguit negen categorieën hebben. Daarom zullen we meestal aannemen dat alle objectparen die dezelfde gelijkeniswaarde hebben gekregen niet noodzakelijk ook precies dezelfde afstand tot elkaar hebben. In dat geval kunnen we dus beter de continue benadering van *ties* kiezen, wat in het algemeen een lagere stress tot gevolg zal hebben.

Een andere analyse-optie die voor de stress van belang is, betreft de *conditionaliteit* van de observaties. Als de nabijheidsdata matrixconditioneel geanalyseerd worden resulteert dat over het algemeen in een hogere stress dan bij de optie rijconditioneel. In een rijconditionele analyse krijgt iedere rij van de nabijheidsmatrix een eigen transformatiefunctie $f_i(o_{ij})$, waardoor de verschillen tussen de afstanden en pseudo-afstanden in het algemeen kleiner zullen worden. Het ligt voor de hand om een rijconditionele analyse uit te voeren als vermoed kan worden dat de paren waarin een bepaald object *h* met andere objecten vergeleken wordt, een ander beoordelingsproces in werking zetten dan de paren waarin object *k* met de andere objecten vergeleken wordt.
Wat de analyse-opties betreft, is vanzelfsprekend ook het veronderstelde *meetniveau* van de data van invloed op de stress. Een metrische analyse zal vrijwel altijd een hogere stress opleveren, omdat de betreffende transformatiefunctie meer beperkingen heeft dan bij een niet-metrische analyse.
Hoge stress kan ook veroorzaakt worden door één of enkele slecht-fittende punten. Stel dat bij de Nederlandse steden ook Alkmaar is vertegenwoordigd. In dat geval zullen de afstanden waarin Alkmaar en Groningen betrokken zijn grotere bijdragen aan de stress leveren dan de andere afstanden. Het kan de moeite waard zijn om dit soort punten (met behulp van het Shepard-diagram) op te sporen en eventueel uit de analyse te verwijderen; niet alleen om de stress te verlagen, maar ook om een betere afbeelding van de andere punten te krijgen.

Lokale minima

Zoals in het vorige hoofdstuk is uitgelegd, kunnen iteratieve MDS-algoritmen blijven steken in zogenaamde *lokale minima*. Dat wil zeggen dat de verkregen configuraties niet optimaal zijn omdat er andere oplossingen bestaan die een lagere stress hebben. Een hoge stress of een vreemde stresscurve kan hiervan het gevolg zijn. Er is maar één manier om na te gaan of er sprake is van een lokaal minimum, namelijk door de MDS-analyse een groot aantal keren te herhalen met verschillende beginconfiguraties. Leveren alle analyses nu dezelfde configuratie met dezelfde stresswaarde op, dan mogen we aannemen dat het algoritme een *globaal minimum* gevonden heeft. Soms zal het voorkomen dat één deel van de analyses naar één identieke oplossing met een lokaal minimum convergeert en een ander deel naar één andere oplossing met een lagere stress. In dat geval kiezen we de oplossing met de laagste stress. Ook is het mogelijk dat er twee of meer verschillende configuraties gevonden worden die (nagenoeg) dezelfde stresswaarde hebben. In dat geval is het de moeite waard om nog meer beginconfiguraties uit te proberen. Blijft het resultaat dan toch hetzelfde, dan zullen inhoudelijke overwegingen de doorslag moeten geven.
Indien men bepaalde inhoudelijke ideeën of hypotheses over de configuratie heeft, dan is het nuttig om in ieder geval ook deze hypothetische configuratie als beginconfiguratie in te voeren. De analyse zal dan laten zien hoe goed deze hypothese bij de data aansluit en of er andere configuraties zijn die een (veel) lagere stress produceren.

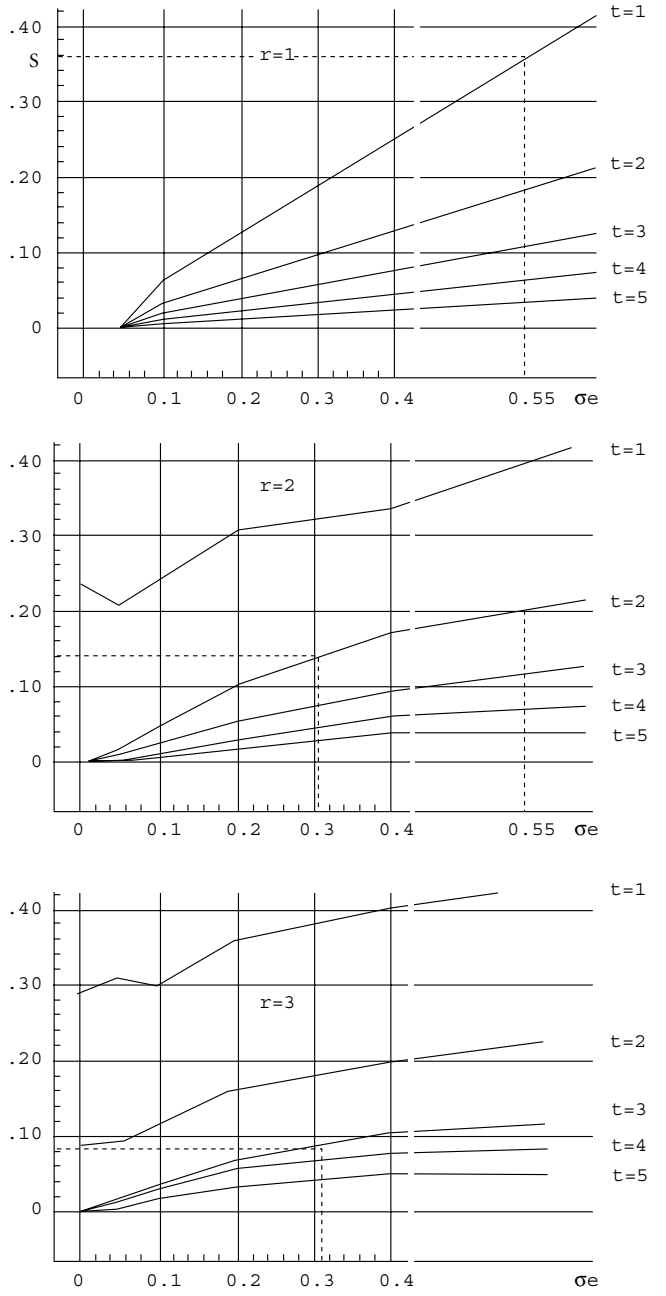
BLOK 7.1 MONTE CARLO-METHODEN VOOR HET BEPALEN VAN DIMENSIONALITEIT

De methode van Wagenaar en Padmos

Zoals hierboven is opgemerkt, noemde Kruskal een *stress* van .20 *poor*, van .10 *fair*, van .05 *good* en van .025 *excellent*. Wagenaar en Padmos (1971) maakten om twee redenen bezwaar tegen het automatisch toepassen van deze kwalificaties. Enerzijds is stress in sterke mate afhankelijk van het aantal objecten. Anderzijds kan men stress niet zomaar beoordelen, omdat een hoge stresswaarde op zijn minst twee oorzaken kan hebben: te weinig dimensies in de oplossingsruimte en gebrek aan overeenstemming tussen de geobserveerde en de 'echte' nabijheden. Deze problemen vormden de aanleiding voor het onderzoek van Wagenaar en Padmos dat tot doel had betere, kwantitatieve richtlijnen voor het beoordelen van stresswaarden te ontwikkelen. Hun onderzoek bestond uit twee delen. In het eerste deel volgden zij een Monte Carlo-procedure die bestond uit de volgende stappen:

1 de constructie van een aantal *r*-dimensionale configuraties van *m* punten (zij noemden deze configuratie de 'psychologische ruimte'). Voor *r* werden de waarden 1, 2 en 3 gekozen, *m* was respectievelijk 8, 10 of 12. De coördinaten van de punten op de dimensies bestonden uit *random* getallen die gekozen werden uit een uniforme verdeling van alle getallen tussen 0 en 1000;
2 het berekenen van de Euclidische afstanden tussen de punten in deze configuraties;
3 het toevoegen van meetfouten (*random error*) aan de afstanden. Iedere afstand d_{ij} werd vermenigvuldigd met een getal k_{ij} dat aselect gekozen was uit een normale verdeling met gemiddelde 1 en standaarddeviatie σ_k. Voor σ_k werden de waarden 0.0, .05, .10, .20 en .40 gekozen.

Al deze keuzen leverden $3 \times 3 \times 5 = 45$ combinaties op. Voor elke combinatie werden 11 verschillende configuraties gegenereerd, die geanalyseerd werden in $t = 1, 2, 3, 4$ en 5 dimensies. In totaal werden dus 2475 MDS-analyses uitgevoerd. Uit deze analyses bleek dat de relatie tussen stress en error op een gecompliceerde manier van *m*, *r* en *t* afhangt. Bijvoorbeeld: als bij de analyse het juiste aantal dimensies gebruikt werd (dus: $t = r$) werden zelfs bij hoge errorniveaus toch lage stresswaarden gevonden (vooral als het aantal punten klein was). Ten tweede: te lage dimensionaliteit (*t*) van de oplossing leidde niet altijd tot hoge stresswaarden. Met name bij tweedimensionale oplossingen van driedimensionale configuraties is de stress vaak vrij laag. Ten slotte bleek dat bij hoge errorniveaus er nauwelijks meer een knik in het stressverloop te onderkennen is. De samenhang tussen stress, *r*, *t* en σ_k voor $m = 12$ is weergegeven in de grafieken van Figuur 7.3.



Figuur 7.3 Het stressverloop als functie van echte dimensionaliteit r*, dimensionaliteit van de oplossing* t *en errornivau* σ *(Wagenaar & Padmos, 1971)*

Wagenaar en Padmos lieten aan de hand van het volgende voorbeeld zien hoe hun resultaten in de praktijk kunnen worden toegepast. Een nabijheidsmatrix van twaalf objecten werd geanalyseerd in één, twee en drie dimensies. De stresswaarden waren .36 .14, en .08. In de grafiek voor $r = 1$ van Figuur 7.3 is na te gaan wat het errorniveau is dat bij een stress van .36 hoort voor een eendimensionale oplossing van een configuratie die in werkelijkheid inderdaad eendimensionaal is. We lezen af dat dit errorniveau zou corresponderen met $\sigma_k \approx .55$. In dat geval zou een tweedimensionale oplossing een stress van ongeveer .20 moeten geven. In de tweedimensionale analyse is echter een stress van .14 gevonden. Dat suggereert dat de ware dimensionaliteit van de configuratie eerder twee dan één is. Bij een stress van .14 voor een tweedimensionale oplossing van een tweedimensionale configuratie hoort volgens de grafiek voor $r = 2$ een errorniveau van $\sigma_k \approx .31$, zodat een driedimensionale analyse van deze configuratie een stress van ongeveer .08 zou moeten hebben. Aangezien dat getal klopt met de gevonden stresswaarde van .07 concluderen we dat de correcte dimensionaliteit van de configuratie inderdaad twee is en dat de observaties een errorniveau $\sigma_k \approx .31$ hadden.

