HOOFDSTUK 11

Preferentierangordeningen

11.1 Preferentierangordeningen

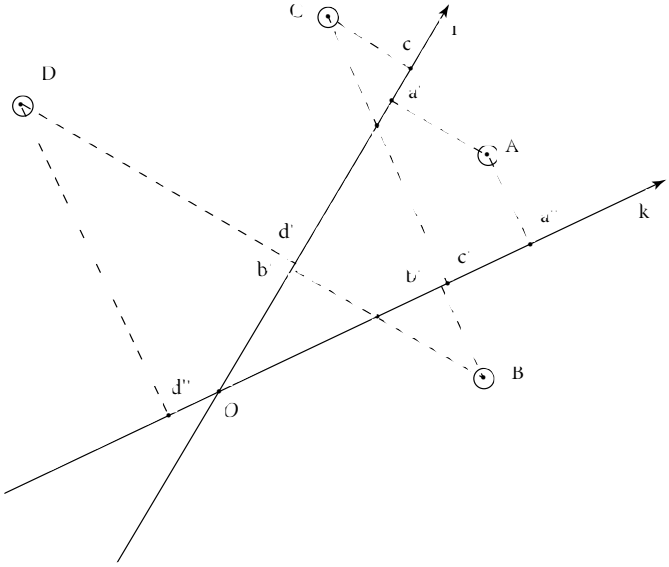
In het vorige hoofdstuk hebben we enkele analysemethoden behandeld voor zogenaamde keuzedata, dat wil zeggen, preferentiedata die uit enen en nullen bestaan. We hebben daarbij twee modellen besproken: het ideaalpuntmodel en het latente-trekmodel. Ook in dit hoofdstuk gaat het om de analyse van preferentiedata, maar nu om gegevens die uit preferentie*rangordeningen* bestaan. Deze gegevens komen tot stand met een zogenaamde *order-k-of-m*-opdracht, waarbij m het aantal objecten is en k kan variëren van 1 (waardoor we in feite weer keuzedata krijgen) tot $m \times 1$ (waardoor een complete rangordening van de m objecten ontstaat). Merk op dat we zulke data op twee manieren kunnen coderen. De meest gebruikelijke is om het meest geprefereerde object het rangnummer 1 toe te kennen. Het object dat op de tweede plaats komt krijgt het getal 2, enzovoort. Het minst geprefereerde object krijgt het rangnummer m. Dus: hoe groter het getal, des te minder de preferentie. Omgekeerd kunnen we het hoogste rangnummer ook aan het meest geprefereerde object toekennen, en het getal 1 aan het minst geprefereerde. In dat geval bestaat er dus een positieve relatie tussen rangnummer en preferentie.
De eerste manier van coderen is vooral handig als we het ideaalpuntmodel gebruiken. Er bestaat dan een positief verband tussen de rangnummers van de objecten en hun afstanden tot het ideaalpunt. De tweede manier is handiger als we van het zogenaamde vectormodel uitgaan (zie Hoofdstuk 7). Er bestaat dan een positief verband tussen de rangnummers van de objecten en hun projecties op bepaalde vectoren (zie later in dit hoofdstuk).
Zoals bij de meeste meerdimensionale schaalproblemen is het mogelijk preferentierangordeningen metrisch en niet-metrisch te analyseren. Niet-metrische analyse ligt voor de hand als de gegevens inderdaad uit rangordeningen

bestaan en de preferenties dus op ordinaal niveau gemeten zijn. Metrische analyse is vooral van toepassing als we ervan uitgaan dat de data een meer dan ordinale betekenis hebben, met name als ze op interval- of rationiveau gemeten zijn. In het laatste geval nemen we aan dat de preferentiedata van een proefpersoon zonder meer evenredig zijn met de onderliggende utiliteit van de objecten. Bij preferentie-op-intervalniveau gaan we er daarentegen van uit dat de waargenomen preferenties een lineaire functie zijn van de utiliteiten. Metrische analyse ligt dus meer voor de hand als de data niet uit rangordeningen bestaan, maar uit andere getallen die de mate van voorkeur van een verzameling personen voor een verzameling objecten weergeven, bijvoorbeeld, hoeveel geld iemand voor de objecten overheeft, hoeveel tijd hij of zij eraan wil besteden of hoe vaak de betreffende objecten gekozen worden. Dit soort gegevens kunnen metrisch geanalyseerd worden op de manier die hieronder beschreven wordt. Met een beetje fantasie kunnen we die gegevens soms opvatten als *frequentiegegevens* (aantallen guldens, aantallen uren en aantallen keuzen). Daardoor kunnen ze ook geanalyseerd worden met een methode die speciaal voor frequentiedata ontworpen is: *correspondentieanalyse*. Deze methode kan overigens altijd gebruikt worden voor data die – zoals frequenties – uit positieve getallen bestaan. Verderop zullen we ook deze methode nader bespreken.

11.2 METRISCHE ANALYSE MET HET VECTORMODEL

Bij de analyse van preferentierangordeningen kan gebruikgemaakt worden van het *vectormodel*, dat al eerder besproken is in Hoofdstuk 7. In het vectormodel worden de rijen en kolommen van een rechthoekige matrix (bijvoorbeeld personen en objecten) samen in één ruimte afgebeeld, waarbij de objecten meestal door middel van punten worden weergegeven en de personen door middel van *vectoren*, dat wil zeggen, lijnen door de oorsprong die een bepaalde richting hebben. Als wordt aangenomen dat de data op rationiveau gemeten zijn, is volgens het vectormodel de preferentie o_{ij} van Persoon i voor Object j *evenredig* aan de *projectie* van het lijnstuk tussen de oorsprong en punt j, op de rechte lijn door de oorsprong die bij Persoon i hoort. Twee van dat soort lijnen, voor Persoon i en Persoon k, zijn in Figuur 11.1 afgebeeld, samen met de projecties van vier objecten op die twee vectoren. Merk op dat Object D op beide vectoren een projectie heeft die links van de oorsprong ligt. Omdat alle twee de vectoren naar rechts wijzen, heeft Object D dus *negatieve* projecties op die vectoren[1].

1 Als de data uit alleen maar positieve getallen bestaan, moet de oorsprong in principe zodanig gekozen worden dat alle projecties positief zijn.



Figuur 11.1 Preferentierangordeningen CADB (Persoon i*) en ACBD (Persoon* k*) volgens het vectormodel*

Algebraïsch komt het model op het volgende neer. Stel, y_{js} is de coördinaat van Object *j* op dimensie *s* en h_{is} is de coördinaat van het eindpunt van de vector voor Persoon *i*. Als we deze vector zodanig kiezen dat hij een lengte van 1.0 heeft, is de *projectie* van het lijnstuk tussen oorsprong en punt *j* op de vector van *i* gelijk aan $p_{ij} = \sum_s h_{is} y_{js}$. Voor iedere rij *i* geldt volgens het model dat de observaties o_{ij} evenredig zijn met p_{ij}, dat wil zeggen, dat $o_{ij} = c_i p_{ij}$ en dus dat

$$o_{ij} = c_i \Sigma_s h_{is} y_{js}. \qquad [11.1]$$

Formule [11.1] kunnen we schrijven als

$$o_{ij} = \Sigma_s c_i h_{is} y_{js} = \Sigma_s x_{is} y_{is} \qquad [11.2]$$

waarin $x_{is} = c_i h_{is}$. De getallen x_{is} zijn de coördinaten op de dimensies $s = 1,...,r$, van een punt dat op de vector van Persoon *i* ligt. Aangezien we weten dat deze vector ook door de oorsprong gaat, is de richting van de gezochte vector vastgelegd met het bepalen van de coördinaten x_{is}. De evenredigheidsfactor c_i is alleen maar van belang voor de lengte van vector *i* en niet voor de richting[2].

2 Zolang we het getal c_i niet precies kennen, is de lengte van de vector arbitrair. Het eenvoudigst is dus om alle vectoren dezelfde standaardlengte (bijv. 1.00) te geven. Maar we zouden lengtes kunnen kiezen die extra informatie weergeven. Het ligt dan voor de hand om de lengte van een vector gelijk te maken aan de *fit* van het model voor de desbetreffende persoon, bijvoorbeeld de correlatiecoëfficiënt van de observaties van Persoon *i* en de benaderingen daarvan volgens $\Sigma_s x_{is} y_{js}$.

In matrixnotatie is Formule [11.2] gelijk aan

$$o_{ij} = \boldsymbol{x}_i' \boldsymbol{y}_j \qquad [11.3)]$$

zodat voor de complete matrix met preferentierangordeningen geldt dat

$$O = XY'. \qquad [11.4]$$

Hiervan is alleen de matrix **O** bekend; bij de analyse gaat het dus om het schatten van de matrices **X** en **Y**. In het metrische geval gebeurt dat met behulp van *singuliere-waardendecompositie* (zie hieronder).

Singuliere-waardendecompositie (SVD)
In Hoofdstuk 2 zijn we al tegengekomen dat iedere willekeurige rechthoekige matrix ontbonden kan worden in drie andere matrices, dat wil zeggen: iedere rechthoekige matrix **M** is te schrijven als het product $\mathbf{U\Lambda V'}$. Daarin zijn **U** en **V** respectievelijk de linker en rechter eigenvectoren van **M** en is Λ een diagonale matrix met de singuliere waarden van **M**. Passen we deze ontbinding toe op **O** dan is

$$O = XY' = U\Lambda V' \qquad [11.5]$$

zodat we kunnen definiëren[3] $\mathbf{X = U}$ en $\mathbf{Y = \Lambda V'}$ waarbij $\mathbf{X'X = XX' = I}$. Voor het berekenen van de desbetreffende matrices uit deze ontbinding bestaan er standaardprogramma's, bijvoorbeeld in de routine `MATRIX - END MATRIX` van SPSS.

Preferenties op intervalniveau
Soms (meestal?) is het niet erg realistisch om aan te nemen dat iemands preferentiedata zonder meer evenredig zijn met de projecties van punten op een vector. Een aanname die minder veeleisend is, is dat de preferentiedata een lineaire functie van de projecties zijn, dus dat

$$o_{ij} = a_i + c_i p_{ij}. \qquad [11.6]$$

De eenvoudigste manier om voor dit geval een oplossing te vinden is door van de rijen van matrix **O** de bijbehorende rijgemiddelden af te trekken. We krijgen dan

$$\begin{aligned} o^*_{ij} &= a_i + c_i p_{ij} - (a_i - c_i \Sigma_j p_{ij}/m) \qquad [11.7] \\ &= c_i p_{ij} - c_i \Sigma_j p_{ij}/m. \end{aligned}$$

3 We kunnen ook afspreken $\mathbf{X = U\Lambda^{1/2}}$ en $\mathbf{Y = V\Lambda^{1/2}}$ of $\mathbf{X = U\Lambda}$ en $\mathbf{Y = V}$.

Substitutie van Formule [11.2] geeft

$$\begin{aligned} o^*_{ij} &= c_i \Sigma_s x_{is} y_{js} - c_i \Sigma_j \Sigma_s x_{is} y_{js}/m \\ &= c_i \Sigma_s x_{is} (y_{js} - \Sigma_j y_{js})/m \\ &= c_i \Sigma_s x_{is} y^*_{js} \end{aligned} \qquad [11.8]$$

waarbij y^*_{js} de coördinaten van de objecten zijn, in afwijking van hun gemiddelde op elke dimensie. Met andere woorden, de observaties in afwijking van hun rijgemiddelden zijn volgens het model evenredig met projecties van punten op vectoren in een ruimte waarvan de oorsprong in de centroïde van de objectpunten ligt.
Het vectormodel voor preferentiedata en de bijbehorende analysemethoden zijn oorspronkelijk geformuleerd door Slater (1960) en Tucker (1960). Een gedetailleerde bespreking is te vinden in Heiser en De Leeuw (1981). Een programma dat speciaal gericht is op de metrische analyse van preferentiedata is MDPREF (Carroll & Chang, 1973). Dit programma heeft vele mogelijkheden om de observaties voorafgaand aan de analyse te transformeren. Een van die opties is het aftrekken van de rijgemiddelden. Voorbeelden van twee metrische analyses worden besproken in Blok 11.1.

BLOK 11.1 METRISCHE ANALYSE VAN PREFERENTIEDATA

In Hoofdstuk 1 is een onderzoek van Van der Kloot en Willemsen (1991) beschreven waarin 40 proefpersonen zes bijdragen aan wetenschappelijk onderzoek paarsgewijs met elkaar moesten vergelijken. Op grond van die vergelijkingen kan voor iedere proefpersoon een dominantiematrix gemaakt worden die weergeeft welke rijobjecten welke kolomobjecten qua belangrijkheid domineren. Voor één proefpersoon (Persoon 35) is zo'n tabel hieronder weergegeven. De rijen en kolommen hebben betrekking op de bijdragen bedenken (BE), leidinggeven (GE), dataverzameling (DV), data-analyse (DA), schrijven (SC) en op de auteurschapsgrens (AG). De rijtotalen van deze matrix geven aan hoe vaak de betreffende rijobjecten belangrijker gevonden werden dan de kolomobjecten; we kunnen deze getallen dus opvatten als belangrijkheidsscores. Omgezet in rangordecijfers vormen deze gegevens een preferentierangordening die de volgorde van belangrijkheid van de bijdragen aangeeft.

	BE	LE	DV	DA	SC	AG		rang
BE	–	1	0	0	0	0	1	2
LE	0	–	0	0	0	0	0	1
DV	1	1	–	0	0	0	2	3
DA	1	1	1	–	0	1	4	5
SC	1	1	1	1	–	1	5	6
AG	1	1	1	0	0	–	3	4

Voor alle proefpersonen uit het onderzoek zijn overeenkomstige rijtotalen en rangordecijfers berekend. De rijtotalen staan in Tabel 11.1.

Tabel 11.1 Belangrijkheidsscores van de auteurschapsgrens en van vijf soorten bijdragen aan een wetenschappelijke publicatie

Persoon	BE	LE	DV	DA	SC	AG
1	2.50[a]	1.50	2.50	2.50	5.00	1.00
2	3.00	2.00	1.50	.50	4.00	4.00
3	3.00	2.00	1.00	3.00	5.00	1.00
4	4.00	3.00	2.00	1.00	5.00	.00
5	2.50	2.00	1.00	.50	5.00	4.00
6	2.50	4.00	.50	2.00	4.00	2.00
7	4.00	1.00	.00	3.00	5.00	2.00
8	4.00	1.50	1.50	.50	4.50	3.00
9	3.50	3.50	.50	1.50	5.00	1.00
10	3.00	3.50	.50	.50	5.00	2.50
11	4.00	3.00	.50	1.50	5.00	1.00
12	.00	4.00	2.00	3.00	5.00	1.00
13	4.50	3.00	1.00	2.50	4.00	.00
14	1.50	1.50	1.50	3.50	5.00	2.00
15	3.00	3.00	.00	2.00	5.00	2.00
16	4.00	1.50	.50	2.00	5.00	2.00
17	3.00	3.00	3.00	2.00	4.00	.00
18	2.00	1.00	.50	2.50	5.00	4.00
19	4.00	4.00	1.00	.00	4.00	2.00
20	3.00	3.00	1.50	.50	5.00	2.00
21	4.50	2.00	.00	2.00	4.50	2.00
22	3.50	4.50	.00	2.50	3.50	1.00
23	1.50	1.50	3.50	1.50	5.00	2.00
24	3.50	2.50	1.00	2.50	4.50	1.00

25	2.00	3.00	.00	4.00	5.00	1.00
26	4.00	3.50	.50	.50	4.50	2.00
27	3.00	4.00	1.50	1.00	4.50	1.00
28	3.50	5.00	1.00	2.00	3.50	.00
29	3.00	4.00	1.50	1.50	5.00	.00
30	4.00	3.50	.00	1.00	4.50	2.00
31	3.00	1.00	2.50	2.50	5.00	1.00
32	4.00	1.00	2.00	3.00	5.00	.00
33	2.00	1.00	4.00	3.00	5.00	.00
34	3.00	4.00	.50	.50	5.00	2.00
35	1.00	.00	2.00	4.00	5.00	3.00
36	3.00	3.00	.00	2.00	5.00	2.00
37	3.50	3.50	.50	.50	4.00	3.00
38	1.00	3.50	.50	3.50	4.50	2.00
39	.00	3.00	3.00	3.00	5.00	1.00
40	4.00	1.00	.00	3.00	5.00	2.00

[a] De getallen eindigend op .50 zijn het gevolg van het feit dat de desbetreffende proefpersoon bij sommige vergelijkingen geen keuze had gemaakt. In dat geval is aan beide objecten van het paar de waarde .50 toegekend.

Als we de data van Tabel 11.1 in de matrix **O** plaatsen, dan geeft de opdracht `CALL SVD(O,U,Labda,V)` uit de SPSS-procedure `MATRIX - END MATRIX` de uitkomsten die hieronder worden weergegeven.