In het tweede deel van hun onderzoek construeerden Wagenaar en Padmos een groot aantal gelijkenismatrices voor respectievelijk $m = 7, 8, 9, 10$, 11 en 12 punten. Deze matrices bevatten de rangnummers 1 tot en met $m(m \times 1)/2$ die steeds in een andere *random* volgorde in de matrices waren ingevuld. Wagenaar en Padmos noemden dit 'errorniveau = ∞'. Van elke matrix werden 50 tot 100 replicaties gegenereerd en elke replicatie werd geanalyseerd in $t = 1, 2, 3, 4$ en 5 dimensies. De stresswaarden die daarbij werden gevonden, bleken een functie van m en t te zijn. Vervolgens is voor elke combinatie van m en t de cumulatieve waarschijnlijkheidsverdeling van de bijbehorende stresswaarden vastgesteld. Daardoor kunnen we voor elke combinatie van m en t nagaan welke stresswaarden een kans van vijf procent of minder hebben om 'toevallig' voor te komen. De grootste van deze waarden noemen we S_{krit}, dat wil zeggen de kritische waarde waar de stress van een bepaalde oplossing onder moet blijven om acceptabel te zijn.

Tabel 7.1 *Kritische stresswaarden S_{krit} waarvoor geldt dat Prob($S \leq S_{krit}$) ≤ .05 als functie van aantal punten (*m*) en dimensionaliteit van de oplossing (*t*)* (Wagenaar & Padmos, 1971)

	t				
m	1	2	3	4	5
7	.200	.070	–	–	–
8	.275	.100	.015	–	–
9	.305	.130	.055	.010	–
10	.340	.150	.070	.030	-
11	.350	.180	.095	.045	.010
12	.395	.205	.100	.065	.035

Wagenaar en Padmos stelden voor de getallen in Tabel 7.1 te gebruiken om uit te maken of een oplossing van een bepaalde dimensionaliteit *significant* is. Zo moet een tweedimensionale oplossing van een 8×8 nabijheidsmatrix een stresswaarde kleiner dan of gelijk aan .10 opleveren om acceptabel te zijn. Een driedimensionale oplossing van dezelfde matrix zou een stress moeten hebben die gelijk is aan of kleiner dan .015.
Wat Tabel 7.1 vooral laat zien is dat er in sommige gevallen stresswaarden voorkomen die (veel) lager zijn dan men intuïtief zou verwachten, bijvoorbeeld alle waarden van de driedimensionale oplossingen. Toch zijn deze waarden minder spectaculair, als men zich realiseert dat ook een $m \times m$-matrix met random nabijheden door middel van niet-metrische MDS perfect (dus met *stress* = 0) kan worden afgebeeld in een $(m - 2)$-dimensionale ruimte. Het is dus niet zo verwonderlijk dat vijf procent van alle random (dus zevendimensionale) configuraties van negen punten redelijk tot goed in drie dimensies kunnen worden afgebeeld. In die gevallen is er kennelijk slechts weinig variatie tussen de punten op de vierde tot en met zevende dimensie.

De waarden die in Tabel 7.1 vermeld staan, zijn stresswaarden die bij een t-dimensionale oplossing van een $(m - 2)$-dimensionale configuratie een kans van .05 of minder hebben om voor te komen. Dat is iets anders dan de kans op een bepaalde stresswaarde voor een r-dimensionale oplossing van een r-dimensionale configuratie. Pas als we over zulk soort kansen beschikken, kunnen we zeggen bij welke stresswaarde het niet langer vol te houden is dat de 'echte' dimensionaliteit gelijk is aan de dimensionaliteit van de oplossing, zodat we over moeten gaan naar een oplossing met meer parameters (= dimensies). De vraag is overigens of deze statistische kwesties eenvoudig door computersimulatie zijn op te lossen. Een andere

aanpak is die van Ramsay (1969, 1977) waarin aannamen gemaakt worden over de statistische verdeling van de residuen $\{d_{ij} - f(o_{ij})\}$.
Een tweede probleem met het toepassen van Tabel 7.1 om significantietoetsen uit te voeren, blijkt als we de uitkomsten van de door Wagenaar en Padmos als voorbeeld gebruikte analyse vergelijken met de getallen uit de tabel. In het voorbeeld ging het om een nabijheidsmatrix van 12 objecten die, bij analyse in één, twee of drie dimensies, stresswaarden van respectievelijk .36, .14 en .08 had. Deze stresswaarden zijn alledrie kleiner dan de kritische waarden (.395, .205 en .10) uit de tabel. Wat moeten we nu concluderen? Is de echte dimensionaliteit één, twee of drie? En wat zouden we moeten doen als alle stresswaarden groter zouden zijn dan de kritische waarden? Het is daarom dus toch niet zo'n goed idee om de getallen uit Tabel 7.1 als significantietoets te gebruiken.

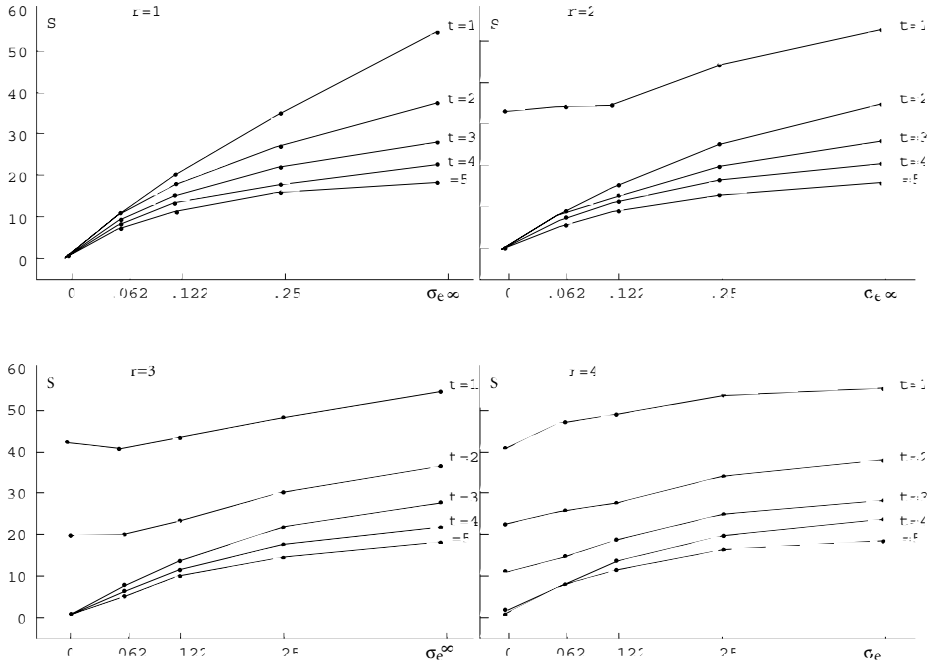
De methode van Spence en Graef

De aanpak van Spence en Graef (1974) is in grote lijnen identiek aan die uit het eerste deel van het onderzoek van Wagenaar en Padmos. De voornaamste verschillen bestaan uit het gebruik van andere aantallen objecten ($m = 12, 18, 26$ en 36), andere 'echte' dimensionaliteiten ($r = 1, 2, 3, 4$) en een andere manier om error aan de artificiële data toe te voegen. Dat laatste gebeurde als volgt: aan iedere door de onderzoekers gegenereerde coördinaat x_{is} van punt i op dimensie s werd elke keer dat object i met een ander object, zeg j, vergeleken werd een andere errorcomponent $\varepsilon_{is/j}$ toegevoegd. De Euclidische afstand met error tussen i en j is dan gelijk aan

$$d_{ij}^{(\varepsilon)} = \sqrt{\sum_{s=1}^{r} \left[(x_{is} + \varepsilon_{is/j}) - (x_{js} + \varepsilon_{js/i}) \right]^2} \qquad [7.2]$$

De errorcomponenten $\varepsilon_{is/j}$ werden aselect getrokken uit een normaalverdeling, met gemiddelde 0 en standaarddeviatie σ_ε, waarbij voor σ_ε de waarden 0.0, .0625, .1225 en .25 gekozen werden. In het onderzoek van Spence en Graef werden dus $4 \times 4 \times 4 = 64$ combinaties van m, r en σ_ε geconstrueerd. Voor elke combinatie werden vijf replicaties gegenereerd. De resulterende nabijheidsmatrices werden geanalyseerd in $t = 1, 2, 3, 4$ en 5 dimensies. De resultaten van dit onderzoek bestonden uit tabellen en grafieken waarin de relatie tussen de gemiddelde stresswaarde per replicatie en de waarden van m, r, σ_ε en t is weergegeven. Vier van zulke grafieken, voor $m = 36$, staan in Figuur 7.4. De vier grafieken (één voor elke ware dimensionaliteit r) bevatten elk vijf curven (voor de oplossingen in respectievelijk één, twee, drie, vier en vijf dimensies) die de relatie tussen stress en errorniveau weergeven. Helemaal rechts in elke grafiek staan de stresswaarden voor errorniveau $= \infty$, dat wil zeggen voor data die uit de getallen 1 tot en met $m(m-1)$ in random volgorde bestonden.

Spence en Graef suggereren een praktische toepassing van deze grafieken die iets anders in zijn werk gaat dan bij Wagenaar en Padmos. Stel dat een MDS-analyse van een 36 × 36-matrix in één tot en met vijf dimensies de stresswaarden .35, .20, .14, .10 en .09 heeft opgeleverd. Zet deze waarden af op de *Y*-as van de grafieken en trek vanuit deze punten lijnen evenwijdig aan de *X*-as. Neem vervolgens een liniaal of driehoek en verschuif die evenwijdig aan de *Y*-as naar rechts. Zoek nu die grafiek waarin de snijpunten van de horizontale lijnen en de liniaal zo dicht mogelijk bij de curven in de grafiek liggen. De betreffende grafiek geeft dan aan wat naar alle waarschijnlijkheid de onderliggende dimensionaliteit van de configuratie is (hier twee). De plaats van de liniaal op de *X*-as correspondeert dan met het betreffende errorniveau in de data (hier $\sigma_\varepsilon \approx .1225$).



Figuur 7.4 Stress als een functie van errorniveau, onderliggende dimensionaliteit (r) en dimensionaliteit van de oplossing (t) voor een configuratie van 36 punten

Discussie

Het belang van bovenstaande methoden is tweeërlei. In de eerste plaats toont zowel het onderzoek van Wagenaar en Padmos als dat van Spence en Graef aan hoe groot de effecten zijn van meetfouten in de data. Los van al het andere, laten beide studies zien hoe belangrijk het is om over

betrouwbare nabijheidsdata te beschikken. We kunnen deze studies opvatten als een pleidooi om uiterste zorgvuldigheid bij de dataverzameling te betrachten teneinde (meet)fouten zoveel mogelijk te voorkomen.
In de tweede plaats schetsen beide studies een procedure om door middel van Monte Carlo-simulatie stresswaarden te genereren waarmee men de stress van empirisch verzamelde nabijheidsdata kan vergelijken. In principe kan men in een bepaald praktijkgeval zelf een (groot) aantal artificiële data genereren en met MDS analyseren. Voorwaarde voor de zinvolheid hiervan is dat de methode waarmee error aan de data wordt toegevoegd overeenkomt met de manier waarop en feitelijke de orde van grootte waarin (meet)fouten ontstaan bij het verzamelen van empirische nabijheidsdata.

7.3 DE INHOUDELIJKE INTERPRETATIE VAN MDS-OPLOSSINGEN

Vier manieren van interpretatie

Zoals in de inleiding van dit hoofdstuk is gezegd, kunnen we vier manieren van aanpak onderscheiden die worden toegepast om een MDS-oplossing inhoudelijk te interpreteren. Deze vier manieren zijn respectievelijk:

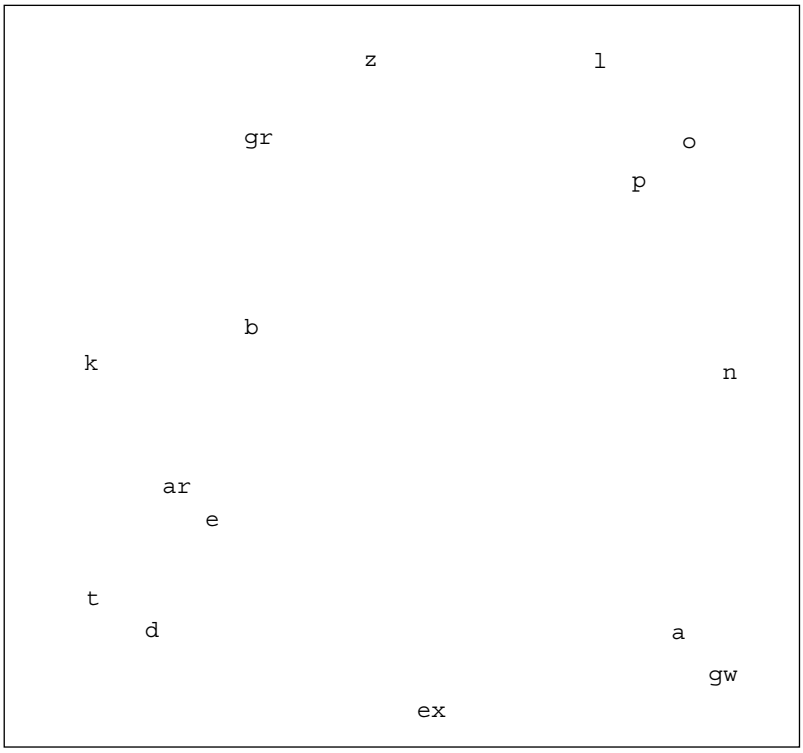
1. de intuïtieve aanpak;
2. additionele analyses op dezelfde nabijheidsdata;
3. interpretatie met behulp van additionele gegevens: externe analyse;
4. het vergelijken van de MDS-oplossingen met andere, hypothetische of in eerder onderzoek gevonden configuraties: Procrustes-analyse.

Coxon (1982) noemt de eerste twee typen *methods of internal interpretation* omdat hierin slechts de oorspronkelijke gegevens gebruikt worden. Dit is in tegenstelling tot *external interpretation* waarin additionele gegevens beschikbaar zijn. Bij interne interpretatiemethoden komt het vooral neer op de toepassing van *grafische* hulpmiddelen om beter inzicht in de configuratie te krijgen. Bovengenoemde manieren van interpreteren worden in de rest van dit hoofdstuk nader besproken.

De intuïtieve aanpak

Naast de stresswaarde en het Shepard-diagram, die indicaties van de *goodness-of-fit* zijn, produceren de meeste MDS-programma's niet alleen een matrix met coördinaten van de objecten op de dimensies, maar leveren ze ook één of meer 'plaatjes' waarin de objecten grafisch zijn afgebeeld. Bij een tweedimensionale oplossing is er natuurlijk maar één zo'n plaatje. Bij een driedimensionale oplossing worden meestal drie plaatjes afgedrukt, namelijk Dimensie 2 (verti-

caal) versus Dimensie 1 (horizontaal), Dimensie 3 versus Dimensie 1 en Dimensie 3 versus Dimensie 2. Bij een vierdimensionale oplossing zijn er $(4 \times 3)/2 = 6$ van zulke afbeeldingen mogelijk.
Het eenvoudigste geval is uiteraard het tweedimensionale plaatje. Maar zelfs dan doen zich enkele problemen voor. Als voorbeeld nemen we de configuratie die al eerder gepresenteerd is (zie Hoofdstuk 1, Figuur 1.2): de afbeelding van zestien persoonlijkheidsadjectieven uit het onderzoek van Van der Kloot en Van Herk (1991). Deze configuratie is hieronder opnieuw afgebeeld in Figuur 7.5. Bij een eerste, intuïtieve interpretatie van deze of een soortgelijke MDS-oplossing kan men twee wegen inslaan: proberen een inhoudelijke interpretatie van de dimensies te vinden, of trachten samenhangende groepjes (*clusters*) objecten aan te wijzen.



Figuur 7.5 De tweedimensionale configuratie van de persoonlijkheidsadjectieven aardig (a), gezellig (g), warm (w), extravert (ex), dominant (d), twistziek (t), eerzuchtig (e), arrogant (ar), koud (k), berekenend (b), gereserveerd (gr), zwijgzaam (z), lui (l), onderworpen (o), pretentieloos (p) en niet-sluw (n).

Dimensies. Om de dimensies te interpreteren moeten we kijken naar de ordening van de stimuli op de assen. Met name moeten we nagaan welke objecten extreme posities op de dimensies innemen. In de configuratie van Figuur 7.5 hebben koud, twistziek, dominant, arrogant en eerzuchtig tamelijk extreme, negatieve coördinaten op de horizontale dimensie,
terwijl aardig, gezellig, warm en niet-sluw de meest extreme positieve coördinaten op deze as hebben. Deze dimensie zou men dus heel goed als een *evaluatie*-dimensie kunnen benoemen: rechts liggen eigenschappen die over het algemeen 'goed' gevonden worden, links ligt een aantal 'slechte' eigenschappen. Op de verticale dimensie hebben vooral extravert en zwijgzaam extreme, tegenovergestelde posities. In de eerste dimensie lijkt het dus om de tegenstelling goed-slecht te gaan; de tweede dimensie geeft de tegenstelling introvert-extravert weer.
Deze interpretatie is overigens niet de enig mogelijke. In Hoofdstuk 5 hebben we gezien dat we assen van de figuur zonder verlies van informatie (dat wil zeggen: zonder de afstanden te veranderen!) mogen draaien. Als we dat doen, dan kunnen we een eerste dimensie vinden die, tussen koud en gereserveerd door, via berekenend naar warm loopt en een tweede as van dominant naar onderworpen. De eerste as zouden we dan sociabiliteit kunnen noemen, de tweede dominantie. We kunnen Figuur 7.5 dus op minstens twee verschillende manieren interpreteren. Dit toont aan, dat de rotatievrijheid bij de interpretatie voor problemen kan zorgen. Zowel goed-slecht en introvert-extravert als sociabiliteit en dominantie, zijn dimensies die vaker in dit soort onderzoek gevonden worden (zie Van der Kloot & Kroonenberg, 1982; Van der Kloot, Kroonenberg & Bakker, 1985).

Clusters. Een andere manier om een MDS-oplossing te interpreteren is door clusters, dat wil zeggen, homogene groepen objecten aan te wijzen. In Figuur 7.5 zijn minstens drie van zulke clusters aanwezig: (a) lui, onderworpen en pretentieloos, (b) aardig, gezellig en warm, en (c) arrogant, eerzuchtig, twistziek en dominant. Het is verleidelijk om ook zwijgzaam en gereserveerd samen te nemen en daarnaast berekenend en koud. Deze groepjes vormen beschrijvingen van een aantal prototypische personen. Ieder individu zou men dan kunnen opvatten als een mengeling van deze prototypen.

Drie- en meerdimensionale oplossingen. Is het al moeilijk om tweedimensionale afbeeldingen te interpreteren, bij drie- en meerdimensionale configuraties zijn de problemen nog veel groter, vooral als men alleen maar over afbeeldingen beschikt waarin de dimensies twee-aan-twee tegen elkaar zijn uitgezet. Gelukkig zijn er tegenwoordig computerprogramma's die het mogelijk maken driedimensionale grafieken te maken en in de ruimte te roteren. Dit zijn belangrijke hulpmiddelen om erachter te komen welke objecten op welke plaats in de ruimte liggen.