U						LABDA					
-.153	-.185	.031	.076	.125	-.019	42.644	.000	.000	.000	.000	.000
-.146	.070	-.277	-.137	.200	.331	.000	9.704	.000	.000	.000	.000
-.161	-.099	.022	.114	-.108	-.076	.000	.000	8.110	.000	.000	.000
-.165	.060	.170	.159	.213	-.160	.000	.000	.000	6.727	.000	.000
-.155	.046	-.298	-.193	.144	-.070	.000	.000	.000	.000	6.207	.000
-.155	.087	.051	-.169	-.127	.205	.000	.000	.000	.000	.000	2.494
-.163	-.048	-.184	.238	-.189	-.041						
-.154	.073	-.235	.070	.233	.122	V					
-.168	.108	.078	.024	-.037	-.216	-.442	.446	-.101	.710	.111	.283
-.164	.154	-.075	-.150	.059	-.202	-.393	.469	.554	-.516	-.127	.189
-.168	.107	.038	.115	-.018	-.197	-.168	-.447	.245	-.059	.750	.382
-.152	-.186	.226	-.364	-.082	-.112	-.287	-.571	.135	.144	-.634	.388
-.160	.065	.171	.259	-.065	.209	-.695	-.228	-.101	-.050	.078	-.668
-.151	-.237	-.065	-.066	-.114	.021	-.238	.067	-.771	-.451	.011	.376
-.165	.061	-.051	-.042	-.146	-.158						
-.163	.012	-.151	.174	-.037	-.082						

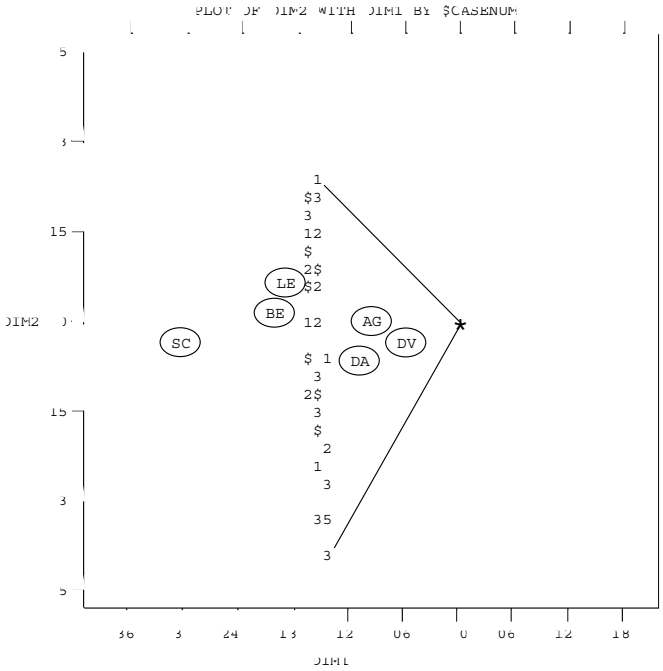
```
-.149   -.067   .242   .073   .201   .267
-.153   -.120  -.342  -.122  -.110   .033
-.159    .251   .014  -.057   .165   .140
-.161    .081  -.031  -.087   .189  -.162
-.163    .094  -.132   .197  -.105   .070
-.157    .156   .167  -.016  -.239   .341
-.146   -.212  -.038  -.126   .332   .016
-.159   -.010   .048   .122  -.065   .075    Y = ΛV'
-.162   -.109   .090  -.038  -.370  -.111 -18.837  4.328   -.816  4.774   .689   .705
-.164    .208  -.033  -.008   .070  -.030 -16.776  4.549  4.495 -3.474  -.786   .471
-.160    .104   .147  -.083   .110  -.026  -7.160 -4.334  1.984  -.396  4.658   .953
-.157    .156   .318  -.007  -.079   .303 -12.238 -5.537  1.098   .966 -3.938   .968
-.165    .056   .244  -.009   .063  -.232 -29.640 -2.214  -.821  -.334   .484 -1.666
-.165    .202  -.040   .007  -.042  -.029 -10.132   .648 -6.252 -3.033   .069   .938
-.154   -.187  -.009   .167   .145   .000
-.160   -.154   .067   .355   .049  -.036
-.147   -.338   .152   .126   .255   .043
-.166    .175   .007  -.155   .048  -.240
-.143   -.378  -.233  -.065  -.081   .156
-.165    .061  -.051  -.042  -.146  -.158
-.156    .204  -.116  -.124   .056   .198
-.153   -.106   .054  -.260  -.291   .097
-.147   -.280   .188  -.296   .059  -.034
-.163   -.048  -.184   .238  -.189  -.041
```

De eerste twee kolommen van **U** (de personen) en **Y** (de bijdragen) zijn afgebeeld in Figuur 11.2. Daartoe zijn eerst de coördinaten van de bijdragen door 100 gedeeld om de afbeelding qua schaal overzichtelijker te maken. Dat maakt immers niets uit voor het vinden van de richtingen van de personen en de (volgorde van de) projecties van de bijdragen op die richtingen.

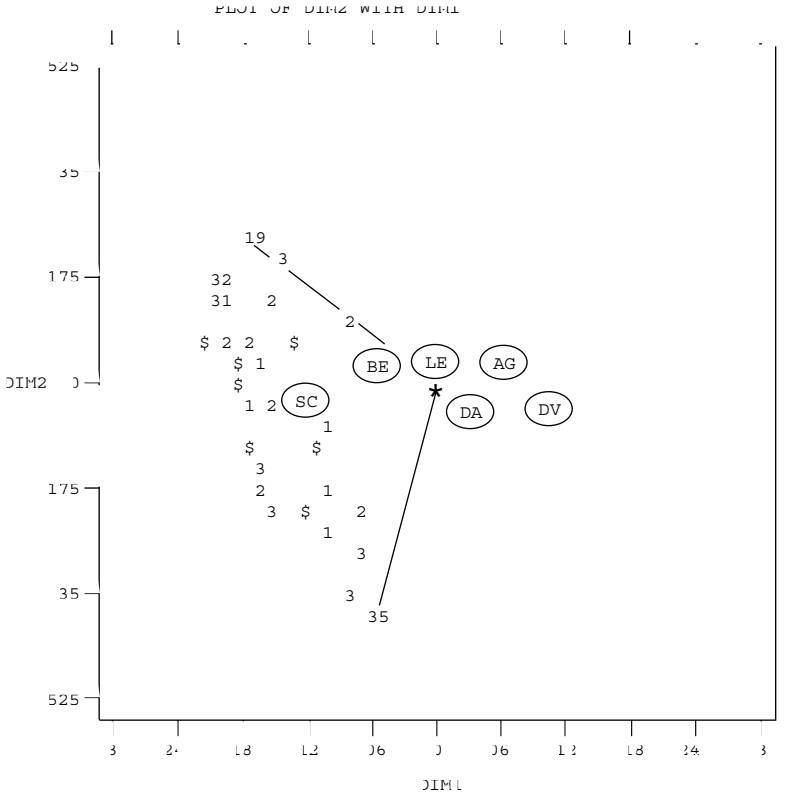
Aan de singuliere waarden van de oplossing zien we dat de eerste dimensie veruit de belangrijkste is. Dat betekent dat de preferentiescores van (de meeste) personen verklaard kunnen worden door de projecties van de bijdragen op de horizontale dimensie. Van rechts naar links zijn die projecties respectievelijk DV, AG, DA, LE, BE en SC, kortom, dezelfde volgorde die we al eerder bij de Thurstone-analyse gevonden hebben. De singuliere waarden van de tweede en hogere dimensies wijzen erop dat er toch wel wát variatie tussen de personen voorkomt. In Figuur 11.2 zien we een waaier van proefpersoonvectoren waarvan de grenzen gevormd worden door de vectoren van Persoon 19 (boven) en Persoon 35 (beneden). De vector van de laatstgenoemde heeft een zodanige richting dat SC en DA er de hoogste projecties op hebben terwijl BE, LE, DV en AG er lager op pro-

jecteren, wat in overeenstemming is met de preferentierangordeningen van die proefpersoon. Op de vector van Persoon 19 hebben BE, LE en SC hoge projecties terwijl de projecties van DV en DA veel lager zijn, met AG ertussenin. Ook dit klopt meer met de scores van die persoon.

Figuur 11.2 Afbeelding volgens het vectormodel van de data uit Tabel 11.1.



Merk op dat alle punten in de afbeelding aan één kant van de oorsprong zijn gelokaliseerd. Dat is het gevolg van het feit dat alle getallen in de O-matrix positief zijn. Een andere oplossing krijgen we als we de data uit Tabel 11.1 voorafgaand aan de analyse, in afwijking van hun rijgemiddelden zetten (bijvoorbeeld: van de scores van Proefpersoon 35 moet het gemiddelde 15/6 = 2.5 worden afgetrokken; dat levert nieuwe belangrijkheidsscores -1.5, -2.5, -.5, 1.5, 2.5 en .5). Wanneer we de nieuwe matrix met afwijkingsscores analyseren, dan levert dat andere uitkomsten voor U, LABDA, V en Y. De afbeelding van de eerste twee dimensies van U en Y is weergegeven in Figuur 11.3.



Figuur 11.3 Afbeelding volgens het vectormodel van de data uit Tabel 11.1 in afwijking van hun rijgemiddelden

De interpretatie van deze oplossing is identiek aan die van Figuur 11.2. We zien weer een waaier van persoonsvectoren die begrensd wordt door Personen 19 en 35. De belangrijkste dimensie is weer de horizontale as, met de volgorde DV, AG, DA, LE, BE en SC voor het belang van de bijdragen. Door het aftrekken van de rijgemiddelden is de oorsprong van de oplossing in de centroïde van de kolomobjecten terechtgekomen. Weer andere oplossingen kunnen we krijgen door niet de rijen maar de kolommen in afwijking van hun gemiddelden te noteren. Ook kunnen we de **O**-matrix *dubbel centreren*, dat wil zeggen, ervoor zorgen dat zowel de rij- als de kolomgemiddelden gelijk aan nul worden. Behalve bewerkingen waarin men de gemiddelden van de preferentiescores verandert, kan men ook de variantie van de scores transformeren. Zo kunnen de kolommen of de rijen eerst in standaardscores worden omgezet voordat er een SVD wordt toegepast.

In principe maakt het wel wat uit of we de **O**-matrix met ruwe gegevens analyseren of eerst een of andere bewerking op de rijen en kolommen van **O** toepassen. Welke van die bewerkingen het meest voor de hand ligt, hangt onder meer af van wat de data precies betekenen, met name qua conditionaliteit en meetniveau. De bewerking dat alle waarnemingen tot afwijkingen van de kolomgemiddelden getransformeerd worden, leidt tot een oplossing die identiek is aan de *principale-componentenanalyse* van de variantie-covariantiematrix van de kolommen. De tweedimensionale afbeelding die daarbij hoort, staat bekend onder de naam BIPLOT. SVD van data die binnen de kolommen in standaardscores zijn omgezet, komt neer op een principale-componentenanalyse van de correlatiematrix van de kolommen, een vorm van exploratieve factoranalyse.
Weer een andere manier om een matrix met positieve getallen te standaardiseren wordt gebruikt in een techniek die bekendstaat als *correspondentie-analyse*. Deze analysemethode is oorspronkelijk bedoeld voor de analyse van kruistabellen met frequenties, maar kan in principe ook gebruikt worden voor iedere willekeurige datamatrix met positieve elementen. Deze methode zullen we verderop in dit hoofdstuk behandelen.

11.3 NIET-METRISCHE ANALYSE MET HET VECTORMODEL

Bestaan de observaties uit echte preferentie*rangordeningen* dan moet men de data *niet-metrisch* analyseren. In dat geval geldt in het vectormodel dat de observaties $\{o_{ij}\}$ een *monotone transformatie f* zijn van de projecties van de punten $j = 1,...,m$ op de vector van Persoon *i*. Omgekeerd geldt dus ook dat deze projecties een monotone transformatie *g* zijn van de observaties, zodat

$$g(o_{ij}) = \Sigma_s x_{is} y_{js}. \qquad [11.9]$$

Zouden, behalve de observatie o_{ij}, ook de waarden van de projecties $p_{ij} = \sum_s x_{is} y_{js}$ bekend zijn, dan zouden we voor de transformatiefunctie *g* weer Kruskals monotone transformatie kunnen kiezen. De meest voor de hand liggende manier om rangordedata niet-metrisch met het vectormodel te analyseren is dus via een iteratief programma waarin beurtelings de functie *g* en de parameters x_{is} en y_{js} bepaald worden.
Opgemerkt moet worden dat bij klassieke preferentie-rangordeningen de data niet matrix- maar *rijconditioneel* zijn. In dat geval is het model dus

$$g_i(o_{ij}) = \Sigma_s x_{is} y_{js} \qquad [11.10]$$

waarin g_i de eigen monotone transformatiefunctie voor de observaties in rij i is. Een computerprogramma dat geschikt is om deze analyse uit te voeren is het zogenaamde PRINCALS-programma (Gifi, 1990) uit de module CATEGORIES van SPSS. Dit programma wordt in de volgende paragraaf wat uitgebreider besproken.

PRINCALS

PRINCALS is een acroniem voor *principal component analysis by alternating least squares* en is een van de door Gifi (1990) ontworpen computerprogramma's voor niet-lineaire multivariate analyse. PRINCALS voert principale-componentenanalyse uit op variabelen van verschillende meetniveaus. De variabelen kunnen op interval, ordinaal en nominaal niveau gemeten zijn[4]. Daarbij wordt elke variabele afzonderlijk *optimaal getransformeerd* op de manier die voor het desbetreffende meetniveau is toegestaan. Deze transformaties zijn dus variabele-conditioneel en omdat variabelen doorgaans corresponderen met de kolommen van de observatiematrix, voert PRINCALS een kolomconditionele analyse uit. In een klassieke matrix met preferentierangordeningen, zijn het echter juist de rijen die een eigen optimale transformatie moeten krijgen. Dus: om PRINCALS op preferentierangordeningen toe te kunnen passen, moet de datamatrix *getransponeerd* worden. De rijen en kolommen van de nabijheidsmatrix moeten verwisseld worden; men spreekt daarom van PRINCALS op een *gekantelde* matrix. Een voorbeeld van zo'n analyse wordt besproken in Blok 11.2 en Blok 11.3. Stel, we hebben preferentierangordeningen van vier studenten voor vier boeken van respectievelijk Reve, Hermans, Mulisch en Wolkers, waarbij het cijfer 1 het meest geprefereerde en 4 het minst geprefereerde boek aanduidt (zie Tabel 11.2).

BLOK 11.2 EEN VOORBEELD VAN PRINCALS

Tabel 11.3 Preferentierangordeningen van vier studenten voor vier boeken

Student	reve	herm	muli	wolk
1	2	1	3	4
6	4	1	2	3
11	1	2	4	3
20	2	1	4	3

Willen we deze gegevens met PRINCALS analyseren, dan moeten we de datamatrix eerst kantelen (transponeren). De SPSS-invoer wordt dan:

4 Als alle variabelen nominaal zijn, komt een PRINCALS-analyse op hetzelfde neer als een HOMALS-analyse. HOMALS is dus eigenlijk een bijzonder geval van PRINCALS.

```
data list table /boek 1 auteur 3-6(a)
      s1 s6 s11 s20 7-14.
begin data.
1 reve 2 4 1 2
2 herm 1 1 2 1
3 muli 3 2 4 4
4 wolk 4 3 3 3
end data.
value labels boeknr 1 `re' 2 `he' 3 `mu' 4 `wo'.
princals variables boeknr s1 s6 s11 s21 (4)
         /analysis s1 s6 s11 s21 (ordinal)
         /dimension=2
         /print default object
         /plot default loadings object (boeknr s1 s6 s11 s20)
          quant (s1 s6 s11 s20).
```

De uitvoer van deze opdrachten wordt hieronder behandeld. Eerst geeft PRINCALS een overzicht van de ingevoerde data.

```
THE NUMBER OF OBSERVATIONS USED IN THE ANALYSIS = 4

LIST OF VARIABLES
==================
                    NUMBER OF     MEASUREMENT
VARIABLE  VARIABLE LABEL        CATEGORIES   LEVEL
-------  -------------      ---------  ----------
S1                           4     ORDINAL
S6                           4     ORDINAL
S11                          4     ORDINAL
S20                          4     ORDINAL

MARGINAL FREQUENCIES
====================
VARIABLE  MISSING  CATEGORIES
             1    2    3    4
S1         0    1    1    1    1
S6         0    1    1    1    1
S11        0    1    1    1    1
S20        0    1    1    1    1
```

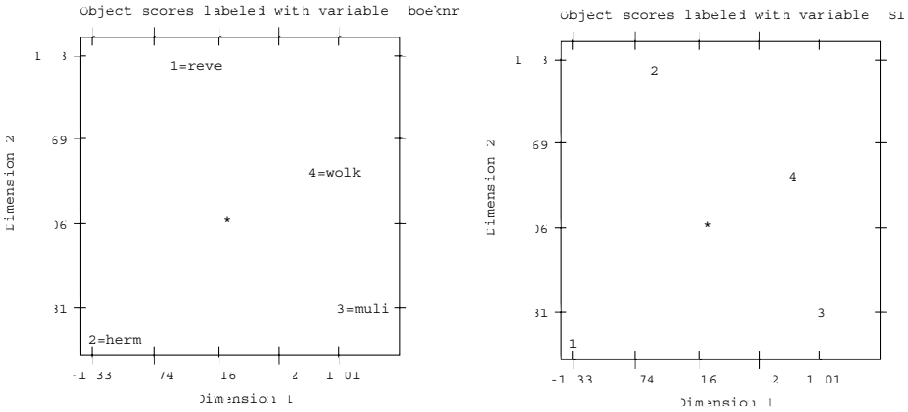
Vervolgens probeert PRINCALS een oplossing in twee dimensies te vinden. Als dat gelukt is, volgt de mededeling:

```
THE ITERATIVE PROCESS STOPS BECAUSE THE CONVERGENCE TEST VALUE IS
REACHED.

DIMENSION   EIGENVALUE
---------  ----------
    1          .6188
    2          .3812
---------  ----------
```

Deze eigenwaarden zijn een aanduiding voor de kwaliteit van de oplossing (de *goodness-of-fit*); de precieze betekenis ervan wordt verderop besproken. Eerst zullen we – in afwijking van de volgorde bij PRINCALS – twee andere onderdelen van de uitvoer bespreken: de coördinaten van de objecten en de richtingen van de persoonsvectoren.
Een van de belangrijkste uitkomsten van de analyse is de tabel met *object-scores*, dat wil zeggen, de coördinaten van de boeken. Dit zijn dus de waarden y_{js} uit de formule $g_i(o_{ij}) = \sum_s x_{is} y_{js}$ (zie Formule [11.10]), die het model beschrijft dat door PRINCALS '*gefit*' wordt. Merk op dat de term *object* hier – net als bij HOMALS – gebruikt wordt om de rijen van de datamatrix aan te geven. Als we de invoermatrix niet gekanteld hadden, waren de objecten hier dus personen geweest.

```
THE OBJECT SCORES ARE:
==========================
OBJECT *        DIMENSION
           1       2
   1 *    -.61     1.43
   2 *   -1.31    -1.10
   3 *    1.18     -.75
   4 *     .73      .42
------------------------
```

Door de opdracht `/plot object (boeknr s1 s6 s11 s20)` levert PRINCALS een aantal plaatjes waarin de objecten zijn afgebeeld: één plaatje met de nummers van de boeken erbij en vier met de rangordecijfers die door de personen zijn toegekend. Hieronder volgt eerst het plaatje met de boeknummers en daarna dat met de rangcijfers van Student 1.