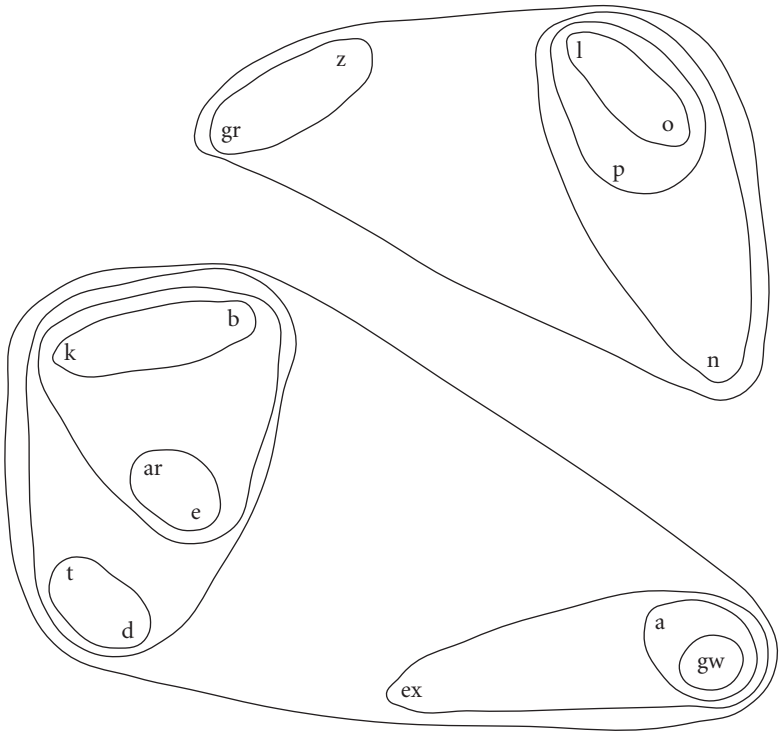
Additionele analyses

Een van de interpretaties die hierboven besproken is, bestond uit het zoeken naar groepjes samenhangende objecten. Dat gebeurde intuïtief, door met een timmermansoog de afstanden tussen de objecten in te schatten om uit te maken welke objecten wel en welke niet bij elkaar in een groepje thuishoren. In plaats van de eigen ogen kan men daarvoor ook een formele analysetechniek gebruiken, met name een van de vele vormen van *clusteranalyse* (zie Everitt, 1993). Dat kan op twee manieren: op de coördinaten van de punten in de oplossing en op de oorspronkelijke nabijheidsdata waarop de oplossing gebaseerd is (zie Kruskal & Wish, 1978).

Clusteranalyse op de coördinaten. Hierbij maken we gebruik van de afstanden tussen de punten in de oplossing. Deze methode is dus niet meer dan een hulpmiddel om de inhoud van de oplossing te verduidelijken en geeft geen nieuwe informatie over de *goodness-of-fit* van de afbeelding. Toch kan zo'n analyse heel verhelderend zijn, vooral als de configuratie veel punten bevat.

Er zijn verschillende soorten (cluster)analyses die men in deze situatie kan gebruiken. In de eerste plaats zijn dat allerlei soorten hiërarchische en niet-hiërarchische clusteranalyses. De resultaten kan men dan in de (twee)dimensionale grafieken afbeelden door de punten die tot hetzelfde cluster behoren te omcirkelen. In het hiërarchische geval ontstaat een afbeelding van elkaar omvattende 'cirkels'. Een voorbeeld is weergegeven in Figuur 7.6 waarin de resultaten van een hiërarchische clusteranalyse volgens de *complete linkage* methode[4] zijn ingetekend. Op het hoogste niveau zien we twee clusters: rechtsboven een cluster met adjectieven die weinig sociale activiteit uitdrukken en linksonder een cluster van eigenschappen die meer sociale activiteit inhouden.

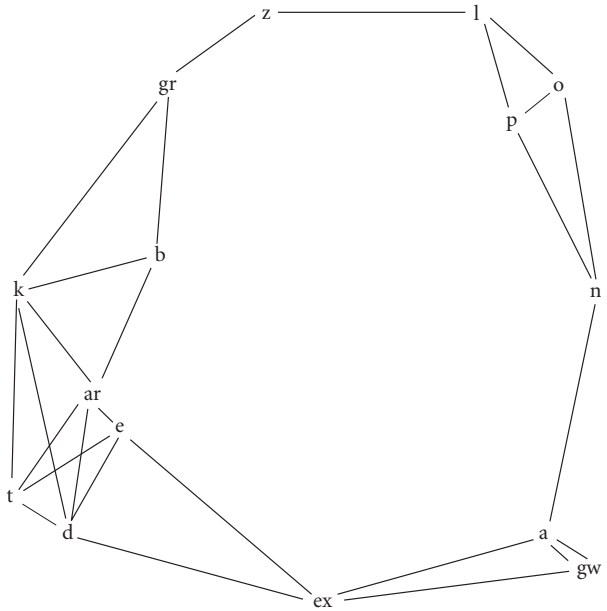
4 De complete linkage methode (ook wel de diametermethode of farthest neighbour methode genoemd) is een vorm van hiërarchische clusteranalyse waarin steeds die clusters bij elkaar worden gevoegd waarvoor geldt dat de twee objecten (uit elk cluster één) die het verst uit elkaar liggen een kleinere afstand hebben dan de verst uit elkaar liggende objecten uit twee andere clusters (zie Everitt, 1993; Meerling, 1988).



Figuur 7.6 De clusterstructuur van zestien persoonlijkheidsadjectieven, afgebeeld in de tweedimensionale MDS*-oplossing*

Een andere methode om clusters objecten te bepalen bestaat eruit dat alle punten met elkaar verbonden worden, die een onderlinge afstand hebben die kleiner is dan een bepaalde waarde (bijvoorbeeld kleiner dan de mediaan van de afstanden). Een voorbeeld is weergegeven in Figuur 7.7. Een vergelijkbare manier van clusteren vindt plaats door ieder punt zodanig met steeds één of twee andere punten te verbinden dat er een *minimal spanning tree* ontstaat, een structuur van paden die gezamenlijk de geringst mogelijke lengte hebben.

Clusteranalyse op de oorspronkelijke nabijheidsdata. Het verschil met de hierboven besproken aanpak is dat er nu clusters gevormd worden op basis van de oorspronkelijke nabijheidsdata, die *niet perfect* hoeven te corresponderen met de afstanden in de oplossingen. Het kan dus voorkomen dat punten die in de oplossing dicht bij elkaar liggen toch niet in één cluster terechtkomen (hun geobserveerde dissimilarity was dan kennelijk groter dan hun afstand in de oplossing). Ook is het mogelijk dat punten die in de oplossing ver van elkaar verwijderd zijn wél in één cluster worden samengevoegd (hun geobserveerde dissimilarity was dus kleiner dan hun afstand in de oplossing). Op deze manier



Figuur 7.7 De padstructuur van zestien persoonlijkheidsadjectieven

geeft de vorm en plaats van de clusters een extra indicatie van de kwaliteit van de oplossing, van de *goodness-of-fit.*
Ook voor de oorspronkelijke nabijheidsdata kunnen veel verschillende vormen van clusteranalyse gebruikt worden. Verbindt men alleen die punten met elkaar die een afstand hebben kleiner dan een bepaalde waarde, dan kan het gebeuren dat er een hoefijzervormige structuur ontstaat. In een deel van de ruimte lopen er dan als het ware geen paden tussen de punten. Dit kan een indicatie zijn dat de 'eigenlijke' structuur van de punten niet twee-, maar eendimensionaal is.

7.4 Externe analyse

Bij externe analyse beschikt men over additionele gegevens over de objecten, dat wil zeggen, over gegevens die niet gebruikt zijn voor het verkrijgen van de MDS-oplossing. Bijvoorbeeld: uit een onderzoek van Brokken (1978) is bekend welke waarden 1200 persoonlijkheidsadjectieven op de schalen *evaluatie*, *activiteit* en *potentie* van Osgoods *semantische differentiaal* hebben. De waarden op deze schalen van de adjectieven uit Figuur 7.5 staan vermeld in Tabel 7.2. Deze extra gegevens kunnen op verschillende manieren in de MDS-oplossing verwerkt worden: als contour, als richting en als punt. De externe gegevens worden als het ware in de MDS-configuratie geprojecteerd. Hieronder zullen we een aantal van die projectiemethoden (zie Heiser & De Leeuw, 1981) bespreken.

Tabel 7.2 Coördinaten[a] en scores[b] van zestien persoonlijkheidsadjectieven op evaluatie, activiteit en potentie

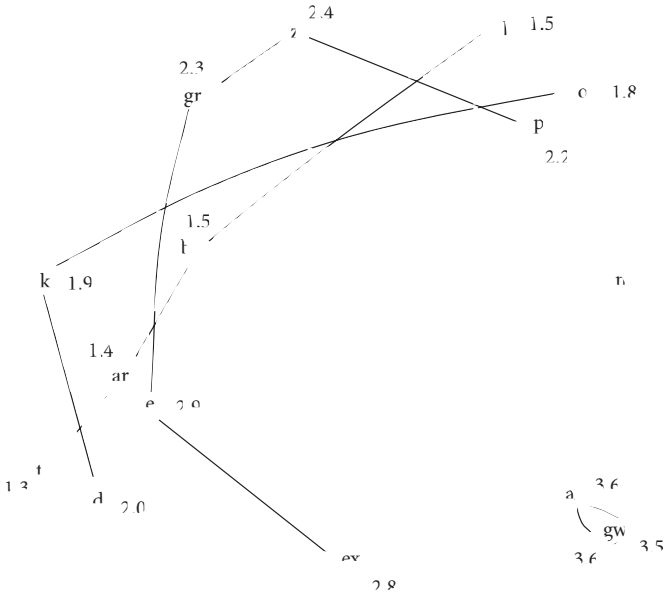
Adjectief[c]	Dimensie 1	Dimensie 2	Evaluatie	Activiteit	Potentie
aardig	.213	-.177	3.55	2.54	2.62
gezellig	.219	-.193	3.55	2.85	2.62
warm*	.219	-.197	3.52	2.85	2.86
extravert	.019	-.258	2.83	3.42	2.86
dominant	-.189	-.190	1.97	3.32	3.24
twistziek	-.250	-.124	1.27	3.38	2.81
eerzuchtig*	-.183	-.071	2.93	3.68	3.19
arrogant	-.196	-.036	1.39	2.80	3.14
koud*	-.251	.047	1.89	1.82	2.38
berekenend*	-.127	.071	1.48	3.03	3.24
gereserveerd	-.135	.204	2.30	1.76	2.19
zwijgzaam	-.047	.252	2.39	1.56	2.29
lui	.142	.238	1.45	1.10	2.10
onderworpen*	.172	.225	1.82	1.55	2.05
pretentieloos	.164	.159	2.24	1.70	2.14
niet-sluw**	.225	.045	–	–	-

[a] ontleend aan Van der Kloot en Van Herk (1991); [b] ontleend aan Brokken (1978); [c] de met een sterretje gemerkte adjectieven kwamen niet als zodanig in de lijst van Brokken voor. Zij werden vervangen door een synoniem: warm door warmvoelend, eerzuchtig door ambitieus, koud door koel, berekenend door geslepen en onderworpen door onderdanig. ** Voor niet-sluw kon geen synoniem worden gevonden.