Voor iedere persoon (variabele) geeft PRINCALS een tabel met *multiple category coordinates*. Voor Student 1 hebben die de volgende waarden (de multipele categoriekwantificaties van de andere studenten worden verderop weergegeven).

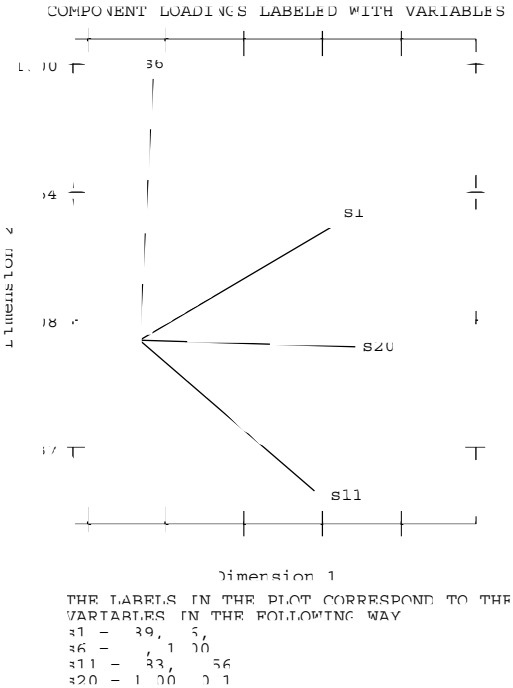
```
MULTIPLE CATEGORY COORDINATES
-------------------------
CATEGORY          DIMENSION
                1       2
     1       -1.31    -1.10
     2        -.61     1.43
     3        1.18     -.75
     4         .73      .42
-------------------------
```

Aan het plaatje hierboven is eenvoudig te zien waar die multipele categoriecoördinaten vandaan komen. Het zijn de coördinaten van de objecten die bij de desbetreffende categorienummers horen. Bijvoorbeeld: Object 2 links onderaan heeft van Student 1 het rangordecijfer 1 gekregen. Daarom krijgt Categorie 1 van deze persoon in PRINCALS de coördinaten van Object 2. In dit voorbeeld is er steeds maar één object per categorie. Zouden er meer objecten in dezelfde categorie zitten, dan worden de multipele categoriecoördinaten gelijk aan de coördinaten van de *centroïde* van de objecten in die groep.

Een van de belangrijkste uitkomsten van PRINCALS is een tabel met coördinaten waaruit we de richtingen van de persoonsvectoren kunnen reconstrueren. Deze coördinaten heten bij PRINCALS de *componentladingen*, dit zijn de waarden x_{is} uit Formule [11.10]. Deze waarden zijn als punten afgebeeld in een grafiek. Elk punt correspondeert met een van de personen. Als we die persoonspunten verbinden met de oorsprong, dan zien we de richtingen of vectoren die de preferentieordeningen van de personen zo goed mogelijk representeren. Bijvoorbeeld, de vector die bij Student 1 hoort, gaat door de oorsprong en door het punt (.886, .464).

COMPONENT LOADINGS

VARIABLE	DIMENSION 1	2
S1	.886	.464
S6	.041	.999
S11	.830	-.558
S20	1.000	.011



Om een totaalbeeld van de oplossing te krijgen zouden we de grafiek van componentladingen en die van de objectscores over elkaar heen moeten afbeelden. Projecteren we vervolgens de objecten op een van de richtingen, dan zou de volgorde van die projecties idealiter overeen moeten komen met de preferentierangorde van de betreffende persoon. Anders gezegd: de projecties van de objecten op de persoonsvector moeten zo hoog mogelijk correleren met een *monotoon stijgende transformatie* ($g_i(o_{ij})$ in Formule [11.10]) van de preferentierangordeningen. De waarden van deze monotoon stijgende transformaties worden door PRINCALS de *quantifications* van de variabelen genoemd. Voor Student 1 zijn deze kwantificaties -1.67, .13, .70 en .84 voor respectievelijk de rangordecijfers 1, 2, 3 en 4. In deze toepassing van PRINCALS corresponderen de 'variabelen' met de proefpersonen die rangordecijfers hebben toegekend aan de objecten. Deze rangordecijfers zijn door PRINCALS opnieuw (maar nu optimaal) gekwantificeerd. Dus: het rangordecijfer 1 (bij PRINCALS: Categorie 1) van de eerste proefpersoon heeft de nieuwe waarde -1.67 gekregen; het rangordecijfer 2 wordt .13, enzovoort. Laten we deze (nieuwe) kwantificaties hier aanduiden met symbool $q_c^{(i)}$, waarbij *i* een persoon aanduidt en *c* één van de door hem of haar gebruikte rangordecijfers of categorieën.
De kwantificaties zijn zodanig gestandaardiseerd dat hun gemiddelde gelijk aan nul is en hun kwadratensom gelijk aan *m* (hier: 4). Ze lopen bovendien altijd van klein naar groot, dat wil zeggen: het laagste rangordecijfer krijgt de laagste waarde, het hoogste rangordecijfer krijgt de hoogste kwantificatie. Dat leidt soms tot een interpretatieprobleem. Als een rangordecijfer van 1 betekent dat het betreffende object het meest geprefereerd wordt (zoals in dit voorbeeld het geval is), dan krijgt juist het meest geprefereerde object de laagste kwantificatie. In het plaatje met componentladingen wijzen de vectoren dan in de richting van het minst geprefereerde object. Dat maakt verder niets uit voor de *fit* van de oplossing maar wel voor de interpretatie!
Het berekenen van de kwantificaties $q_c^{(i)}$ gaat als volgt. Ieder object *j* heeft een projectie op de vector van persoon *i*. Die projecties vinden we door vanuit *j* de loodlijn neer te laten op vector *i*. Het punt waar die loodlijn de vector *i* snijdt heeft coördinaten op de dimensies van de oplossing, die we hier $z_{js}^{(i)}$ zullen noemen. Als nu object *j* door persoon *i* in categorie *c* is gestopt, dan kunnen we afspreken dat we categorie *c* als kwantificatie op dimensie *s* de waarde $q^{(i)}{}_{cs|j=c} = z_{js}^{(i)}$ toekennen. Deze kwantificaties worden in PRINCALS de *single category quantifications* genoemd. Er zijn voor elke categorie en elke persoon evenveel van zulke kwantificaties als er dimensies zijn. De relatie tussen de single category quantifications en de quantifications is nu dat de laatste verkregen worden door de single category quantifications te delen door de componentlading van persoon *i* op dimensie *s*, dus:
$q_c^{(i)} = (q^{(i)}{}_{cs|j=c.})/x_{is}$.

Hierboven hebben we gezien dat de single category quantifications coördinaten zijn van punten op de vector van persoon i. Wat gebeurt er nu als de volgorde van de projecties van de objecten op die vector niet overeenstemt met de preferentierangnummers die persoon *i* aan hen heeft toegekend? Stel dat vier objecten de preferentievolgorde 1, 2, 3 en 4 hadden terwijl hun projecties op de vector van persoon *i* in de volgorde 2, 1, 4 en 3 liggen. De single category quantifications worden dan gevonden door categorie 1 en categorie 2 als kwantificatie op elke dimensie het gemiddelde toe te kennen van de coördinaten van de projecties van object 1 en object 2 op vector *i*. Categorie 3 en 4 krijgen dan de gemiddelde coördinaten van de projecties van object 3 en object 4.
Uit het bovenstaande blijkt dat er nauwe relaties bestaan tussen allerlei aspecten van de PRINCALS-uitvoer, met name tussen de multiple category quantifications en de object scores, en tussen de single category quantifications, de quantifications, de projecties van de objecten en de component loadings. Hieronder volgt de complete uitvoer van PRINCALS met betrekking tot de kwantificaties.

```
VARIABLE: S1
_ _ _ _ _ _
TYPE: ORDINAL  MISSING: 0
CATEGORY:       MARGINAL FREQUENCY     QUANTIFICATION
--------     ----------------    -------------
     1        1                -1.67
     2        1                 .13
     3        1                 .70
     4        1                 .84
SINGLE CATEGORY COORDINATES
-------------------------
CATEGORY         DIMENSION
           1      2
     1   -1.48    -.77
     2    .11     .06
     3    .62     .32
     4    .75     .39

MULTIPLE CATEGORY COORDINATES
---------------------------
CATEGORY         DIMENSION
           1      2
     1   -1.31   -1.10
     2    -.61    1.43
     3    1.18    -.75
     4     .73     .42
================================================================

VARIABLE: S6
_ _ _ _ -
TYPE: ORDINAL  MISSING: 0
CATEGORY:       MARGINAL FREQUENCY     QUANTIFICATION
--------     ----------------    -------------
     1        1                -1.16
     2        1                -.70
     3        1                 .45
     4        1                1.40
```

```
SINGLE CATEGORY COORDINATES
-----------------------
CATEGORY        DIMENSION
          1      2
    1    -.05   -1.16
    2    -.03    -.70
    3     .02     .45
    4     .06    1.40
MULTIPLE CATEGORY COORDINATES
-------------------------
CATEGORY        DIMENSION
          1      2
    1   -1.31   -1.10
    2    1.18    -.75
    3     .73     .42
    4    -.61    1.43
================================================================

VARIABLE: S11
--------
TYPE: ORDINAL  MISSING: 0
CATEGORY:      MARGINAL FREQUENCY     QUANTIFICATION
--------    ----------------    -------------
    1        1                -1.30
    2        1                 -.47
    3        1                  .37
    4        1                 1.40
SINGLE CATEGORY COORDINATES
-----------------------
CATEGORY        DIMENSION
          1      2
    1   -1.08     .72
    2    -.39     .26
    3     .31    -.21
    4    1.16    -.78
MULTIPLE CATEGORY COORDINATES
-------------------------
CATEGORY        DIMENSION
          1      2
    1    -.61    1.43
    2   -1.31   -1.10
    3     .73     .42
    4    1.18    -.75
================================================================

VARIABLE: S20
- - - - -
TYPE: ORDINAL  MISSING: 0
CATEGORY:      MARGINAL FREQUENCY     QUANTIFICATION
--------    ----------------    -------------
    1        1                -1.32
    2        1                 -.59
    3        1                  .74
    4        1                 1.17

SINGLE CATEGORY COORDINATES
-----------------------
CATEGORY        DIMENSION
          1      2
    1   -1.32    -.01
    2    -.59    -.01
    3     .74     .01
    4    1.17     .01
```

```
MULTIPLE CATEGORY COORDINATES
-----------------------------
CATEGORY        DIMENSION
          1      2
    1   -1.31   -1.10
    2    -.61    1.43
    3     .73     .42
    4    1.18    -.75
================================================================
```

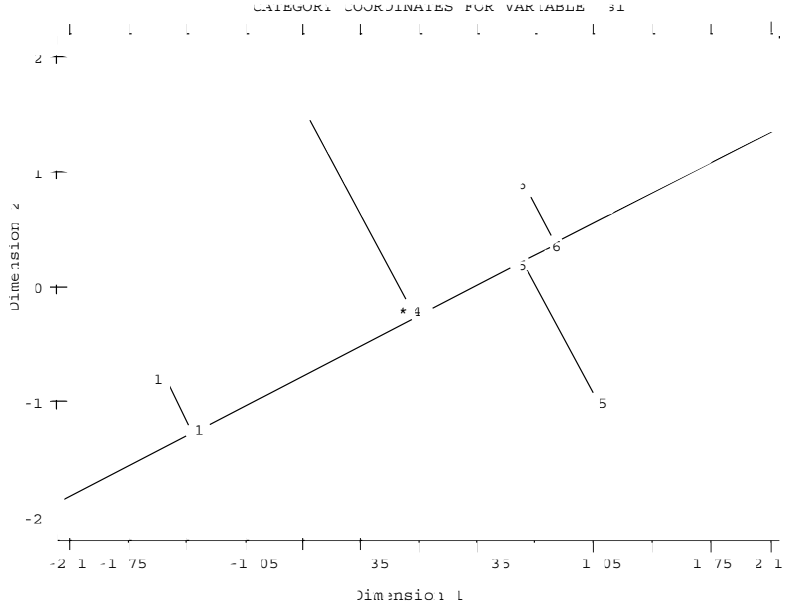
Naast de eerdergenoemde eigenwaarden geeft PRINCALS nog een ander, gedetailleerder, overzicht van de *fit* van de oplossing. De *fit* van de oplossing bestaat uit de *multiple fit* en de *single fit*, die per variabele en per dimensie berekend worden. De multiple *fit* van een variabele op een dimensie is identiek aan wat bij HOMALS de *discriminatiewaarde* heet. In dit voorbeeldje zijn alle discriminatiewaarden gelijk aan 1.00 omdat iedere categorie slechts één object bevat (immers: ieder rangcijfer is maar aan één object toegekend). Daardoor is de SS_W gelijk aan 0.00 en is de SS_B gelijk aan SS_T, zodat de discriminatiemaat SS_B/SS_T gelijk aan 1.00 is.

```
MULTIPLE FIT
------------
VARIABLE   ROW SUMS      DIMENSION
--------             1     2
S1          2.000     1.000   1.000
S6          2.000     1.000   1.000
S11         2.000      1.000   1.000
S20         2.000      1.000   1.000

MEAN:       2.000      1.000   1.000

SINGLE FIT
----------
VARIABLE   ROW SUMS      DIMENSION
--------             1     2
S1         1.000      .784    .216
S6         1.000      .002    .998
S11        1.000      .689    .311
S20        1.000     1.000    .000

MEAN:      1.000      .619    .381
```

De *single fit* van een variabele (hier: persoon) op een dimensie is gelijk aan het kwadraat van de bijbehorende componentlading op dezelfde dimensie. Zo'n componentlading x_{is} is gelijk aan de correlatie tussen de objectscores van de objecten op dimensie *s* en de kwantificaties van de categorieën die persoon *i* aan de betreffende objecten heeft toegekend. De single *fit* is dus een gekwadrateerde correlatiecoëfficiënt en geeft aan hoe goed de oorspronkelijke rangordecijfers van persoon *i* verklaard kunnen worden uit de objectscores op dimensie *s*. De gemiddelden van de kolommen van de 'single *fit*'-tabel zijn gelijk aan de eigenwaarden van de PRIN-

CALS-oplossing. Deze waarden staan onder de tabel. De som van de single *fit* over alle dimensies is het kwadraat van de *multipele correlatie* tussen de kwantificaties van de objecten en hun coördinaten op de verschillende assen. Deze waarden staan in de kolom met *row sums*.
Merk op dat voor alle variabelen (personen) de *fit* gelijk aan 1.00 is: de projecties van de objecten op de vectoren staan bij alle personen dus in dezelfde volgorde als hun preferentierangordeningen. Dat is geen toeval, maar ligt aan dit voorbeeld: de preferenties van *vier* personen voor *vier* objecten kunnen met een *ordinaal vectormodel* altijd perfect in *twee* dimensies worden afgebeeld. Als er meer objecten zijn, dan geeft een afbeelding in twee dimensies meestal geen perfecte *fit*.

Merk ook op dat de lengte van een persoonsvector gelijk is aan de multipele correlatie tussen de kwantificaties en de projecties van de objecten voor een bepaalde persoon. In de samengestelde grafiek van objecten en vectoren kunnen we aan de lengte van een vector dus ook nog zien hoe goed de *fit* is. Dezelfde standaardisatie zijn we al eerder tegengekomen: bij externe analyse volgens het vectormodel (Hoofdstuk 7) en in ALSCAL bij de gewichten van ASCAL en INDSCAL.



Figuur 11.4 Object scores, categoriekwantificaties en projecties op de preferentievector van Student 1 (NB: hoge rangcijfers komen overeen met lage preferenties).

Met de opdracht /plot quant (s1 s6 s11 s20) worden plaatjes verkregen waarin drie verzamelingen punten samen zijn afgebeeld: de objectscores, de projecties van de objecten op de persoonsvector, en de posities van de kwantificaties op de persoonsvector. Het plaatje voor Student 1 is weergegeven in Figuur 11.4. Omdat alle variabelen in ons voorbeeld een perfecte *fit* hebben (de correlatie tussen kwantificaties en projecties is 1.00), vallen de projecties van de objecten samen met de kwantificaties. In het algemeen is dat natuurlijk niet zo.