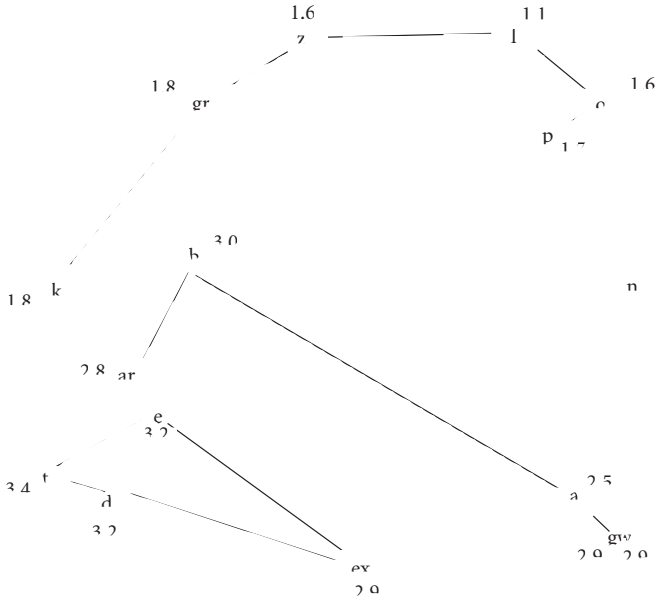
Gelijkheidscontouren

Eén manier om de additionele gegevens uit Tabel 7.2 te gebruiken is door bij elk van de adjectieven in Figuur 7.5 de bijbehorende waarde van een externe variabele, bijvoorbeeld evaluatie, te noteren. Vervolgens kunnen we punten met gelijke waarden met elkaar verbinden, zodat er een afbeelding met gelijkheidscontouren ontstaat. Schiffman, Reynolds en Young (1981) noemen dit *iso-attribute contours*, te vergelijken met hoogtelijnen, isobaren en isothermen op een landkaart (in ons voorbeeld: *iso-evaluatie*contouren). Een andere term die vaak voor dit soort lijnen gebruikt wordt, is het woord *isopreferentie*contouren. De achtergrond hiervan is dat de additionele gegevens die men wil afbeelden vaak uit preferenties bestaan, dat wil zeggen uit scores die aangeven in welke mate een object gewenst of geprefereerd wordt. Heiser en De Leeuw (1981) hebben voor dit soort lijnen de term *isochresten* voorgesteld: lijnen die objecten van gelijk nut met elkaar verbinden (Grieks: $\chi\rho\varepsilon o\varsigma$ = behoefte).

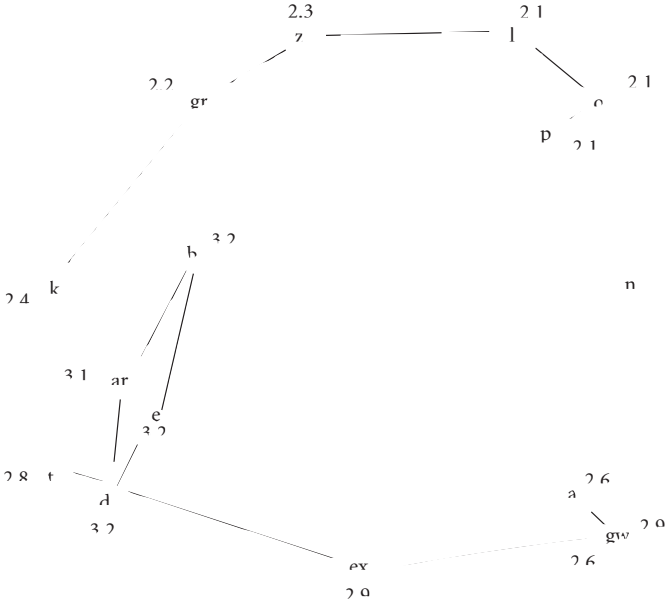
In de Figuren 7.8, 7.9 en 7.10 zijn de contouren voor evaluatie, activiteit en potentie in het plaatje van de zestien persoonlijkheidsadjectieven afgebeeld. Deze figuren laten zien wat we al vermoedden: naarmate adjectieven meer van linksboven naar rechtsonder in de figuur gaan, worden ze positiever gewaardeerd, al zien we in deze richting vooral het onderscheid tussen aardig, gezellig en warm (met evaluatiescores van 3.5) en de rest van de adjectieven (met lagere scores). In Figuur 7.9 zijn vijf contouren getekend die van rechtsboven naar linksonder hogere activiteitsscores hebben. In Figuur 7.10 zijn drie gebieden te onderscheiden die van boven naar beneden gaande weliswaar een (gemiddelde) toename in potentiescores laten zien, maar waarvan het gebiedje met de hoogste potentiescores eigenlijk tussen de andere twee in ligt (daarover meer in de volgende paragrafen).



Figuur 7.8 De iso-evaluatiecontouren van zestien persoonlijkheidsadjectieven



Figuur 7.9 De iso-activiteitscontouren van zestien persoonlijkheidsadjectieven



Figuur 7.10 De iso-potentiecontouren van zestien persoonlijkheidsadjectieven

Additionele gegevens als richting: het vectormodel

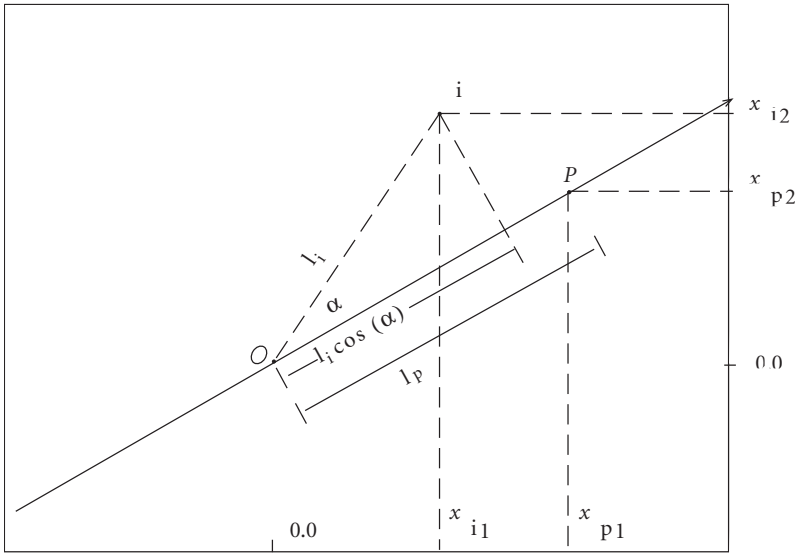
In overeenstemming met de intuïtieve interpretatie wijzen de gelijkheidscontouren erop dat de evaluatiescores van de adjectieven hoger worden naarmate de adjectieven meer naar rechts in de MDS-oplossing liggen. Met andere woorden, evaluatie correspondeert met een *richting* in de MDS-ruimte en hetzelfde geldt voor activiteit en potentie.

Hieronder zullen we een methode beschrijven waarmee we zo'n richting of *vector* exact kunnen bepalen. In die methode gaat het erom de coördinaten te berekenen van een punt dat samen met de oorsprong van de ruimte de gezochte richting vastlegt. Dat gaat als volgt. Stel dat we een *r*-dimensionale MDS-oplossing voor *m* objecten met coördinaten $\{x_{is}\}$ hebben en stel dat deze objecten scores $\{y_i\}$ hebben op een of andere externe variabele. Laten we aannemen dat deze externe variabele op intervalniveau gemeten is. Deze variabele beelden we af als een richting in de ruimte die door *O* en *P* gaat. Volgens het vectormodel moet nu gelden dat de *y*-scores na aftrek van een constante *c* (om het correcte nulpunt van *Y* te vinden) evenredig zijn aan de projecties van de objecten op de vector *OP* (zie Figuur 7.11 voor een tweedimensionaal voorbeeld). In dat geval is

$$y_i - c = k.l_i cos(\alpha) \qquad [7.3]$$

zodat

$$y_i = k.l_i cos(\alpha) + c \qquad [7.4]$$



Figuur 7.11 Externe analyse volgens het vectormodel

waarbij l_i de lengte is van de lijn uit de oorsprong naar punt *i*. Als we de lengte van de vector *OP* aanduiden als l_p kunnen we Formule [7.4] herschrijven als

$$y_i = \left(\frac{k}{l_p}\right) l_p l_i cos(\alpha) + c. \qquad [7.5]$$

Nu is bekend (zie Hoofdstuk 2) dat $l_p l_i \cos(\alpha) = \sum_s x_{ps} x_{is}$ zodat

$$y_i = \left(\frac{k}{l_p}\right) \sum_{s=1}^{r} x_{ps} x_{is} + c \qquad [7.6]$$

$$= \left(\frac{kx_{p1}}{l_p}\right) x_{i1} + \left(\frac{kx_{p2}}{l_p}\right) x_{i2} + ... + \left(\frac{kx_{pr}}{l_p}\right) x_{ir} + c.$$

In Formule [7.6] zijn van alle *m* objecten de scores y_i en de coördinaten $\{x_{is}\}$ bekend; wat onbekend is, zijn de coëfficiënten

$$\left(\frac{kx_{ps}}{l_p}\right)$$

en de constante *c*. Formule [7.6] heeft dus de vorm $Y = b_1X_1 + b_2X_2 + \ldots + b_rX_r + c$ van een multipele-regressieprobleem. *Y* komt in dit geval overeen met de scores op de externe variabele; X_1 tot en met X_r corresponderen met de coördinaten van de objecten op de *r* dimensies. Dit regressieprobleem kan op de standaardmanier worden opgelost. De regressiecoëfficiënten b_1 tot en met b_r zijn dan schattingen van de onbekende waarden

$$\left(\frac{kx_{p1}}{l_p}\right) \text{ tot en met } \left(\frac{kx_{pr}}{l_p}\right).$$

De multipele-correlatiecoëfficiënt $R_{Y.X1X2...Xr}$, met name het kwadraat ervan (kortweg R^2), geeft aan hoe goed de *Y*-scores op een lineaire manier uit de coördinaten van de objecten te voorspellen zijn en is dus een maat voor de *fit* van het vectormodel.

Merk op dat de regressiecoëfficiënten niet gelijk zijn aan de coördinaten van het punt *P*, maar er wel met een factor k/l_p mee evenredig zijn. Met andere woorden: de regressiecoëfficiënten kunnen we opvatten als de coördinaten van een ander punt (zeg: *P'*) dat echter wel op de gezochte lijn door *P* en *O* ligt. Immers, de richting van een lijn wordt bepaald door de verhouding van de coördinaten van de punten op die lijn. Voor *P'* is de verhouding van coördinaten op de eerste twee dimensies gelijk aan

$$\left(\frac{kx_{p1}}{l_p}\right) \div \left(\frac{kx_{p2}}{l_p}\right) = x_{p1} \div x_{p2}.$$

De lijn door *O* en *P'* is dus de gezochte richting van *Y*. Omdat alleen de richting van belang is, kunnen we deze lijn net zo kort of net zolang maken als we

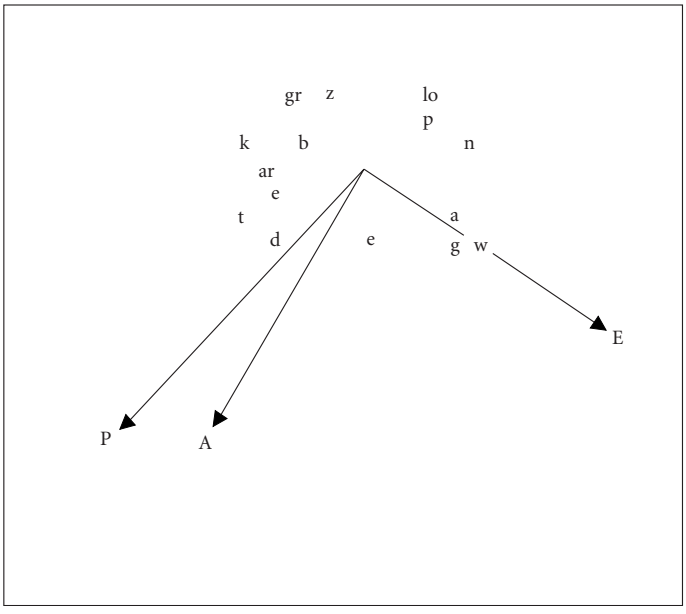
zelf willen. Het is nu gebruikelijk om de lengte (l) van de lijn zodanig te kiezen dat $l = R$. Daartoe moeten we de regressiecoëfficiënten delen door

$$\sqrt{\sum_{s=1}^{r} b_s^2} = \sqrt{\sum_{s=1}^{r} \left(\frac{kx_{ps}}{l_p}\right)^2}$$

(de wortel uit de som van hun gekwadrateerde waarden) en vermenigvuldigen met R. De lengte van de lijn is dan een visuele indicatie van de *fit*.
Voor de gegevens uit Tabel 7.2 zijn de regressiecoëfficiënten en multipele correlaties berekend met behulp van de SPSS-opdracht

```
REGRESSION VARIABLES=DIM1 DIM2 EVA ACTI POT
          /DEPENDENT EVA ACTI POT /ENTER DIM1 DIM2.
```

De regressiecoëfficiënten zijn 2.44 en -2.16 voor evaluatie (R = .752), -1.51 en -3.78 voor activiteit (R = .901) en -1.05 en -1.63 voor potentie (R = .818). De desbetreffende richtingen zijn in Figuur 7.12 afgebeeld. De lengten van de lijnstukken van de punten E (.56, -.50), A (-.33, -.84) en P (-.54, -.84) naar de oorsprong zijn gelijk aan de bijbehorende correlatiecoëfficiënten. We zien dus: evaluatie wijst naar rechtsonder, terwijl activiteit en potentie van rechtsboven naar linksonder gaan.



Figuur 7.12 Evaluatie, activiteit en potentie als vectoren in de ruimte van persoonlijkheidsadjectieven

Additionele gegevens als punt: het ideaalpuntmodel

De belangrijkste inhoudelijke eigenschap van het vectormodel is dat extreme scores op de additionele variabelen altijd aan de buitenkant van de configuratie moeten liggen. Naarmate de coördinaten van de objecten extremer worden (positief of negatief) worden hun projecties op een richting of vector noodzakelijkerwijs groter. In sommige gevallen kan dat een bezwaar zijn. We kunnen ons voorstellen dat het meest gewaardeerde of geprefereerde object niet aan de buitenkant van een configuratie ligt, maar meer in het midden. Bijvoorbeeld: als we alle politieke partijen van Nederland schalen, dan vinden we waarschijnlijk een dimensie van extreem links naar extreem rechts. De gemiddelde waardering (zoals die onder andere blijkt uit het aantal stemmen) voor partijen aan de uiteinden van deze dimensie is minder dan die voor partijen die meer in het politieke midden liggen. Het vectormodel is dan geen realistisch model voor de relatie tussen waardering en politieke positie. Een beter model is dan het zogenaamde *ideaalpunt-model*. De variabele 'waardering' wordt dan afgebeeld als een (ideaal)punt te midden van de objecten en wel zodanig dat de objecten die het meest gewaardeerd worden het dichtst bij dit ideaalpunt liggen. Met andere woorden: waardering is een functie van de afstand tussen het ideaalpunt en de objecten en dus zijn de (kwadraten van de) waarderingsscores ook een functie van de gekwadrateerde afstanden tussen ideaalpunt en objecten. Een aantal inhoudelijke aspecten van dit model komt in Hoofdstuk 11 nader aan de orde. Als we de variabele met additionele gegevens weer Y noemen[5] en het bijbehorende ideaalpunt P, dan kunnen we het ideaalpuntmodel als volgt formuleren:

$$y_i = kd^2_{pi} + c = k\sum_{s=1}^{r}(x_{ps} - x_{is})^2 + c. \qquad [7.7]$$

In het tweedimensionale geval vereenvoudigt [7.7] tot

$$y_i = k(x_{p1} - x_{i1})^2 + k(x_{p2} - x_{i2})^2 + c \qquad [7.8]$$

$$= kx^2_{p1} + kx^2_{i1} - 2kx_{p1}x_{i1} + kx^2_{p2} + kx^2_{i2} - 2kx_{p2}x_{i2} + c$$

zodat

$$y_i = k(x^2_{p1} + x^2_{p2}) + k(x^2_{i1} + x^2_{i2}) - 2kx_{p1}x_{i1} - 2kx_{p2}x_{i2} + c \qquad [7.9]$$

$$= (-2kx_{p1})x_{i1} + (-2kx_{p2})x_{i2} + k(x^2_{i1} + x^2_{i2}) + [k(x^2_{p1} + x^2_{p2}) + c].$$

Ook formule [7.9] heeft weer de vorm van een multipele-regressievergelijking waarin de scores $\{y_i\}$ voorspeld worden uit de scores op drie predictoren: $\{x_{i1}\}$,

5 Onder Y verstaan we een verzameling (al dan niet gekwadrateerde) additionele gegevens waarvan we verwachten dat die een lineair verband heeft met de gekwadrateerde afstanden tussen het ideaalpunt en de objecten.

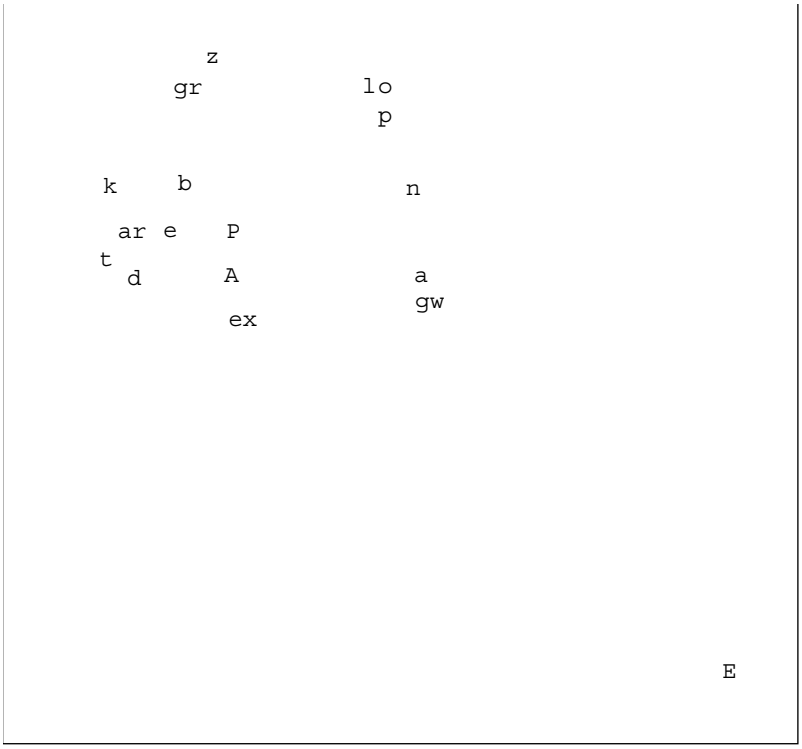
$\{x_{i2}\}$ en $\{x_{i1}^2 + x_{i2}^2\}$. De bijbehorende regressiecoëfficiënten b_1, b_2 en b_3 zijn dan schattingen van $-2kx_{p1}$, $-2kx_{p2}$ en k, zodat we de coördinaten van P kunnen vinden als $x_{p1} = -b_1/2b_3$ en $x_{p2} = -b_2/2b_3$. Merk op dat $b_3 = k$ een negatief getal moet zijn, want bij een kleine afstand is de waardering groot en bij een grote afstand is de waardering klein (NB: let op de manier waarop 'waardering' gemeten is! Drukt een groot getal inderdaad veel waardering uit of juist weinig?). Soms vindt men echter een positieve waarde voor k. In dat geval heeft men te maken met een *anti-ideaalpunt*. Dat wil zeggen: de waardering is het kleinst op de plaats van punt P en neemt toe naarmate men verder van punt P verwijderd raakt.
In Tabel 7.3 staan de resultaten van de regressie-analyses voor evaluatie, activiteit en potentie op de coördinaten van de zestien persoonlijkheidsadjectieven met behulp van de SPSS-opdrachten

```
COMPUTE DIM12 = DIM1**2 + DIM2**2.
REGRESSION VARIABLES=DIM1 DIM2 DIM12 EVA ACTI POT
   /DEPENDENT EVA ACTI POT/ENTER DIM1 DIM2 DIM12.
```