PRINCALS met ontbrekende gegevens

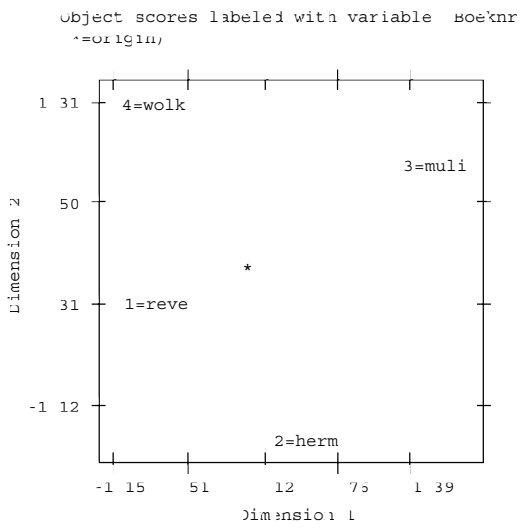
Soms zijn niet van alle personen de rangordeningen voor alle objecten beschikbaar. Dat kan het geval zijn omdat sommige proefpersonen 'vergeten' of niet in staat waren alle objecten te rangordenen, of het kan het gevolg zijn van het soort instructie dat aan de proefpersonen gegeven is. Bij een taak van het type *order k of m*, krijgt men per respondent slechts k rangordecijfers en zijn er $m - k$ ontbrekende getallen. Als we deze de code 0 toekennen, betrekt PRINCALS ze niet in de berekeningen.
In het voorbeeldje van boeken en studenten hebben we Student 20 vervangen door Student 21, die alleen de eerste drie boeken gelezen had. De PRINCALS-invoer bestaat nu uit:

```
data list table /boeknr 1 auteur 3-11 (a)
          s1 s6 s11 s21 12-19.
begin data.
1 vh reve   4 6 2 2
2 hermans   1 1 3 1
3 mulisch   5 2 5 3
4 wolkers   6 4 4 0
end data.

value labels boeknr 1 `re' 2 `he' 3 `mu' 4 `wo'.
princals variables boeknr s1 s6 s11 s21 (6)
      /analysis s1 s6 s11 s21 (ordinal)
      /dimension=2
      /print default object
      /plot default loadings object (boeknr s1 s6 s11 s21) quant (s1 s6 s11 s21).
```

De belangrijkste uitvoer van PRINCALS volgt hieronder.

The object scores are

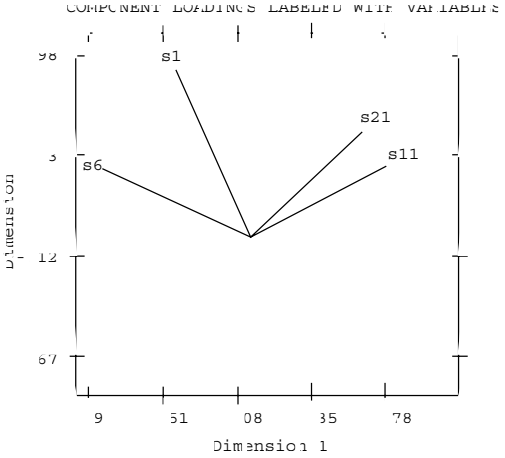
Object # Dimension
1 .
2 .
3 .
. .



Het plaatje van de objectscores laat vrijwel dezelfde structuur zien als bij de vorige analyse, zij het dat deze configuratie ongeveer een kwartslag linksom geroteerd is. Deze rotatie vinden we ook terug bij de componentladingen. De vector van Student 1 loopt nu naar linksboven in plaats van naar rechtsboven, S11 loopt enigszins naar rechtsboven in plaats van naar rechtsonder. S6 loopt een beetje naar linksboven in plaats van loodrecht omhoog.

COMPONENT LOADINGS

VARIABLE	DIMENSION 1	2
S1	-.46	.98
S6	-.94	.44
S11	.93	.39
S21	.75	.66



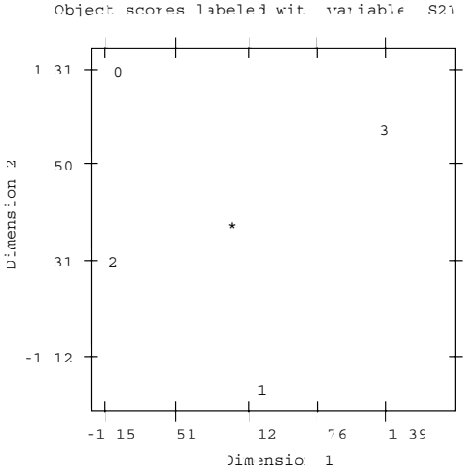

```
DIMENSION    EIGENVALUE
- - - - -     - - - - -
   1         .6254
   2         .4336
MULTIPLE FIT
- - - - - -
VARIABLE    ROW SUMS      DIMENSION
- - - -                 1       2
S1          2.172      1.065   1.107
S6          2.172      1.065   1.107
S11         2.172      1.065   1.107
S21         1.483       .805    .679

MEAN:        2.000      1.000   1.000
- - - - - - - - - - - - - - - - - - - -
SINGLE FIT
- - - - -
VARIABLE    ROW SUMS      DIMENSION
- - - -                 1       2
S1          1.162       .209    .953
S6          1.069       .879    .191
S11         1.010       .858    .152
S21          .995       .556    .439
MEAN:       1.059       .625    .434
```

De eigenwaarden en de rij-sommen van de tabellen van *single* en *multiple fit* suggereren dat de *fit* van deze oplossing nog beter is dan van de vorige, maar dat kan natuurlijk niet (de vorige was immers perfect). Dat de rij-sommen groter dan 1.00 worden, ligt aan het ene ontbrekende getal bij Student 21. Net als bij HOMALS worden de objectscores nu per dimensie

zodanig gestandaardiseerd dat $\sum_j v_j y^2_{js} = n.m.$ Daardoor is de multipele *fit* van een variabele niet meer gelijk aan de 'normale' discriminatiewaarde en is de enkelvoudige *fit* niet gelijk aan een correlatiecoëfficiënt. Enige voorzichtigheid bij de interpretatie is dus geboden, zeker als er relatief veel ontbrekende getallen zijn.

Hiernaast volgen van Student 21 de categoriekwantificaties, de enkelvoudige en multipele categoriecoördinaten en de afbeelding daarvan.

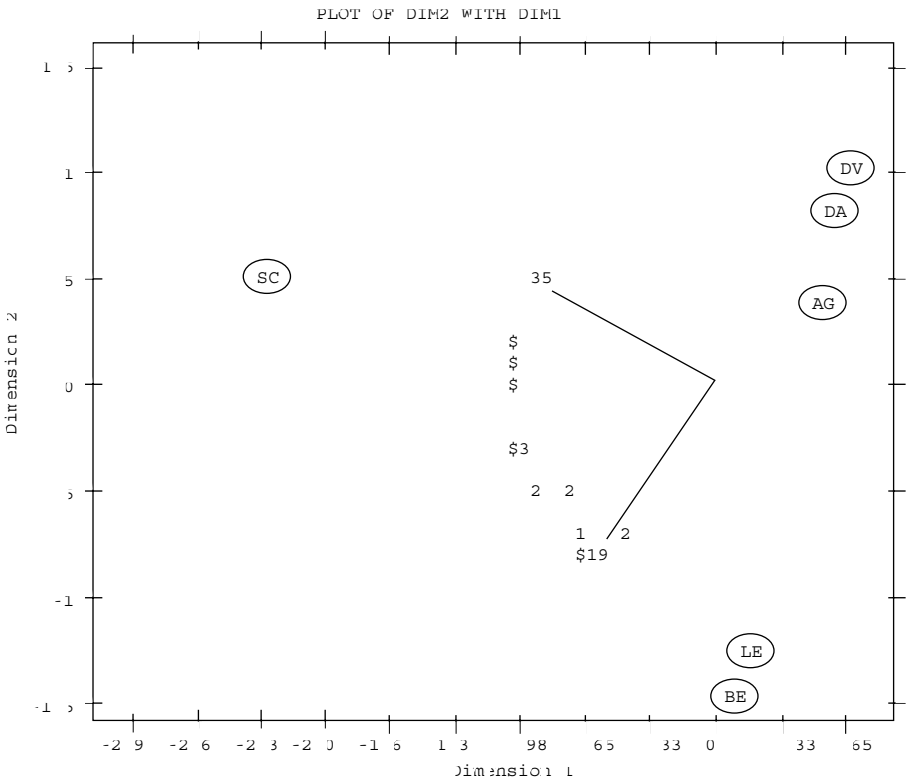
```
VARIABLE: S21
- - - - -
TYPE: ORDINAL  MISSING:      1
CATEGORY:      MARGINAL FREQUENCY            QUANTIFICATION
-----          ------------------            --------------
    1       1                  -.84
    2       1                  -.84
    3       1                  1.61

SINGLE CATEGORY COORDINATES
- - - - - - - - - - - - - - -
CATEGORY         DIMENSION
          1      2
    1    -.63    -.56
    2    -.63    -.56
    3    1.20     1.06

MULTIPLE CATEGORY COORDINATES
- - - - - - - - - - - - - - - -
CATEGORY         DIMENSION
          1      2
    1     .26   -1.44
    2    -.98    -.28
    3    1.48     .74
```



BLOK 11.3 PRINCALS OP DE RANGORDENINGEN VAN AUTEURSBIJDRAGEN

In Blok 11.1 is de metrische analyse besproken van de preferentiescores van 40 personen voor zes auteursbijdragen uit Tabel 11.1. Hieronder volgen de belangrijkste resultaten van een niet-metrische analyse van dezelfde data door middel van PRINCALS. Om zo'n analyse uit te kunnen voeren moeten de data eerst op twee manieren aanpassen. In de eerste plaats moeten de preferentiescores worden omgezet in rangordeningen waarvan het laagste getal een 1 is (een 0 wordt door PRINCALS als ontbrekende observatie beschouwd!) en die verder uit gehele getallen bestaan. Dat laatste betekent dat we voor *ties* niet zomaar het gemiddelde van een aantal rangordescores kunnen kiezen. De handigste oplossing is om *opeenvolgende rangnummers* te kiezen. Bijvoorbeeld: de scores {0, 1, 1, 4} worden niet omgezet in {1, 2.5, 2.5, 4} maar in {1, 2, 2, 3}. In de tweede plaats moet de datamatrix worden gekanteld, getransponeerd. Voor de auteurschapsdata betekent dat, dat we een nieuwe matrix krijgen met zes rijen (de auteursbijdragen zijn nu dus de objecten geworden) en veertig kolommen (de personen zijn nu de variabelen). Beide aanpassingen kunnen eenvoudig in de WINDOWS-versie van SPSS gerealiseerd worden, respectievelijk met `TRANSPOSE` en `RANK`.

Een PRINCALS-analyse in twee dimensies levert eigenwaarden van .846 en .135. De eerste eigenwaarde is de belangrijkste, zodat we kunnen concluderen dat de rangordeningen van de meeste proefpersonen tot één en dezelfde belangrijkheidsvolgorde van de bijdragen herleid kunnen worden. De proefpersoonsvectoren (de componentladingen) en de bijdragen (de objectscores) zijn afgebeeld in Figuur 11.5. Deze figuur lijkt in hoge mate op Figuur 11.3, maar er zijn ook verschillen. We zien weer een waaier van persoonsvectoren die begrensd wordt door Persoon 19 (beneden) en Persoon 35 (boven), maar de waaier is minder gespreid. We zien één losse proefpersoon (Persoon 35) en drie samenklonterende groepjes personen. Ook de configuratie van bijdragen uit PRINCALS lijkt op die van de metrische analyse. Er is echter een belangrijk, inhoudelijk verschil. In de PRINCALS-oplossing is de ordening van de bijdragen langs de eerste dimensie (van rechts naar links) DV, DA, AG, LE, BE en SC terwijl dat in de metrische oplossing DV, AG, DA, LE, BE en SC was. De auteurschapsgrens heeft dus een andere positie gekregen: volgens deze analyse is data-analyse op zichzelf niet voldoende om voor het auteurschap van een publicatie in aanmerking te komen.



Figuur 11.5 PRINCALS*-oplossing van de in rangnummers omgezette data uit Tabel 11.1*

11.4 CORRESPONDENTIEANALYSE

Een oorspronkelijk in Frankrijk ontwikkelde analysetechniek, die sinds enige tijd ook in Nederland en de Engelssprekende landen meer bekendheid heeft gekregen, is *analyse des correspondances* (Benzécri, 1966). Deze techniek is in feite een speciale toepassing van het vectormodel en singuliere-waardendecompositie. Oorspronkelijk gaat het in deze techniek om het analyseren van een zogenaamde *kruistabel* met frequenties die aangeeft hoe vaak elke combinatie voorkomt van de categorieën van twee variabelen. Zo'n tabel heeft *r* rijen en *k* kolommen, waarbij *r* en *k* het aantal categorieën van de twee variabelen aanduiden. Heeft men eenmaal zo'n tabel verzameld, dan is het gebruikelijk om na te gaan of er een *verband* of *associatie* tussen de rijvariabele en de kolomvariabele bestaat. Daartoe berekent men zogenaamde *expected frequencies*,

dat wil zeggen, de frequenties die men zou verwachten als de categorieën van de rij- en kolomvariabelen onafhankelijk van elkaar waren. Noemen we de geobserveerde frequenties o_{ij} en de bijbehorende verwachte frequenties e_{ij} dan is $e_{ij} = (\sum_i o_{ij})(\sum_j o_{ij})/(\sum_i \sum_j o_{ij})$. De toetsingsgrootheid voor onafhankelijkheid van rijen en kolommen is, zoals bekend

$$\chi^2 = \sum_i^r \sum_j^k \left(\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \right).$$

Hoewel deze toets laat zien of er een samenhang is tussen de rij- en kolomvariabelen, zegt hij niets over de precieze vorm en inhoud van die samenhang. Welke categorieën van de ene variabele gaan er nu meer of minder vaak samen dan men op grond van toeval kan verwachten?
Een antwoord op deze vragen wordt gegeven door correspondentieanalyse. Het doel van zo'n analyse is het vinden van een aantal dimensies (met kwantificaties voor de rij- en kolomelementen) die ieder voor zich een deel van de associatie (de χ^2-waarde) tussen de rijen en kolommen verklaren. Dat kunnen we op twee manieren uitleggen: algebraïsch en grafisch. Waar het algebraïsch op neerkomt, is dat van de cellen o_{ij} van de geobserveerde kruistabel de bijbehorende verwachte frequenties e_{ij} worden afgetrokken waardoor een matrix met residuen r_{ij} ontstaat. Vervolgens worden deze residuen gedeeld door $\sqrt{(N \times e_{ij})}$ waarbij N het totaal aantal observaties in alle cellen is. De resulterende matrix (laten we die **R** noemen) wordt via singuliere-waardendecompositie (SVD) ontbonden in singuliere waarden $\mathbf{\Lambda}$ en linker en rechter eigenvectoren **U** en **V**, zodat $\mathbf{R} = \mathbf{U\Lambda V'}$. Voor de singuliere waarden geldt dat $N \times \sum_s \lambda^2_s = \chi^2$; de totale χ^2-waarde van de tabel is nu dus opgedeeld in een aantal onafhankelijke componenten die bij de verschillende dimensies van de SVD horen. We spreken daarom van χ^2-decompositie. Elke SVD-dimensie 'verklaart' als het ware een onafhankelijk deel van de samenhang tussen de rijen en kolommen van de kruistabel. Het aantal SVD-dimensies is bij correspondentieanalyse altijd kleiner dan of gelijk aan het aantal categorieën van de rijvariabele min 1 of aan het aantal categorieën van de kolomvariabele min 1, dat wil zeggen, aan de kleinste van de twee.
Na het berekenen van de SVD-oplossing worden de waarden in de kolommen van **U** en **V** vervolgens op een bepaalde manier gestandaardiseerd, waardoor twee nieuwe matrices **W** en **Z** ontstaan. De standaardisatie is zodanig dat $\sum_i \{w^2_{is}(\sum_j o_{ij})\} = N$ en $\sum_{js} \{z^2_{js}(\sum_i o_{ij})\} = N$. Dat gebeurt door de waarden in de eigenvectoren te vermenigvuldigen met de wortel uit N, gedeeld door het bijbehorende rij- of kolomtotaal: $w_{is} = u_{is}\sqrt{(N/\sum_j o_{ij})}$ en $z_{js} = v_{js}\sqrt{(N/\sum_i o_{ij})}$. De gestandaardiseerde matrix **W** kunnen we gebruiken om de categorieën van de rijvariabele af te beelden. Vervolgens willen we dan ook de kolomcategorieën afbeelden in dezelfde ruimte. Hierbij zijn er weer verschillende mogelijkheden: (a) voor de rijcategorieën nemen we **W** en voor de kolomcategorieën $\mathbf{Z\Lambda}$, (b) voor de rijcategorieën nemen we $\mathbf{W\Lambda}$ en voor de kolomcategorieën **Z**, (c) voor

de rijcategorieën nemen we $\mathbf{Z\Lambda}^{1/2}$ en voor de kolomcategorieën $\mathbf{W\Lambda}^{1/2}$ (daarover later meer).
Hiermee zijn we toe aan de grafische benadering van correspondentieanalyse. Stel, we willen de rij-elementen van de kruistabel afbeelden als punten in de ruimte. Om dat te kunnen doen, zouden we moeten weten wat de afstanden tussen die punten zijn, zodat we bijvoorbeeld de Young-Householdermethode zouden kunnen gebruiken om coördinaten te vinden. Het gaat er dus om een geschikte maat te definiëren om de afstand tussen twee rijen met frequenties aan te duiden. Het idee is nu dat twee rijen uit een kruistabel de afstand nul hebben, als de frequenties uit de ene rij evenredig zijn met de frequenties uit de andere rij. Voor twee rijen i en h geldt dan dat $o_{ij} = c(o_{hj})$ voor alle kolomcategorieën $j = 1,...,k$. Naarmate de frequenties uit de ene rij minder evenredig met de frequenties uit de andere rij zijn, neemt de afstand tussen die rijen toe. De afstandsfunctie waar het hier om gaat is de zogenaamde χ^2-afstand, die gedefinieerd is als

$$d^2_{ih(\chi^2)} = \sum_{j=l}^{k} \left(\frac{1}{\sum_i o_{ij}} \right) \left(\frac{o_{ij}}{\sum_j^k o_{ij}} - \frac{o_{kj}}{\sum_j^k o_{kj}} \right).$$

Metrische MDS met behulp van de methode van Young en Householder levert dan dezelfde afbeelding op die met correspondentieanalyse verkregen wordt als daarbij $\mathbf{W\Lambda}^{1/2}$ als kwantificatie van de rijcategorieën gekozen wordt.
Zoals eerder gezegd, is correspondentieanalyse een heel algemene analysetechniek die op allerlei data toegepast kan worden. Alle vormen van Nishisato's (1980, 1994) *dual scaling* zijn toepassingen van correspondentieanalyse. Een algemene inleiding is van Greenacre (1984).
Hieronder zullen we in Blok 11.4 een klein rekenvoorbeeld van correspondentieanalyse behandelen, namelijk van een 6×2 kruistabel. Daarna, in Blok 11.5, bespreken we de resultaten van een correspondentieanalyse van een grotere tabel.