Tabel 7.3 Regressie-analyses voor de afbeelding van evaluatie, activiteit en potentie volgens het ideaalpuntmodel

	regressiecoëfficiënt			coördinaat		
eigenschap	b_1	b_2	$b_3 = k$	dim 1	dim 2	multipele correlatie R
evaluatie	2.55	-2.20	- 1.98	.64	-.56	.753
activiteit	- .80	-4.09	-13.77	-.03	-.15	.939
potentie	- .45	-1.89	-11.53	-.02	-.08	.918

De desbetreffende ideaalpunten zijn afgebeeld in Figuur 7.13. De ideaalpunten voor potentie en activiteit liggen tamelijk centraal tussen de eigenschappen. 'Evaluatie' ligt echter zeer ver naar buiten. Dat betekent dat 'evaluatie' net zo goed door middel van het vectormodel zou kunnen worden afgebeeld. Immers, de afstanden tussen het punt E en de objecten zijn bijna evenredig met de projecties van de objecten op de lijn door E en O. Dat deze conclusie correct is, blijkt uit de multipele-correlatiecoëfficiënten voor het vectormodel en het ideaalpuntmodel. Voor evaluatie zijn deze respectievelijk .752 en .753; de extra parameter van het ideaalpuntmodel levert voor evaluatie dus geen betere *fit* op. Voor activiteit en potentie geeft het ideaalpuntmodel (respectievelijk $R = .939$ en $R = .918$) wel een betere *fit* dan het vectormodel (respectievelijk $R = .901$ en $R = .818$).



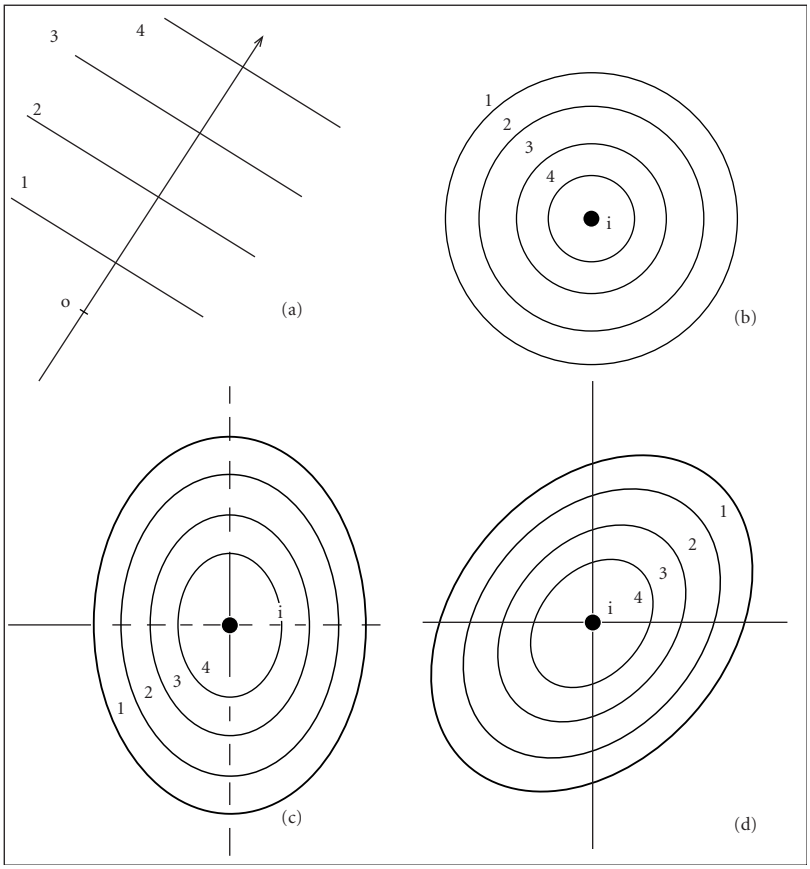
Figuur 7.13 De ideaalpunten van evaluatie (E), activiteit (A) en potentie (P) te midden van zestien persoonlijkheidsadjectieven

In de Formules [7.8] en [7.9] is het tweedimensionale geval uitgewerkt. In het driedimensionale geval is

$$y_i = k(x^2_{p1} + x^2_{p2} + x^2_{p3}) + k(x^2_{i1} + x^2_{i2} + x^2_{i3}) - 2kx_{p1}x_{i1} - 2kx_{p2}x_{i2} - 2kx_{p3}x_{i3} + c \qquad [7.10]$$

$$= (-2kx_{p1})x_{i1} + (-2kx_{p2})x_{i2} + (-2kx_{p3})x_{i3} + k(x^2_{i1} + x^2_{i2} + x^2_{i3}) + [k(x^2_{p1} + x^2_{p2} + x^2_{p3}) + c].$$

We hebben dan een regressievergelijking met vier predictoren $\{x_{i1}\}$, $\{x_{i2}\}$, $\{x_{i3}\}$ en $\{x_{i1}^2 + x_{i2}^2 + x_{i3}^2\}$ en vier coëfficiënten $b_1 = -2kx_{p1}$, $b_2 = -2kx_{p2}$, $b_3 = -2kx_{p3}$ en $b_4 = k$. Uitbreidingen naar vier en meer dimensies gaan analoog.



Figuur 7.14 Vier modellen voor de afbeelding van externe gegevens: (a) vectormodel, (b) ideaalpuntmodel, (c) gewogen ideaalpuntmodel, (d) IDIOSCAL-model

Het gewogen ideaalpuntmodel

Zowel het vectormodel als het ideaalpuntmodel kan beschreven worden in termen van gelijkheidscontouren. In het ideaalpuntmodel bestaan de isopreferentiecontouren uit concentrische cirkels rondom het ideaalpunt. In het vectormodel bestaan de contouren uit rechte lijnen die loodrecht staan op de richting van de vector (NB: zulke lijnen kan men opvatten als cirkels rond een punt in het oneindige. Vergelijk het punt voor evaluatie in Figuur 7.12 en 7.13). Beide modellen zijn afgebeeld in Figuur 7.14 (a en b). Op het ideaalpunt zijn twee varianten mogelijk: het *gewogen ideaalpuntmodel* en het *gewogen en geroteerde ideaalpuntmodel* (het zogenaamde *IDIOSCAL-model*).

In het gewogen ideaalpuntmodel bestaan de isopreferentiecontouren uit concentrische *ellipsen* rond het ideaalpunt, waarbij de hoofdassen van de ellipsen

samenvallen met de assen van de configuratie. Inhoudelijk betekent dit dat de afstanden tussen de objecten en het ideaalpunt op de ene dimensie niet in dezelfde mate voor de waardering meetellen als de afstanden op de andere dimensies. Ook in het IDIOSCAL-model zijn de isochresten concentrische ellipsen, maar hier zijn de hoofdassen van de ellipsen geroteerd ten opzichte van de dimensies van de configuratie. De isochresten die bij deze twee modellen horen, zijn afgebeeld in Figuur 7.14 (c en d).
Het gewogen ideaalpuntmodel kan als volgt geformuleerd worden:

$$y_i = kd^{*2}_{pi} + c = k\sum_{s=1}^{r} w_s(x_{ps} - x_{is})^2 + c. \qquad [7.11]$$

In het tweedimensionale geval geldt dus dat

$$\begin{aligned} y_i &= kw_1(x_{p1} - x_{i1})^2 + kw_2(x_{p2} - x_{i2})^2 + c \qquad [7.12] \\ &= kw_1x^2_{p1} + kw_1x^2_{i1} - 2kw_1x_{p1}x_{i1} + kw_2x^2_{p2} + kw_2x^2_{i2} - \\ &\quad 2kw_2x_{p2}x_{i2} + c \\ &= (-2kw_1x_{p1})x_{i1} + (-2kw_2x_{p2})x_{i2} + kw_1x^2_{i1} + kw_2x^2_{i2} + \\ &\quad [kw_1x^2_{p1} + kw_2x^2_{p2} + c]. \end{aligned}$$

Ook nu hebben we weer een multipele-regressieprobleem. In dit geval zijn er vier predictoren $\{x_{i1}\}$, $\{x_{i2}\}$, $\{x_{i1}{}^2\}$ en $\{x_{i2}{}^2\}$ en dus vier regressiecoëfficiënten $b_1 = -2kw_1x_{p1}$, $b_2 = -2kw_2x_{p2}$, $b_3 = kw_1$ en $b_4 = kw_2$, zodat $x_{p1} = -(b_1/2b_3)$ en $x_{p2} = -(b_2/2b_4)$.

Het IDIOSCAL-model
Het IDIOSCAL-model is op te vatten als ideaalpuntmodel waarin de assen van de configuratie geroteerd en gewogen worden. Dit model kan als volgt geformuleerd worden:

$$y_i = kd^{*2}_{pi} + c = k\sum_{s=1}^{r} w_s(x_{ps}{}^* - x_{is}{}^*)^2 + c \qquad (7.13)$$

waarin $x_{ps}{}^* = \sum_{a=1}^{r} t_{as}x_{pa}$ en $x_{is}{}^* = \sum_{a=1}^{r} t_{as}x_{ia}$ (zie Hoofdstuk 5).

In het tweedimensionale geval geldt dus dat

$$\begin{aligned} y_i &= (-2kw_1x_{p1}{}^*)x_{i1}{}^* + (-2kw_2x_{p2})x_{i2}{}^* + kw_1(x_{i1}{}^*)^2 + kw_2(x_{i2}{}^*)^2 \\ &\quad + [kw_1x^2_{p1} + kw_2x^2_{p2} + c]. \qquad [7.14] \end{aligned}$$

Werken we deze formule verder uit, dan krijgen we weer een multipele-regressievergelijking. In dit geval zijn er vijf predictoren $\{x_{i1}\}$, $\{x_{i2}\}$, $\{x_{i1}^2\}$, $\{x_{i2}^2\}$ en $\{x_{i1}x_{i2}\}$ en dus vijf regressiecoëfficiënten b_1 tot en met b_5 waaruit x_{p1} en x_{p2} zijn op te lossen.