BLOK 11.4 MISDAAD EN STRAF

In Tabel 11.3 staat een kruistabel (ontleend aan Willemsen, 1990) die weergeeft hoe vaak ongewenst gedrag van meisjes en jongens bestraft of niet bestraft werd. Het ging in dit onderzoek om 40 meisjes en 35 jongens die in zogenaamde leefgroepen van de residentiële jeugdhulpverlening woonden. Deze leefgroepen worden begeleid door groepsleiders en -leidsters. Aan hen werd gevraagd of er sprake was geweest van ongewenste gedragingen van de jongere tijdens zijn of haar verblijf in de leefgroep, en zo ja, welke straf

hiervoor was uitgedeeld. Ongewenst gedrag is onderverdeeld in drie categorieën: geweld, vermogensmisdrijven (inbraak, stelen, zwart reizen) en overtreding van de huisregels (spijbelen, weglopen, nacht wegblijven, dronkenschap, ongewenste seksuele contacten). Straf kon bestaan uit huisarrest, aangifte bij de politie of boete betalen; de categorie geen straf bestond uit geen straf of een gesprek.

Tabel 11.3 Frequenties f_{ij} van drie soorten ongewenst gedrag en de hiervoor verkregen straf

Ongewenst gedrag	geen straf	straf	totaal
geweld door meisje	15	5	20
geweld door jongen	20	20	40
vermogensmisdrijf, meisje	8	2	10
vermogensmisdrijf, jongen	8	8	16
regelovertreding, meisje	41	25	66
regelovertreding, jongen	27	29	56
totaal	119	89	208

Tabel 11.4 Verwachte waarden e_{ij} van drie soorten ongewenst gedrag en de hiervoor verkregen straf

Ongewenst gedrag	geen straf	straf	totaal
geweld door meisje	11.44	8.56	20
geweld door jongen	22.88	17.12	40
vermogensmisdrijf, meisje	5.72	4.28	10
vermogensmisdrijf, jongen	9.15	6.85	16
regelovertreding, meisje	37.76	28.24	66
regelovertreding, jongen	32.04	23.96	56
totaal	119	89	208

In Tabel 11.4 staan de verwachte waarden van de cellen uit Tabel 11.3, waarbij is aangenomen dat er geen associatie tussen de rij- en kolomcategorieën zou bestaan. De bijbehorende χ^2-waarde is 8.398 (df = 5; p = .136). In Tabel 11.5 staan de verschillen van Tabel 11.3 en 11.4 nadat ze gedeeld zijn door $\sqrt{(N \times e_{ij})}$. In dezelfde tabel staan de linker en rechter eigenvectoren **U** en **V** en de singuliere waarde λ. Aangezien het aantal rijcategorieën gelijk aan zes is en het aantal kolomcategorieën gelijk aan

twee, is het aantal dimensies gelijk aan 2 – 1 = 1. Ook zijn in deze tabel de gestandaardiseerde eigenvectoren **W** en **Z** weergegeven.

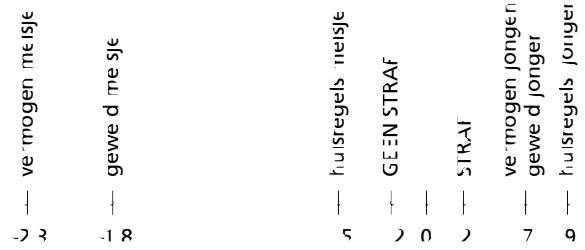
Tabel 11.5 Genormeerde residuen $(f_{ij} - e_{ij})/(\sqrt{N} \times e_{ij})$, ongestandaardiseerde en gestandaardiseerde eigenvectoren en singuliere waarde Λ

Ongewenst gedrag	geen straf	straf	U	W
geweld door meisje	.0729	-.0843	-.555	-1.789
geweld door jongen	-.0418	.0483	.318	.725
vermogensmisdrijf, meisje	.0661	-.0764	-.503	-2.292
vermogensmisdrijf, jongen	-.0264	.0306	.201	.725
ongewenst gedrag, meisje	.0366	-.0423	-.278	-.494
ongewenst gedrag, jongen	-.0617	.0714	.470	.905
V	-.6541	.7564		
Z	-.8650	1.1560		
λ = .2009				

Gaan we uit van de gestandaardiseerde eigenvector **W** als kwantificatie voor de rijcategorieën, dan kunnen we een kwantificatie voor de categorie straf vinden door de kwantificaties van de gedragingen te *wegen* met de proportie van het aantal keren dat deze gedragingen bestraft werden. We berekenen op die manier de *gemiddelde categoriekwantificatie* van alle gedragingen die tot straf geleid hebben. De kwantificatie voor straf wordt dan (5/89)(-1.789) + (20/89)(.725) + (2/89)(-2.292) + (8/89)(.725) + (25/89)(-.494) + (29/89)(.905) = .232. Het punt voor straf wordt dus gelokaliseerd in het gewogen gemiddelde van de bijbehorende gedragskwantificaties. Merk op: deze waarde kunnen we ook direct berekenen, namelijk als $z_2 \times \lambda$ = (1.1560)(.2009) = .232. Op soortgelijke manier kunnen we de kwantificatie voor geen straf berekenen, deze wordt (15/89)(-1.789) + (20/89)(.725) + (8/89)(-2.292) + (8/89)(.725) + (41/89)(-.494) + (27/89)(.905) = -.174.

Met behulp van bovenstaande kwantificaties kunnen we nu de eendimensionale afbeelding maken die in Figuur 11.6 is weergegeven. Van links naar rechts gaand zien we in deze figuur de categorieën vermogensdelict door meisje, geweld door meisje, overtreding van huisregels door meisje, geweld en vermogensdelict door jongen (vallen samen) en overtreding van huisregels door jongen. De categorie geen straf ligt links in de figuur, de categorie straf rechts. Dit laat zien (a) dat meisjes over het algemeen minder

vaak gestraft worden dan jongens, (b) dat meisjes vaker dan verwacht gestraft worden voor overtreding van de huisregels en (c) dat meisjes minder vaak dan verwacht gestraft worden voor vermogensdelicten.



Figuur 11.6 Kwantificatie van ongewenste gedragingen en straf

Een soortgelijke afbeelding kunnen we krijgen door **Z** als kwantificaties voor straf en geen straf te kiezen en vervolgens de gedragingen tussen deze twee waarden in te plaatsen. De coördinaat voor geweld door een jongen wordt dan (20/40)(-.865) + (20/40)(1.156) = .146. Zie weer dat $w_2 \times \lambda =$ (.725)(.2009) = .146. Er zijn dus twee gelijkwaardige alternatieven om de rijen en kolommen te kwantificeren. In het ene geval worden de kolomcategorieën geprojecteerd op de ruimte (hier een lijn) van de rijcategorieën, in het andere geval worden de rijcategorieën op de ruimte van de kolommen geprojecteerd. Deze mogelijkheden liggen voor de hand als men respectievelijk de kolomvariabele uit de rijvariabele wil 'voorspellen' of de rijvariabele uit de kolomvariabele.[5]

Twee andere manieren om de rij- en kolomcategorieën te kwantificeren zijn respectievelijk de *canonische* of symmetrische manier en de *principale* of Franse manier. In de canonische aanpak kiest men $\mathbf{W\Lambda}^{1/2}$ en $\mathbf{Z\Lambda}^{1/2}$ als rij- en kolomkwantificaties; in de principale of Franse aanpak kiest men $\mathbf{W\Lambda}$ en $\mathbf{Z\Lambda}$. In het laatste geval 'mag' men de rij- en kolomkwantificaties strikt genomen niet meer in één en dezelfde figuur afbeelden.

Een prettige eigenschap van correspondentieanalyse is dat men in de uiteindelijke oplossing allerlei extra gegevens kan afbeelden, dat wil zeggen, gegevens die niet aanwezig waren in de kruistabel die geanalyseerd is. Bijvoorbeeld: stel dat we zouden weten dat precies de helft van alle straffen uit huisarrest bestaat. We kunnen huisarrest dan kwantificeren door middel van (2.5/89)(-1.789) + (10/89)(.725) + (1/89)(-2.292) + (4/89)(.725) + (22.5/89)(.905) = .116 (voor het gemak nemen we aan dat er ook halve

5 In het ANACOR-programma van SPSS heten deze opties respectievelijk *column principal* en *row principal normalization*.

keren gestraft kan worden). Huisarrest kan op deze manier dus ook in de afbeelding gelokaliseerd worden. Dit soort extra gegevens wordt in de Franse literatuur *les éléments supplémentaires* genoemd.
Correspondentieanalyse is niet alleen bruikbaar voor frequentiematrices of kruistabellen. In principe is deze techniek toepasbaar op iedere matrix met positieve getallen. Correspondentie-analyse kan worden uitgevoerd met behulp van het computerprogramma ANACOR dat tot de door Gifi (1990) beschreven programma's voor niet-lineaire multivariate analyse[6] behoort en dat is opgenomen in de SPSS-module CATEGORIES. Een voorbeeld van een correspondentie-analyse met ANACOR wordt in Blok 11.5 besproken.

BLOK 11.5 EEN INHOUDSANALYSE VAN KOOKBOEKEN

Van dertien bekende en minder bekende, traditionele en moderne, regionale en algemene, Franse kookboeken hebben Van der Kloot en Willemsen (1985) genoteerd hoeveel bladzijden er van elk boek gewijd waren aan twaalf verschillende onderwerpen[7]. De verzameling kookboeken bestond uit boeken van Escoffier (ESC), Pellaprat (PEL), twee boeken van Bocuse (BO1 en BO2), twee boeken van Guèrard 2 (GU1 en GU2), Maximin (MAX), Olympe (OLY), Troisgros (TRO), Vergé (VER) en drie regionale boeken uit Bretagne (BRE), de Périgord (PER) en France-occidentale (OCC). Deze boeken vormen de kolommen van Tabel 11.6 waarin de resultaten van de tellingen zijn weergegeven. De rijen in deze tabel corresponderen met de volgende onderwerpen: (1) sauzen en soepen, (2) hors d'oeuvres, patés, quiches, galettes en pizza's, (3) fruits de mer, schaaldieren, slakken en kikkers, (4) eieren, (5) vissen, (6) gevogelte en ganzenlever, (7) wild en (tam) konijn, (8) vlees, (9) groenten en pasta, (10) gebak, pannenkoeken en wafels, (11) ijs, pudding en soufflés, (12) fruit.

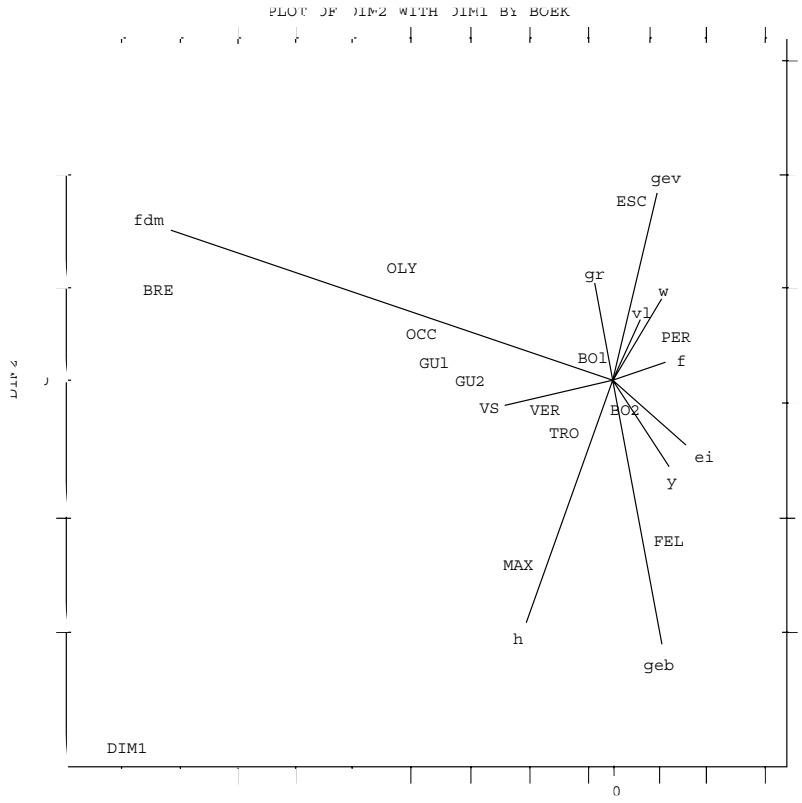
6 ANACOR op een $r \times c$ kruistabel is identiek aan HOMALS op twee nominale variabelen met respectievelijk r en c categorieën

7 Deze methode van inhoudsanalyse is ontleend aan Gifi (1990) die op deze manier de verschillen en overeenkomsten tussen een aantal leerboeken over multivariate analyse onderzocht.

Tabel 11.6 Aantallen bladzijden per kookboek per onderwerp

	Kookboeken													
	ESC	PEL	BRE	OCC	PER	BO1	BO2	GU1	GU2	MAX	OLY	TRO	VER	
1 Saus	200	262	30	21	42	27	146	24	42	15	12	23	23	867
2 Hors d'euvres	112	392	34	31	15	22	67	22	30	30	14	14	13	796
3 Fruits de mer	60	62	62	19	3	3	30	16	14	4	17	5	8	303
4 Eieren	107	174	8	11	23	14	58	1	4	2	4	8	6	420
5 Vis	175	272	60	38	25	27	60	13	11	21	14	14	11	741
6 Gevogelte	276	185	12	15	32	11	56	17	14	10	17	7	6	658
7 Wild	127	109	5	19	24	7	53	5	1	10	6	10	4	380
8 Vlees	386	388	36	33	51	36	146	16	15	14	25	17	11	1174
9 Groenten	307	259	48	23	58	45	179	20	20	14	10	21	22	1026
10 Gebak	63	315	19	13	43	12	119	11	2	14	7	13	9	640
11 IJs	209	305	5	6	10	8	81	17	15	18	9	7	16	706
12 Fruit	97	104	2	5	6	5	44	2	10	6	4	5	8	298
	2119	2827	321	234	332	217	1039	164	178	158	139	144	137	8009

Correspondentieanalyse van deze tabel levert onderstaande afbeelding (zie Figuur 11.7) in twee dimensies die samen 68 procent van de totale χ^2-waarde verklaren. In deze grafiek zijn de kookboeken als punten afgebeeld en de onderwerpen als vectoren. We zien dat er in het plaatje een soort driehoek gevormd wordt met *fruits de mer*, gebak en *hors d'oeuvres*, en gevogelte als hoekpunten. Sommige kookboeken laten een duidelijke voorkeur voor een van deze onderwerpen zien. Enerzijds ging het in dit onderzoek om een poging 'd'établir une liaison entre la haute cuisine et l'analyse des correspondances' (Van der Kloot & Willemsen, 1985, p. 131). Anderzijds ging het erom de hypothese te onderzoeken dat wat men in de jaren tachtig als de *nouvelle cuisine* betitelde, gekenmerkt werd door een grotere nadruk op voor- en tussengerechten en minder op vleesgerechten. Deze hypothese wordt slechts ten dele bevestigd. Weliswaar besteden de boeken van Olympe, Guérard, Vergé, Troisgros en Maximin meer of zelfs aanzienlijk meer aandacht aan, respectievelijk, *fruits de mer*, vis en *hors d'oeuvres*, maar de boeken van Bocuse, die zichzelf als exponent van de *nouvelle cuisine* beschouwt, blijken veel traditioneler en meer algemeen van aard te zijn.