De *fit* van externe-analysemodellen

Merk op dat we steeds ingewikkelder modellen met steeds meer *parameters* zijn gaan gebruiken. In het vectormodel voor twee dimensies moesten er twee onbekende parameters geschat worden, in het ideaalpuntmodel drie, in het gewogen ideaalpuntmodel vier en in het IDIOSCAL-model vijf. De modellen vormen een *hiërarchie*, dat wil zeggen: elk model met minder parameters is als het ware een bijzonder geval van een model met meer parameters. Het zal duidelijk zijn dat de *fit* van deze modellen toeneemt met het aantal parameters. Om uit te maken wat in een bepaald geval het beste model is, kan men, naast inhoudelijke overwegingen, ook de toename in *fit* bekijken. We kunnen daartoe de proportie-verklaarde variantie R^2 van het ene model vergelijken met dat van het andere. Is de R^2 van een ingewikkelder model niet (veel) groter dan die van een simpeler model, dan ligt het voor de hand om het eenvoudigere model te gebruiken.

Niet-metrische externe analyse

Hierboven is de zogenaamde *metrische* aanpak van externe-analyseproblemen behandeld. Daarbij werden de *Y*-scores opgevat als een lineaire functie van de objectcoördinaten en konden de onbekende parameters geschat worden met behulp van multipele regressie-analyse. Naast deze metrische aanpak is er ook een *niet-metrische* aanpak, waarbij het erom gaat de *rangorde* van de *Y*-waarden te voorspellen. Het verschil met de metrische aanpak is dat we niet de scores $\{y_i\}$ zelf willen voorspellen, maar een *monotoon stijgende transformatie* ervan, $f(y_i)$. Om dat voor elkaar te krijgen moeten we het volgende iteratieve algoritme toepassen.

Stap 1: kies één van de externe-analysemodellen. Bereken de regressiecoëfficiënten die bij de voorspelling van *Y* uit dit model horen en bereken R^2.

Stap 2: gebruik de regressiecoëfficiënten om $\hat{Y}$, de voorspelde waarde van *Y*, uit te rekenen.

Stap 3: gebruik Kruskals methode (zie Hoofdstuk 3) om de geobserveerde *Y*-scores zodanig monotoon te transformeren dat ze zoveel mogelijk met de voorspellingen $\hat{Y}$ overeenkomen. Dus: $f(y_i) \approx \hat{Y}_i$.

Stap 4: normaliseer de berekende $\{f(y_i)\}$ zodanig dat hun kwadratensom gelijk aan m (het aantal objecten) wordt, dat wil zeggen, deel $f(y_i)$ door $\sqrt{(\sum_i [f(y_i)]^2/m)}$. Deze genormaliseerde en getransformeerde scores noemen we $f'(y_i)$.

Stap 5: bereken de regressiecoëfficiënten die bij de voorspelling van $f'(Y)$ horen en bereken R^2.
Is de nieuwe R^2 groter dan de vorige, ga dan naar Stap 2. Is de nieuwe R^2 (nagenoeg) gelijk aan de vorige, dan is de optimale oplossing gevonden.

Bovenstaande stappen zijn onderdeel van verschillende computerprogramma's voor het afbeelden van preferenties en andere eigenschappen van de objecten. Bekende programma's zijn PREFMAP en PREFMAP-2 (Chang & Carroll, 1972), PREFMAP-3 (Meulman, Heiser, & Carroll, 1986) en PROFIT (Chang & Carroll, 1968). PREFMAP is een acroniem voor *preference mapping*; deze programma's kunnen metrische en niet-metrische externe analyses uitvoeren volgens het vectormodel, het ongewogen en gewogen ideaalpuntmodel en het IDIOSCAL-model. PROFIT – de naam is afgeleid van *property fitting* – is een programma voor metrische en niet-metrische externe analyse volgens het vectormodel. Merk op dat in het algemeen de *fit* van een niet-metrische analyse groter of op zijn minst even groot zal zijn als de metrische *fit* van hetzelfde model. Is de niet-metrische *fit* van een eenvoudiger model (nagenoeg) even groot als de metrische *fit* van een ingewikkelder model, dan ligt het voor de hand om het simpelere model te accepteren.

7.5 PROCRUSTESANALYSE

De vierde aanpak om een MDS-oplossing inhoudelijk te interpreteren bestaat uit het vergelijken ervan met een bekende configuratie, bijvoorbeeld een oplossing die in eerder onderzoek gevonden is, of een configuratie die *a priori* uit een hypothese of theorie is afgeleid. Een primitieve vorm daarvan zijn we tegengekomen in Hoofdstuk 1 waarin de configuratie van persoonlijkheidstrekken vergeleken werd met een cirkelvormige structuur die in eerder onderzoek was gevonden. In aanvulling op, of in plaats van zo'n intuïtieve aanpak kan men beter een meer formele methode toepassen waarin twee (of meer) matrices met coördinaten van dezelfde objecten met elkaar worden vergeleken.
Stel, we hebben twee matrices, **X** en **Y**, met de coördinaten in twee configuraties van m objecten op r dimensies. Een van de manieren waarop we de overeenstemming tussen beide matrices zouden kunnen nagaan is door twee correlatiecoëfficiënten te berekenen: (a) tussen de coördinaten van de objecten op de eerste dimensie van de ene en op de eerste dimensie van de andere configuratie en (b) tussen de coördinaten van de objecten op de tweede dimensie van de ene en de tweede dimensie van de andere configuratie. Het zal duidelijk zijn dat deze aanpak een groot aantal bezwaren heeft. Die berusten allemaal op het feit dat twee identieke configuraties, dat wil zeggen, twee configuraties waarin de objecten dezelfde afstandsrelaties tot elkaar hebben toch geheel verschillende matrices met coördinaten kunnen hebben. Immers, elk van de twee identieke configuraties kan op een aantal verschillende, afstandsbehoudende manieren getransformeerd zijn (zie Hoofdstuk 5).
Ten opzichte van de ene configuratie kan van de andere configuratie (a) het nulpunt verschoven zijn (translatie), (b) één of meer van de assen van teken gewisseld zijn (reflectie), (c) één of meer van de assen van volgorde gewisseld zijn (permutatie), en (d) kunnen de assen orthogonaal geroteerd zijn. Ten slotte kunnen alle coördinaten van de ene configuratie allemaal met een constante

factor vermenigvuldigd zijn (centrale dilatie). Al deze transformatiemogelijkheden maken het moeilijk de overeenkomst tussen twee configuraties goed in te schatten, zelfs als ze identiek zijn. Bij het berekenen van correlaties tussen corresponderende assen heeft men weliswaar geen last van translatie en centrale dilatie, maar door reflectie verandert het teken van de correlatiecoëfficiënt en door permutatie en rotatie kunnen de correlatiecoëfficiënten er totaal anders uit gaan zien. Deze problematiek wordt nog vele malen groter, als er in plaats van twee configuraties **X** en **Y** drie of meer matrices met elkaar vergeleken moeten worden.

Een manier om uit te maken of twee configuraties via afstandsbehoudende transformaties tot elkaar herleid kunnen worden, is door binnen elke configuratie de afstanden tussen de objecten te berekenen en deze beide verzamelingen afstanden met elkaar te correleren. We berekenen dus de correlatiecoëfficiënt $r_{dij^{(X)},dij^{(Y)}}$ over de $m(m-1)/2$ objectparen $\{i,j\}$. Twee identieke configuraties hebben, ook al zijn de assen geroteerd, gespiegeld, gepermuteerd verschoven en uniform gedilateerd, een correlatie van 1.0. Is de correlatiecoëfficiënt (veel) lager, dan moet men concluderen dat de twee configuraties niet door middel van afstandsbehoudende transformaties uit elkaar kunnen worden afgeleid.

Als de afstanden binnen twee configuraties een hoge correlatie met elkaar hebben, dan weten we weliswaar dat de ene configuratie een afstandsbehoudende transformatie van de andere is, maar we weten niet hoe de desbetreffende transformatiematrices eruitzien. Om daar achter te komen moeten we een methode toepassen die onder de naam (*gegeneraliseerde*) *Procrustesanalyse* bekendstaat.

Bij deze methode om twee of meer configuraties met elkaar te vergelijken maakt men gebruik van de coördinaten op alle dimensies tegelijkertijd. Daarbij is het de bedoeling één of meer transformatiematrices te ontdekken die ervoor zorgen dat de ene configuratie na transformatie zoveel mogelijk lijkt op de andere configuratie. Het oplossen van dit probleem wordt ook wel *matching* genoemd. Bij Procrustesanalyse worden meestal drie hypothesen of modellen onderzocht:

1 De ene configuratie (**X**) is tot de andere configuratie (**Y**) te herleiden door middel van de afstandsbehoudende transformaties translatie, centrale dilatie, spiegeling, permutatie en orthogonale rotatie. We gaan dan uit van het model $\mathbf{X} \approx k\mathbf{YRPT} + c\mathbf{I}$ en proberen een transformatiematrix $\mathbf{G} = \mathbf{YRPT}$, een dilatiefactor k en een translatieparameter c te vinden.

2 De ene configuratie is uit de andere af te leiden door middel van de afstandsbehoudende transformaties translatie, centrale dilatie, spiegeling en permutatie enerzijds (dus geen rotatie!) en een niet-afstandsbehoudende weging van de dimensies anderzijds (zie ook het gewogen-afstandsmodel of INDSCAL-model uit Hoofdstuk 4 en het gewogen-ideaalpuntmodel uit dit hoofdstuk). Het model is dan $\mathbf{X} \approx k\mathbf{YWRP} + c\mathbf{I}$, waarin **W** een diagonale matrix met gewichten is.

3 De ene configuratie is tot de andere herleidbaar door middel van translatie, centrale dilatie, reflectie en permutatie enerzijds en de niet-afstandsbehoudende

transformatie van rotatie en weging van de geroteerde dimensies (zie ook het IDIOSCAL-model hierboven). Het model is dan $\mathbf{X} \approx k\mathbf{YRPTW} + c\mathbf{I}$.
Het voert te ver om de oplossingen van deze problemen hier te behandelen. Voor een meer technische behandeling wordt de lezer verwezen naar Borg en Lingoes (1987) en Commandeur (1991). Een meer inhoudelijke bespreking van de problematiek is te vinden bij Coxon (1982) en bij Schiffman, Reynolds en Young (1981). Computerprogramma's voor Procrustesanalyse zijn onder andere PINDIS van Borg en Lingoes (1977) en Lingoes en Borg (1976, 1977, 1978), MATCHALS van Commandeur (1991) en een procedure van Gower (1975).