Figuur 11.7 Gezamenlijke afbeelding van kookboeken en onderwerpen

11.5 HET ONTVOUWINGSMODEL

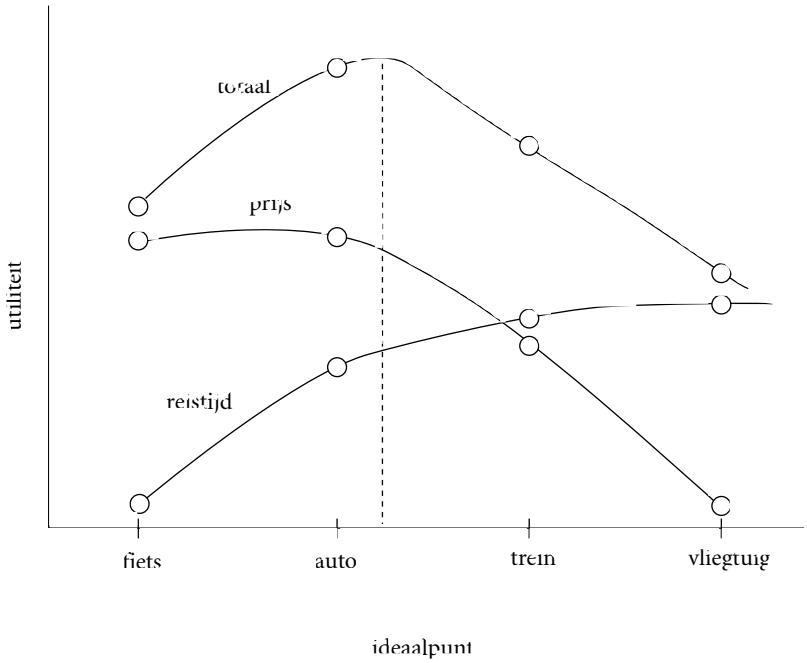
Om een matrix met preferentierangordeningen te analyseren kan men in plaats van het vectormodel ook het zogenaamde *ontvouwingsmodel* of *ideaalpuntmodel* gebruiken. Ook in dit geval probeert men een *joint plot* te maken, dat wil zeggen, een multidimensionale afbeelding waarin de personen en objecten gezamenlijk zijn gerepresenteerd. Het verschil met het vectormodel is dat de personen nu als *(ideaal)punten* en niet als richtingen worden afgebeeld. In dit model wordt aangenomen dat het rangordecijfer van de preferentie van persoon *i* voor object *j* een monotone functie is van de *afstand* tussen het ideaalpunt van *i* en het objectpunt van *j*, zodat
$o_{ij} = f(d_{ij})$ en dus ook

$$g(o_{ij}) = d_{ij} \qquad [11.11]$$

De keuze van het ideaalpuntmodel voor het afbeelden van preferenties berust vooral op een aantal inhoudelijke, theoretische aannamen over keuzegedrag die verschillen van de uitgangspunten van het vectormodel. In Hoofdstuk 7 hebben we al gezien dat we het vectormodel als een speciaal geval van het ideaalpuntmodel kunnen beschouwen, namelijk het geval waarin het ideaalpunt van een persoon in het oneindige ligt. Het vectormodel is dus voor te stellen als een ideaalpuntmodel waarin de personen extreem aan de buitenkant van de configuratie moeten liggen. Daarentegen kunnen in de algemene versie van het ideaalpuntmodel de personen in principe iedere willekeurige plaats in de configuratie innemen. Ze kunnen aan de buitenkant liggen, maar ook centraal tussen de objectpunten in.
Het vectormodel impliceert dus ook dat de meest en de minst geprefereerde objecten aan de buitenkant van de configuratie zullen liggen. Alleen dan kunnen ze op een bepaalde vector de hoogste, respectievelijk laagste projecties hebben. Bij het ideaalpuntmodel is dat niet zo. Er hoeft alleen maar te gelden dat de meest geprefereerde objecten dicht bij de ideaalpunten van de desbetreffende personen liggen. Waar de minst geprefereerde objecten liggen, doet er dan niet toe, als ze maar ver van die ideaalpunten af liggen. Dit impliceert dat een object dat door een groep personen gemiddeld hoog gewaardeerd wordt, in het midden van de configuratie zal moeten liggen. Objecten die door sommigen wel en door anderen niet gewaardeerd worden, zullen dus meer aan de buitenkant liggen.

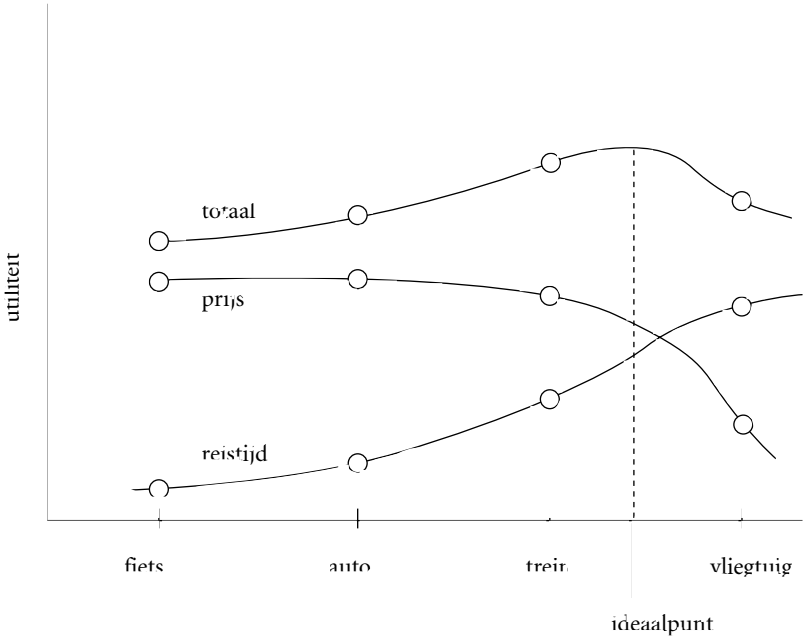
Twee basisaannamen

Het ideaalpuntmodel berust op twee basisaannamen. Deze zullen we aan de hand van een voorbeeld proberen uit te leggen. Stel, we willen op vakantie naar Zuid-Frankrijk en we moeten beslissen hoe we daar naartoe zullen gaan. De mogelijkheden zijn fiets, auto, trein en vliegtuig. Het goede van fietsen is onder andere dat het goedkoop is; de auto is duurder, de trein nog duurder en vliegen is het allerduurst. Het nadeel van fietsen is dat het erg lang duurt om in Zuid-Frankrijk te komen. De auto en de trein zijn sneller en vliegen is uiteraard het allersnelst (op die afstand). We hebben nu dus twee eigenschappen, prijs en snelheid, met bijbehorende utiliteitswaarden die op een monotone maar tegengestelde manier met de vier vervoerstypen verbonden zijn. In Figuur 11.8 zijn twee curven getekend die de utiliteiten van respectievelijk prijs en snelheid van de vier vervoermiddelen weergeeft. De utiliteitscurve voor prijs begint hoog (bij fietsen) en daalt tussen auto en trein scherp naar beneden. De curve van snelheid daarentegen begint laag, loopt snel omhoog om naar het eind toe af te vlakken.



Figuur 11.8 Voorbeeld van twee utiliteitsfuncties voor vier vervoerstypen

Uitgaande van deze curves kunnen we een curve voor de totale utiliteit van de vervoermiddelen tekenen, namelijk als de som van de utiliteiten van de afzonderlijke eigenschappen. Deze somcurve heeft ergens een piek in het midden, tussen de auto en de trein in. Aan beide kanten van die piek, dus zowel bij lagere als bij hogere snelheid en prijs, neemt de totale utiliteit en dus de preferentie voor de desbetreffende stimuli geleidelijk af. Naarmate vervoermiddelen (qua prijs en snelheid) minder lijken op het vervoermiddel met de piek, neemt de preferentie verder af. Dit soort curves noemde Coombs (1964) *single peaked preference functions*. De ééntoppigheid volgt uit het feit dat elk object dat men uit een verzameling zou kunnen kiezen zowel een of meer positieve eigenschappen heeft (Coombs & Avrunin, 1980: *the good things*) als een of meer negatieve eigenschappen (*the bad things*). Die eigenschappen zijn op een monotone manier verbonden met de posities die de objecten op de dimensies innemen. Een van de basisaannamen van het ontvouwingsmodel is dat de stimuli zodanig geordend kunnen worden dat de relevante dimensies met *ééntoppige preferentiefuncties* gevonden kunnen worden.

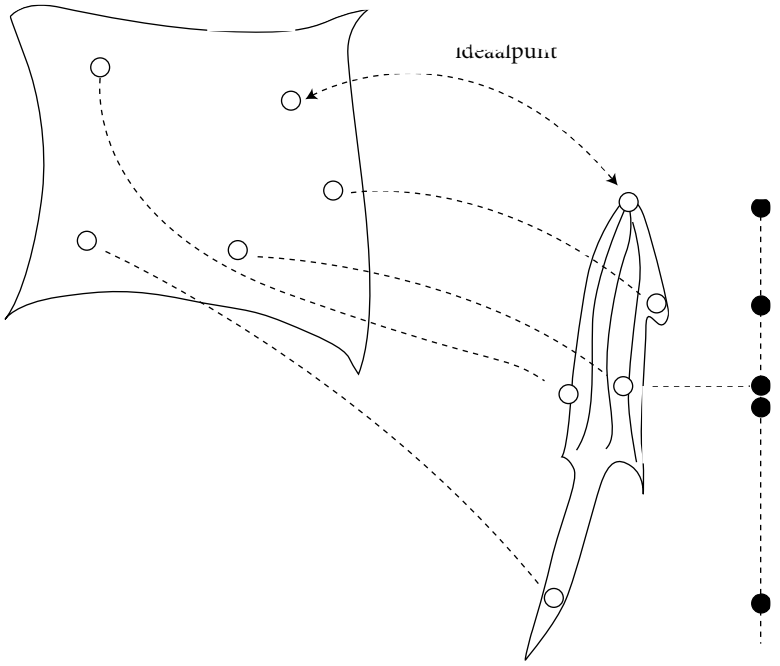


Figuur 11.9 Voorbeeld van twee utiliteitsfuncties van een welvarend persoon voor vier vervoerstypen

De tweede basisaanname is dat de piek van de totale utiliteitscurve voor iedere persoon in principe ergens anders kan liggen. Als geld voor iemand geen rol speelt maar tijd wel, dan loopt de utiliteitscurve voor prijs veel vlakker dan hierboven en de curve voor reistijd veel steiler. De som van beide curven ziet er dan ook anders uit, met een piek die tussen trein en vliegtuig ligt (zie Figuur 11.9).

De gedachte achter het ontvouwingsmodel is nu dat iemands ideaalpunt samenvalt met het punt waar de samengestelde utiliteitscurve het hoogst is. Het 'object' dat het dichtst bij dit ideaalpunt in de buurt ligt, wordt het meest geprefereerd. Naarmate een object – in welke richting dan ook – verder van dit ideaalpunt verwijderd is, geniet het minder de voorkeur en komt het op een latere plaats in de preferentievolgorde. Dit is als volgt aanschouwelijk te maken.

Stel dat zowel het ideaalpunt van Persoon *i* als de punten van een aantal objecten worden voorgesteld door knopen die op een uitgespreide zakdoek zijn genaaid (zie Figuur 11.10). De preferentievolgorde van persoon *i* kunnen we dan te weten komen door die zakdoek in het ideaalpunt beet te pakken en op te tillen. De zakdoek zal dan gaan hangen en als we onze hand van boven naar beneden langs de zakdoek laten glijden, dan voelen we eerst de knoop van het meest geprefereerde object, daarna die van het object dat op de tweede plaats komt, enzovoort.



Figuur 11.10 Het ontstaan van een preferentierangordening door het opvouwen van de onderliggende configuratie

Geven we een aantal personen de opdracht *order m – 1 of m*, dan krijgen we van elke respondent als het ware een opgevouwen zakdoek toegereikt. Ons analyseprobleem bestaat nu uit het opnieuw openvouwen van die zakdoeken; vandaar de naam ontvouwing (*unfolding*). Of eigenlijk: we krijgen niet de zakdoeken zelf in handen, we krijgen alleen maar de beschrijving van de volgorde waarin de knopen van de zakdoek te voelen zijn. Uitsluitend op grond van die volgordegegevens moeten we proberen de vorm van de gemeenschappelijke zakdoek te reconstrueren.

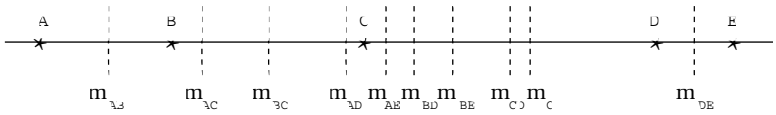
Klassiek ontvouwen (CMDU)

In de meest klassieke vorm van het ontvouwingsprobleem beschikken we over *ordinale* nabijheidsgegevens, die bovendien *conditioneel* zijn per proefpersoon. Immers, het feit dat Object *j* zowel voor Persoon *i* als Persoon *k* op de eerste (of tweede, of derde) plaats komt, impliceert niet dat dit object ook dezelfde *afstanden* ten opzichte van de ideaalpunten van *i* en *k* heeft. Dit probleem kan men op verschillende manieren proberen op te lossen. In de eerste aanpak, die eigenlijk alleen in het eendimensionale geval enigszins werkt, probeert men een configuratie van objecten te vinden die het grootste aantal van de geobserveerde voorkeursrangordeningen kan verklaren. Dit wordt de *combinatorische* aanpak genoemd, die oorspronkelijk door Coombs (1964) bedacht is. Nieuwere, eendimensionale methoden zijn onder andere ontwikkeld door

Van Blokland-Vogelesang (1990) en Van Schuur (1984). Een voorbeeld van Coombs benadering wordt besproken in Blok 11.6.
De tweede aanpak probeert, zoals alle tot nu toe behandelde methoden, een configuratie te vinden door optimalisatie van een kleinste-kwadratencriterium voor de *goodness-of-fit*. Een aantal van deze technieken, die op *numerieke optimalisatie* berusten en ook voor meerdimensionaal ontvouwen geschikt zijn, wordt verderop behandeld.

BLOK 11.6 EENDIMENSIONAAL ONTVOUWEN

In een eendimensionaal ontvouwingsmodel worden stimuli en individuen als punten op een rechte lijn afgebeeld. Als we de plaats van de stimuli op die lijn kennen (dat is in de praktijk natuurlijk niet het geval!) dan is de voorkeursrangorde van een individu uitsluitend afhankelijk van de plaats van zijn of haar ideaalpunt op diezelfde lijn. Stel, er zijn vijf stimuli die in de volgorde ABCDE op de eendimensionale schaal liggen. Een individu met een ideaalpunt dat extreem links op deze schaal ligt, zal dus de voorkeursrangorde ABCDE hebben. Deze rangorde geldt voor alle individuen met een ideaalpunt dat links van het midden van het lijnstuk AB ligt. Het midden van het lijnstuk AB duiden we aan met m_{AB}. Een individu met een ideaalpunt iets aan de rechterkant van m_{AB} ligt nu dichter bij B dan bij A, zodat de voorkeursrangorde die bij dat ideaalpunt hoort, BACDE is. Een ideaalpunt dat niet alleen rechts van m_{AB} maar ook rechts van m_{AC} ligt, correspondeert, als het nog wel links van m_{BC} ligt, met de voorkeursrangorde BCADE. Met andere woorden, iedere keer dat een ideaalpunt een midden 'passeert', levert dat een andere voorkeursrangordening op. De middens van de lijnstukken verdelen de lijn in een aantal segmenten; met elk van die segmenten correspondeert een bepaalde, unieke voorkeursrangorde. Ideaalpunten die in verschillende segmenten liggen, 'hebben' verschillende voorkeursrangordeningen. Ideaalpunten die in hetzelfde segment liggen hebben dezelfde voorkeursrangorde. In Figuur 11.11 is als voorbeeld een eendimensionale schaal met vijf stimuli getekend. Deze stimuli hebben $(5 \times 4)/2 = 10$ middens die de schaal in elf segmenten onderverdelen. Bij deze onderliggende ordening van vijf stimuli zijn er dus elf verschillende voorkeursrangordeningen mogelijk. Van links naar rechts zijn dat: ABCDE, BACDE, BCADE, CBADE, CBDAE, CBDEA, CDBEA, CDEBA, DCEBA, DECBA en EDCBA. Iedere keer dat een ideaalpunt een midden 'passeert', wordt de volgorde van twee stimuli in de rangorde omgewisseld.



Figuur 11.11 Eendimensionale schaal met vijf stimuli en zeven ideaalpunten

Voor het eendimensionale ontvouwingsmodel is de algemene regel: bij m stimuli zijn er $m(m-1)/2$ middens en zijn er dus *maximaal* $m(m-1)/2 + 1$ verschillende voorkeursrangordeningen mogelijk. Indien de stimuli niet eendimensionaal te ordenen zijn, zijn er veel meer voorkeursrangordeningen mogelijk. In het uiterste geval (bij m–1 dimensies!) zelfs $m(m-1)(m-2)(m-3)...(2)(1)$, wat bij vijf stimuli neerkomt op 120 mogelijke rangordeningen.

Zoals gezegd, wanneer de plaats van de stimuli op de eendimensionale schaal bekend is en ook de plaats van iemands ideaalpunt, dan kan de rangordening van die persoon eenvoudig gegenereerd worden door de stimuli op te noemen in de volgorde van hun afstand tot het desbetreffende ideaalpunt. We pakken als het ware de schaal op in het ideaalpunt en gebruiken dat als scharnier om de schaal *dubbel te vouwen*. Beschikken we echter alleen maar over de voorkeursrangordeningen (de zogenaamde *I-schalen*; I slaat op individu) van een aantal personen, dan bestaat ons schaalprobleem juist uit het bepalen van de plaats van de stimuli en de individuen op de onderliggende schaal. We moeten de verzameling voorkeursrangordeningen dus *ontvouwen* om de gezamenlijke schaal (de *J-schaal*; J staat voor *joint*) van stimuli en ideaalpunten te reconstrueren (de terminologie is van Coombs, 1964).

Als een verzameling I-schalen perfect ontvouwbaar is, dan mogen er in de eerste plaats niet meer dan $m(m-1)/2 + 1$ verschillende rangordeningen geobserveerd worden. In de tweede plaats mogen de geobserveerde ordeningen op niet meer dan twee verschillende stimuli eindigen (alle I-schalen uit het genoemde voorbeeld eindigen of op A of op E). In de derde plaats moeten de I-schalen die ontstaan door één of meer stimuli uit de ordeningen weg te laten, ook de twee hierboven genoemde eigenschappen hebben.

Wanneer aan de drie bovengenoemde voorwaarden voldaan is, is het mogelijk de I-schalen eendimensionaal te ontvouwen. Daartoe sporen we eerst de twee stimuli op die aan de uiteinden van de schaal moeten liggen (in het voorbeeld A en E). Vervolgens kijken we of er een I-schaal voorkomt waarin één van deze stimuli als eerste en de andere als laatste komt. Bij precies $m(m-1)/2 + 1$ verschillende I-schalen moeten twee van zulke ordeningen voorkomen die elkaars spiegelbeeld zijn (in ons voorbeeld ABCDE en EDCBE). Met deze twee I-schalen hebben we meteen de onderlig-

gende J-schaal gevonden, dat wil zeggen, de *kwalitatieve* J-schaal: alleen de *volgorde* van de stimuli op de schaal is nu bekend. We moeten de afstanden tussen de stimuli nu zo kiezen dat alle geobserveerde I-schalen verklaard kunnen worden. Op die manier ontstaat een *kwantitatieve* J-schaal. Om uit een kwalitatieve J-schaal een kwantitatieve te verkrijgen, verschuiven we de stimuli zodanig dat hun middens in de 'juiste' volgorde komen te liggen, dat wil zeggen, in de volgorde die zoveel mogelijk I-schalen verklaart. Ook de termen kwalitatieve en kwantitatieve J-schaal stammen van Coombs (1964). Men zou hier beter de termen ordinaal en geordend metrisch kunnen gebruiken (zie Meerling, 1988, 1989).

Een voorbeeld van eendimensionaal ontvouwen

In een onderzoek van Ritsema en Van der Kloot (1980) over cognitieve gelijkheid tussen docenten en studenten werden preferentierangordeningen verzameld met betrekking tot vier stimuli, die uit onderstaande uitspraken bestonden.

Uitspraak E:	*Gedrag wordt veel sterker ingegeven door* EMOTIES *dan door rationele overwegingen.*
Uitspraak O:	*Mensen kunnen in praktisch elke denkbare richting veranderd worden als de* OMGEVING *op de juiste wijze gecontroleerd wordt.*
Uitspraak A:	AANGEBOREN *eigenschappen zijn in sterke mate bepalend voor het soort persoon dat iemand later wordt.*
Uitspraak I:	*De belangrijkste voorwaarde voor mensen om te veranderen is dat ze een helder* INZICHT *in hun situatie hebben.*

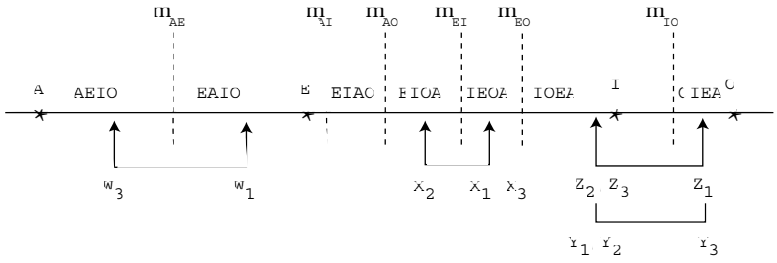
Deze uitspraken werden in drietallen aangeboden aan vier docenten die een tweedejaarscursus psychologie doceerden. De docenten werd gevraagd voor ieder drietal aan te geven met welke hij of zij het meest instemde en met welke uitspraak hij of zij het het minst eens was. Uit deze triadische vergelijkingen werd (door het aantal keren te noteren dat de ene uitspraak een andere domineerde) voor iedere docent een voorkeursrangorde voor de uitspraken afgeleid. In totaal werd deze taak drie keer aan de docenten voorgelegd: eenmaal aan het begin, eenmaal halverwege en één maal aan het eind van de cursus. Omdat één docent één keer niet in staat was aan de vergelijkingstaak mee te doen, zijn er in totaal elf voorkeursrangordeningen verzameld. Deze staan vermeld in Tabel 11.7.

Tabel 11.7 Voorkeursrangordeningen voor vier psychologische uitspraken

Docent	tijdstip	rangorde	Docent	tijdstip	rangorde
X	1	IEOA*	Z	1	OIEA
X	2	EIOA*	Z	2	IOEA
X	3	IEOA	Z	3	IOEA
Y	1	IOEA*	W	1	EAIO*
Y	2	IOEA	W	2	-
Y	3	OIEA*	W	3	AEIO*

* de zes verschillende antwoordpatronen die voorkomen.

In Tabel 11.7 is te zien dat er slechts zes verschillende voorkeursrangordeningen gegeven zijn (deze zijn met een sterretje aangeduid). Dit maakt het aannemelijk dat de docenten en de uitspraken op één onderliggende J-schaal te ordenen zijn. Vier van de zes verschillende rangordeningen eindigen op A en de overige twee op O. Dat suggereert dat A en O de uiteinden van de gezochte J-schaal vormen. Een van de I-schalen, AEIO, begint met A en eindigt met O, waardoor dit waarschijnlijk de gezochte J-schaal is.
Merk op dat er van deze J-schaal twee varianten mogelijk zijn: een waarin het midden van EI (m_{EI}) links van het midden van AO (m_{AO}) ligt en een waarin m_{EI} juist rechts van m_{AO} ligt. Het verschil tussen beide varianten is dat de eerste de rangordening IEAO verklaart, terwijl de tweede dat niet doet. De tweede variant verklaart daarentegen wel de rangordening EIOA. Omdat IEAO niet in de observaties voorkomt en EIOA wel, is het duidelijk dat we de J-schaal zo moeten construeren dat m_{EI} rechts van m_{AO} ligt. Zo'n schaal is weergegeven in Figuur 11.12.



Figuur 11.12 J-schaal van vier psychologische uitspraken en zeven ideaalpunten

De afgebeelde schaal heeft zeven segmenten. Wanneer we de voorkeursrangordeningen bekijken voor ideaalpunten in deze segmenten, dan zien we dat deze J-schaal respectievelijk de I-schalen AEIO, EAIO, EIAO, EIOA, IEOA, IOEA en OIEA op kan leveren.Van deze rangordeningen zijn er zes geobserveerd en één (EIOA) niet.
De gevonden J-schaal lijkt iets te meten als 'basisovertuiging met betrekking tot het *nature-nurture* probleem'. Docenten links op de schaal geloven sterker in de rol van aangeboren eigenschappen en emoties. Individuen rechts op de schaal hechten meer belang aan omgevingsinvloeden en aan inzicht. Een van de aardige dingen die Figuur 11.12 laat zien, is dat de docenten, die over de hele cursusperiode niet steeds dezelfde rangordening opleverden, qua ideaalpunt steeds hoogstens één segment over de schaal heen en weer schoven.

11.6 NUMERIEKE OPLOSSINGEN VOOR HET ONTVOUWINGSPROBLEEM

Het hierboven geschetste ontvouwingsprobleem (volledig: *classical multidimensional unfolding*, CMDU) zou in principe met een MDS-programma zoals ALSCAL aangepakt kunnen worden. Het gaat dan om een matrix met n rijen (meestal personen) en m kolommen (meestal objecten) met *rijconditionele* nabijheidsgegevens die op *ordinaal niveau* gemeten zijn. Met ALSCAL of een ander geschikt programma proberen we een oplossing te vinden met coördinaten voor de rijen (de ideaalpunten) en coördinaten voor de kolommen (de objectpunten). Daarbij moet gelden dat

$$d_{ij} \approx g_i(o_{ij}) \qquad [11.12]$$

waarbij '≈' betekent dat de coördinaten $\{x_{is}\}$ en $\{x_{js}\}$ zodanig gekozen worden dat de gebruikte stressfunctie geminimaliseerd wordt. Hieronder volgt de typische ALSCAL-aansturing van het CMDU-probleem, waarbij we gebruikmaken van de data van Ritsema en Van der Kloot (1980), die in Blok 11.6 beschreven zijn.

```
DATA LIST /PERSOON 1 A E I O 3-6.
begin data.
1 1234 (AEIO)
2 2134 (EAIO)
3 4123 (EIOA)
4 4213 (IEOA)
5 4312 (IOEA)
```

```
6 4321 (OIEA)
end data.
ALSCAL variables=A E I O
  /shape=rectangular
  /model=euclid
  /level=ordinal
  /condition=row.
```

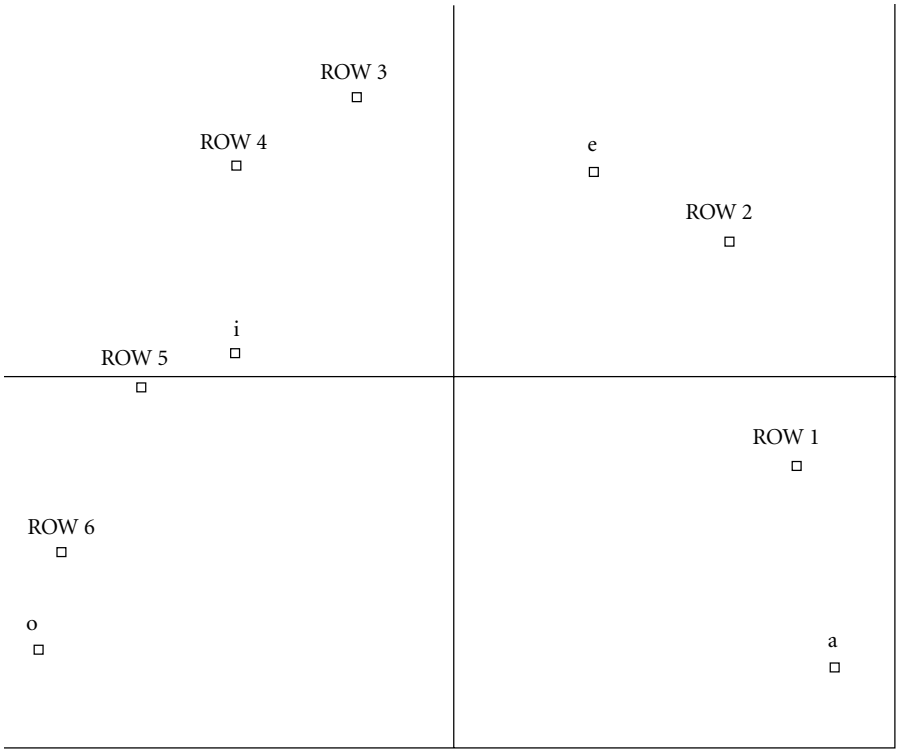
Bij de aansturing van een MDS-programma moet men er goed op letten wat de observaties precies voorstellen. Is 1 het rangcijfer van het meest geprefereerde of van het minst geprefereerde object? Bij bovenstaande data geeft 1 de meeste preferentie aan. De observaties zijn dus *dis*similarities die op een monotoon stijgende manier met afstanden gerelateerd zijn (hoe groter het rangcijfer des te groter de afstand tot het ideaalpunt). In het omgekeerde geval, dus wanneer het kleinste rangcijfer de minste voorkeur aanduidt, is er dus sprake van een monotoon dalende relatie met afstand. De observaties zijn dan *sim*ilarities, wat aan ALSCAL moet worden meegedeeld met het subcommando `/level = ordinal (similar)`.
De ALSCAL-analyse van de data van Ritsema en Van der Kloot levert onderstaande uitvoer:

```
-> ALSCAL
-> VARIABLES= a e i o
->    /SHAPE=RECTANGULAR /INPUT ROWS(6)
->    /LEVEL=ORDINAL
->    /CONDITION=ROW
->    /MODEL=EUCLID
->   /CRITERIA=CONVERGE(.0001) STRESSMIN(.005) ITER(75) CUTOFF(0)        DIMENS(2,2)
->    /PLOT=DEFAULT
->    /PRINT=DATA HEADER .


Alscal Procedure Options

Data Options-
Number of Rows (Observations/Matrix).    6
Number of Columns (Variables) . . .    4
Number of Matrices  . . . . . .   1
Measurement Level . . . . . .    Ordinal
Data Matrix Shape . . . . . .    Rectangular
Type . . . . . .  Dissimilarity
Approach to Ties . . . . . .    Leave Tied
Conditionality . . . . . .   Row
Data Cutoff at . . . . . .    .000000

Model Options-
Model . . . . . .  Euclid
Maximum Dimensionality . . . . .   2
Minimum Dimensionality . . . . .   2
Negative Weights . . . . . .   Not Permitted

Output Options-
```

```
Job Option Header . . . . . .   Printed
Data Matrices . . . . . .   Printed
Configurations and Transformations .   Plotted
Output Dataset . . . . . .   Not Created
Initial Stimulus Coordinates . . .   Computed
Initial Column Stimulus Coordinates .   Computed

Algorithmic Options-
Maximum Iterations . . . . . .   75
Convergence Criterion . . . . .   .00010
Minimum S-stress . . . . . .   .00500
Missing Data Estimated by . . . .   Ulbounds
Tiestore . . . . . 100


          Raw (unscaled) Data for Subject 1

             1      2      3      4
        1   1.000  2.000  3.000  4.000
        2   2.000  1.000  3.000  4.000
        3   4.000  1.000  2.000  3.000
        4   4.000  2.000  1.000  3.000
        5   4.000  3.000  1.000  2.000
        6   4.000  3.000  2.000  1.000


Iteration history for the 2 dimensional solution (in squared distances)
Young's S-stress formula 2 is used.

          Iteration     S-stress      Improvement
           1          .00000

              Iterations stopped because
              S-stress is less than   .005000

            Stress and squared correlation (RSQ)

                  Matrix   1
                (Row Stimuli Only)
    Stimulus    Stress      RSQ   Stimulus    Stress      RSQ

       1        .000    1.000       2        .000    1.000
       3        .000    1.000       4        .000    1.000
       5        .000    1.000       6        .000    1.000

    Averaged (rms) over stimuli
  Stress  =   .000      RSQ = 1.000

      Configuration derived in 2 dimensions
           Stimulus Coordinates
              Dimension
Stimulus  Stimulus    1       2
 Number      Name
 Column
   1     A          1.8048  -1.2256
   2     E           .6647    .8206
   3     I          -.6966    .1065
   4     O         -1.3995  -1.0938
 Row
   1               1.6453   -.3886
   2               1.2089    .5308
   3               -.1495   1.1216
   4               -.6754    .8552
   5              -1.0824   -.0267
```

```
6              -1.3202   -.7000

Optimally scaled data (disparities) for subject   1

     1       2       3       4
1     .852    1.557   2.394   3.125
2    1.855     .617   1.952   3.073
3    3.054     .868   1.153   2.544
4    3.237    1.340    .749   2.079
5    3.126    1.942    .408   1.113
6    3.169    2.500   1.019    .402
```

Uit bovenstaande uitvoer blijkt dat ALSCAL in één iteratie een perfecte oplossing heeft gevonden, die precies overeenkomt met de 'ware' structuur van de data die we in Blok 11.6 hebben besproken. De tweedimensionale configuratie is weergegeven in Figuur 11.13. In deze figuur is het bekende hoefijzer te herkennen, wat erop wijst dat de configuratie in feite eendimensionaal is. Zowel de stimuli als de personen liggen in de verwachte volgorde op het hoefijzer.
Aan de tabel met pseudo-afstanden zien we dat de rijconditionele transformatieoptie zinvol is geweest, omdat iedere proefpersoon een (iets) andere trans-



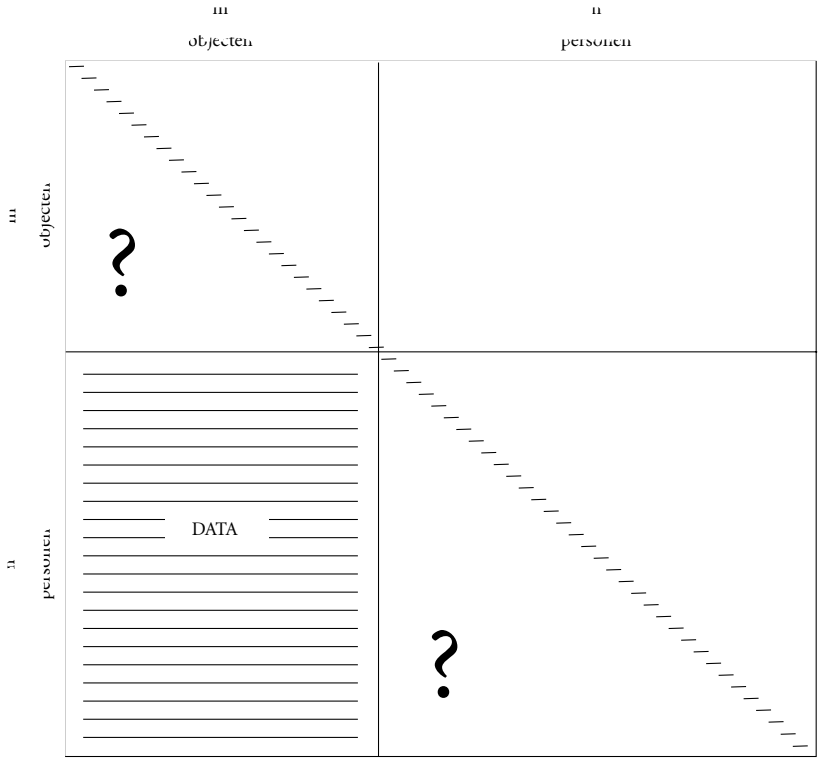
formatie van zijn of haar rangordecijfers heeft gekregen.

Figuur 11.13 ALSCAL-oplossing van een eendimensionaal ontvouwingsprobleem

Dat we zo'n mooie oplossing gekregen hebben, ligt aan het feit dat het hier om een klein voorbeeld (weinig personen, weinig stimuli) gaat, met *error*-loze data die perfect door middel van een eendimensionale schaal zijn weer te geven. Dat is natuurlijk niet altijd het geval. In de volgende paragrafen bespreken we een aantal oorzaken waardoor ontvouwing door middel van ALSCAL (of een ander MDS-programma) soms op een mislukking uit kan lopen.

Problemen van CMDU

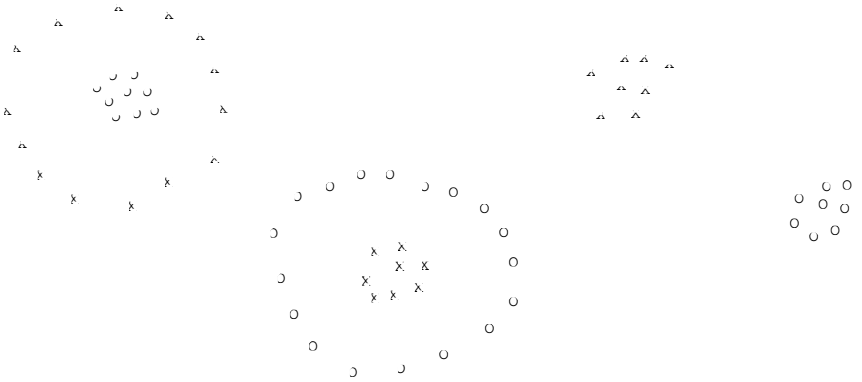
Hoewel bovengenoemde benadering in principe niet afwijkt van het klassieke MDS-probleem van Hoofdstuk 6, lukt het in de praktijk vaak niet om een goede oplossing te vinden. De reden daarvan is dat het CMDU-probleem eigenlijk een CMDS-probleem is van een nabijheidsmatrix met een groot aantal, volgens een systematisch patroon ontbrekende, waarden. We willen in feite coördinaten vinden van $n + m$ punten, die samen $(n + m)(n + m - 1)/2 = (n^2 + m^2 + 2nm - n - m)/2$ onderlinge afstanden hebben. Over slechts nm van deze onderlinge afstanden zijn gegevens geobserveerd; gegevens over de $n(n-1)/2$ onderlinge afstanden van de ideaalpunten en de $m(m-1)/2$ onderlinge afstanden van de



objecten ontbreken volledig (zie Figuur 11.14)!

Figuur 11.14 Patroon van geobserveerde en ontbrekende nabijheidsdata bij een CMDU-probleem

Dit gevoegd bij het feit dat de gegevens die wél geobserveerd zijn uit rijconditionele rangordeningen bestaan, maakt dat het heel moeilijk is een geschikte oplossing te vinden. Vaak vindt een programma als ALSCAL zogenaamde *gedegenereerde oplossingen*, dat zijn oplossingen die wel een goede *fit* hebben, maar een onzinnig of niet-informatief plaatje opleveren. Enkele voorbeelden staan



in Figuur 11.15.

Figuur 11.15 Voorbeelden van gedegenereerde unfolding-oplossingen

In de linkerfiguur zien we dat alle objecten op een kluitje in het midden terecht zijn gekomen en dat de personen er in een cirkel omheen liggen. In de middelste figuur zien we het omgekeerde: de ideaalpunten van de personen liggen allemaal in het midden en de objecten liggen daar in een cirkel omheen. De meest rechtse figuur laat een andere gedegenereerde oplossing zien: alle objecten op een kluitje aan de ene kant en alle personen op een kluitje aan de andere kant van de configuratie. Deze figuren, die ook in allerlei mengvormen voor kunnen komen, zeggen dus niets meer over de precieze relaties tussen de personen en de objecten. Evenmin bevatten ze zinvolle informatie over de relaties tussen de personen onderling en tussen de objecten.

Het probleem met gedegenereerde oplossingen is niet dat een programma als ALSCAL geen goede oplossing kan vinden, maar juist dat het veel te gemakkelijk is om een goede, dat wil zeggen een oplossing met lage (s)stress, te verkrijgen. Het grote aantal ontbrekende gegevens plus het ordinale meetniveau en de rijconditionaliteit van de geobserveerde gegevens geven het MDS-programma veel te veel vrijheid om de coördinaten van de punten te kiezen.

Een oplossing voor bovengenoemde problemen moeten we dan ook zoeken in het opleggen van meer en/of stringentere *restricties* aan de oplossing. Bijvoorbeeld: we kunnen de rijconditionaliteit van de data laten vervallen, dat wil zeggen de data analyseren alsof zij matrixconditioneel zijn. Of: we kunnen de ordinaliteit laten vallen en doen alsof de data op intervalniveau of rationiveau gemeten zijn. In ALSCAL kunnen we bij zo'n metrische analyse nog kiezen tus-

sen rechtlijnige of curvilineaire transformatiefuncties. Ten slotte kunnen we tegelijkertijd de conditionaliteit en het meetniveau aanpassen. We weten dan wel dat we onze observaties misschien enigszins geweld aandoen (wat ook stressverhogend werkt) maar ook dat we zonder deze kunstgrepen een niet te interpreteren oplossing krijgen. We moeten dus kiezen uit twee of meer kwaden.
Een andere manier om restricties aan de oplossing op te leggen is door eisen te stellen aan de plaats waar iemands ideaalpunt terecht moet komen. Men kan bijvoorbeeld eisen dat (a) iemands ideaalpunt samenvalt met het object dat hij of zij het meest prefereert, of (b) dat dit ideaalpunt in de centroïde ligt van de objecten die de grootste en de op-één-na-grootste voorkeur hebben, of (c) dat het in de centroïde ligt van de objecten die als eerste, als tweede en als derde gekozen zijn, enzovoort. Deze mogelijkheden om restricties in te voeren zitten niet in ALSCAL, maar wel in het programma PROXSCAL van Heiser (1988) en Busing, Commandeur en Heiser (1997). Dit laatste programma heeft ook de mogelijkheid additionele, *externe* gegevens over de objecten of de personen in de analyse 'mee te nemen'. In dat geval is de configuratie niet uitsluitend op de voorkeursrangordeningen gebaseerd, maar ook op die extra variabelen.
Opgemerkt moet worden dat degeneratieproblemen zich meer voordoen naarmate de data meer *error* bevatten en naarmate het aantal dimensies dat voor de oplossing gekozen is (veel) kleiner is dan de 'werkelijke' dimensionaliteit. Ook nemen deze problemen toe naarmate de datamatrix rechthoekiger wordt, dat wil zeggen, naarmate het aantal rijen (veel) groter of kleiner wordt dan het aantal kolommen van de matrix.

Andere oplossingen

Aangezien het ideaalpuntmodel op zichzelf een elegant model is, dat om inhoudelijke redenen vaak de voorkeur boven het vectormodel verdient, zou het jammer zijn als we het om bovengenoemde praktisch-technische redenen niet zouden kunnen gebruiken. Hieronder worden drie manieren besproken waarop men een matrix met voorkeursrangordeningen toch volgens het ideaalpuntmodel kan analyseren. In de eerste twee methoden wordt de informatie die in de voorkeursrangordeningen aanwezig is, gebruikt om een andersoortige datamatrix te creëren, die vervolgens met HOMALS geanalyseerd wordt. In de derde methode is er sprake van een '*quick and dirty*'-aanpak, waarin de preferentierangordeningen worden omgezet in keuzedata.

HOMALS op paarsgewijze vergelijkingen

In de data van Ritsema en Van der Kloot (1980) heeft de derde proefpersoon voor de stimuli A, E, I en O de voorkeursrangorde EIOA gegeven. Dat betekent dus dat E geprefereerd wordt boven I, O en A, dat I geprefereerd wordt boven O en A, en dat O geprefereerd wordt boven A. Als de ene stimulus boven een andere geprefereerd wordt, kunnen we dat interpreteren als 'de eerste stimulus domineert de ander in aantrekkelijkheid of utiliteit'. Voor de persoon met voorkeursrangorde EIOA zouden we dus een matrix met enen en nullen kunnen

maken die er als volgt uitziet:

	A	E	I	O
A	–	0	0	0
E	1	–	1	1
I	1	0	–	1
O	1	0	0	-

Een 1 in deze matrix betekent dat de rijstimulus de kolomstimulus domineert. Een dergelijke dominantiematrix kunnen we voor elke persoon construeren. Zouden we die matrices bij elkaar optellen, dan krijgen we het soort data – zogenaamde paarsgewijze vergelijkingen – dat we in Hoofdstuk 1 volgens de methode van Thurstone geanalyseerd hebben. Echter, met die methode kregen we geen inzicht in mogelijke individuele verschillen tussen de proefpersonen, wat we in dit geval nu juist wel willen. Daarom construeren we een supermatrix waarin de enen en nullen van één persoon naast elkaar en van de verschillende personen onder elkaar zijn neergezet. Die matrix ziet er als volgt uit:

Persoon	AE	AI	AO	EI	EO	IO
1 (AEIO)	1	1	1	1	1	1
2 (EAIO)	0	1	1	1	1	1
3 (EIOA)	0	0	0	1	1	1
4 (IEOA)	0	0	0	0	1	1
5 (IOEA)	0	0	0	0	0	1
6 (OIEA)	0	0	0	0	0	0

De kolommen van deze matrix komen overeen met alle verschillende paren die uit de stimuli te vormen zijn. Een 1 in deze matrix betekent dat de eerstgenoemde stimulus van het paar geprefereerd is boven de tweede, een 0 betekent dat de tweede stimulus boven de eerste wordt verkozen. Deze matrix kunnen we met HOMALS analyseren (zie Heiser, 1981), maar omdat nullen voor HOMALS ontbrekende gegevens zijn moeten we eerst de gegevens hercoderen, bijvoorbeeld: `RECODE AE TO IO (0=2)`. De data hebben dan de betekenis gekregen: 1 = eerste stimulus boven tweede verkozen; 2 = tweede stimulus boven eerste geprefereerd.
HOMALS op zo'n soort datamatrix geeft een configuratie van personen (de objectscores) en van de categorieën 1 en 2 van de variabelen AE tot en met IO. Categorie 1 van elke variabele ligt in de centroïde van de personen die deze categorie 'gekozen' hebben, categorie 2 ligt in de centroïde van de personen die 2 'geantwoord' hebben. Daarmee hebben we echter nog geen afbeelding van de

afzonderlijke stimuli. De coördinaten van A kunnen we achteraf bepalen door de centroïden te berekenen van de 1-categorieën van AE, AI en AO. De locatie van E is de centroïde van de 2-categorie van AE en de 1-categorieën van EI en EO. De posities van de overige stimuli worden op analoge manier bepaald. Een

BLOK 11.7 ONTVOUWEN MET HOMALS: PAARSGEWIJZE CODERINGEN

voorbeeld wordt behandeld in Blok 11.7.
Hieronder volgen de SPSS-commando's waarmee de rangordedata van Ritsema en Van der Kloot (zie Blok 11.6) worden omgezet in paarsgewijze dominantiedata.

```
data list/persoon 1 A E I O 3-6.
begin data.
1 1234
2 2134
3 4123
4 4213
5 4312
6 4321
end data.
compute AE=0.
compute AI=0.
compute AO=0.
compute EI=0.
compute EO=0.
compute IO=0.
if (A lt E) AE=1.
if (A lt I) AI=1.
if (A lt O) AO=1.
if (E lt I) EI=1.
if (E lt O) EO=1.
if (I lt O) IO=1.
list /variables persoon AE to IO.
```

Dit deel van het SPSS-programma levert onderstaande datamatrix met enen en nullen op die na hercodering met HOMALS geanalyseerd gaat worden.

PERSOON	AE	AI	AO	EI	EO	IO
1	1.00	1.00	1.00	1.00	1.00	1.00
2	.00	1.00	1.00	1.00	1.00	1.00
3	.00	.00	.00	1.00	1.00	1.00
4	.00	.00	.00	.00	1.00	1.00
5	.00	.00	.00	.00	.00	1.00
6	.00	.00	.00	.00	.00	.00

De rest van het SPSS-programma luidt:

```
recode AE to IO (0=2).
HOMALS variables AE to IO (2)
 /dimensions 2
 /print default object
 /plot discrim object quant
 /save object (2).
```

De belangrijkste uitvoer van dit programma is:

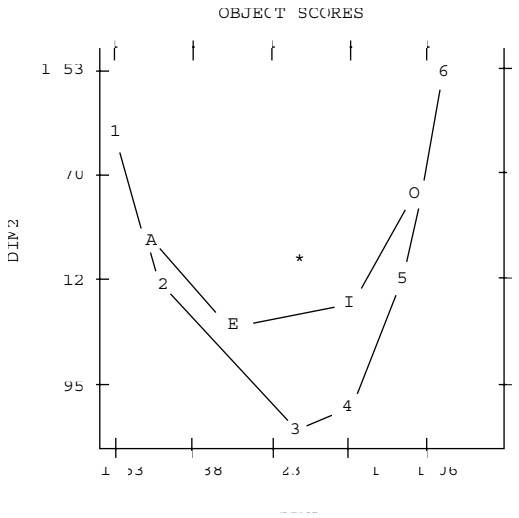
```
THE OBJECT SCORES ARE:
======================
OBJECT *     DIMENSION
           1      2
   1 *   -1.53     .97
   2 *   -1.04    -.07
   3 *    -.02   -1.27
   4 *     .44   -1.05
   5 *     .87    -.11
   6 *    1.28    1.53
----------------------

CATEGORY QUANTIFICATIONS
================================================
 VARIABLE: AE
 -------
 CATEGORY   DIMENSIONS
 -------  ---------
           1      2
    1    -1.53     .97
    2     .31    -.19
============================
 VARIABLE: AI
 -------
 CATEGORY   DIMENSIONS
 -------  ---------
           1      2
    1    -1.28     .45
    2     .64    -.23
============================
```

```
VARIABLE: AO
--------
CATEGORY    DIMENSIONS
--------  ----------
            1       2
     1    -1.28     .45
     2     .64    -.23
============================
 VARIABLE: EI
---------
 CATEGORY    DIMENSIONS
 --------  ----------
            1       2
     1     -.86    -.12
     2      .86     .12
============================
 VARIABLE: EO
---------
 CATEGORY    DIMENSIONS
---------  ----------
            1       2
     1     -.54    -.35
     2     1.07     .71
============================
 VARIABLE: IO
---------
 CATEGORY    DIMENSIONS
---------  ----------
            1       2
     1     -.26    -.31
     2     1.28    1.53
============================
```

Om vervolgens de coördinaten $\{y_{js}\}$ van de stimuli te berekenen gebruiken we de categoriekwantificaties $\{q_{vcs}\}$ van de variabelen AE tot en met IO (v duidt de betreffende variabele aan, c de categorie en s de dimensie). Dus:

$y_{A1} = (q_{AE11} + q_{AI11} + q_{AO11})/3 = (-1.53 - 1.28 - 1.28)/3 = -1.37;$
$y_{A2} = (q_{AE12} + q_{AI12} + q_{AO12})/3 = (-.97 + .45 + .45)/3 = -.02;$
$y_{E1} = (q_{AE21} + q_{EI11} + q_{EO11})/3 = (.31 - .86 - .54)/3 = -.36;$
$y_{E2} = (q_{AE12} + q_{EI12} + q_{EO12})/3 = (-.19 - .12 - .35)/3 = -.22;$
$y_{I1} = (q_{AI21} + q_{EI21} + q_{EO11})/3 = (.64 + .86 - .26)/3 = .41;$
$y_{I2} = (q_{AI22} + q_{EI22} + q_{EO12})/3 = (-.23 + .12 - .31)/3 = -.14;$
$y_{O1} = (q_{AO21} + q_{EO21} + q_{IO21})/3 = (.64 + 1.07 + 1.28)/3 = .77;$
$y_{A2} = (q_{AO12} + q_{EO22} + q_{IO22})/3 = (-.23 + .71 + 1.53)/3 = .67.$

Tekenen we deze punten in de grafiek met objectscores, dan ontstaat onderstaande figuur waarin we de tot hoefijzer verbogen eendimensionale schaal van stimuli en personen herkennen. Gezien deze hoefijzervorm is het voldoende om alleen de coördinaten op de eerste dimensie te gebruiken.




```
==============================
 VARIABLE: EI
--------
 CATEGORY    DIMENSIONS
 --------  ----------
           1       2
     1    -.86    -.12
     2     .86     .12
==============================
 VARIABLE: EO
--------
 CATEGORY    DIMENSIONS
--------  ----------
           1       2
     1    -.54    -.35
     2    1.07     .71
==============================
 VARIABLE: IO
---------
```

In het geval dat er meer stimuli zijn, hebben de eerste drie variabelen meer categorieën en moet er ook een vierde nieuwe variabele geconstrueerd worden (en eventueel een vijfde, zesde, enzovoort). De vierde variabele heeft dan evenveel categorieën als er viertallen zijn (dus: $m(m-1)(m-2)(m-3)/24$) en geeft aan welke vier stimuli het eerste, tweede, derde en vierde rangordenummer hebben. Deze variabele noemen we dan PREF1234.
De variabelen PREF1, PREF12 en PREF123 van ons voorbeeld worden met HOMALS geanalyseerd. Dat levert coördinaten voor de personen (de objectscores) en coördinaten voor alle categorieën van alle variabelen. Als coördinaten voor de stimuli kunnen we nu het beste de categoriekwantificaties van variabele PREF1 nemen, deze liggen immers in de centroïden van de personen die de meeste voorkeur voor de desbetreffende stimuli hebben. Merk op dat niet alle categorieën van alle variabelen ook daadwerkelijk hoeven voor te komen. Het kan best gebeuren dat een bepaald drietal stimuli nooit als eerste, tweede en derde geprefereerd wordt. Voor HOMALS maakt dat niet uit; deze categorieën doen niet mee in de analyse. Ze krijgen allemaal 0.00 als categoriekwantificatie. In Blok 11.8 wordt deze methode toegepast op de gegevens van Ritsema en Van der Kloot (1980).

BLOK 11.8 UNFOLDING MET HOMALS: DE METHODE HEISER

Hieronder geven we de SPSS-programmaregels om de data te hercoderen volgens de methode van Heiser (1981).

```
if (A=1) PREF1=1.
if (E=1) PREF1=2.
if (I=1) PREF1=3.
if (O=1) PREF1=4.
if ((A=1 and E=2) or (A=2 and E=1)) PREF12=1.
if ((A=1 and I=2) or (A=2 and I=1)) PREF12=2.
if ((A=1 and O=2) or (A=2 and O=1)) PREF12=3.
if ((E=1 and I=2) or (E=2 and I=1)) PREF12=4.
if ((E=1 and O=2) or (E=2 and O=1)) PREF12=5.
if ((I=1 and O=2) or (I=2 and O=1)) PREF12=6.
if ((A=1 and E=2 and I=3) or (A=1 and E=3 and I=2) or
   (A=2 and E=1 and I=3) or (A=3 and E=1 and I=2) or
   (A=3 and E=2 and I=1) or (A=2 and E=3 and I=1) PREF123=1.
if ((A=1 and E=2 and O=3) or (A=1 and E=3 and O=2) or
   (A=2 and E=1 and O=3) or (A=3 and E=1 and O=2) or
   (A=3 and E=2 and O=1) or (A=2 and E=3 and O=1) PREF123=2.
if ((A=1 and I=2 and O=3) or (A=1 and I=3 and O=2) or
   (A=2 and I=1 and O=3) or (A=3 and I=1 and O=2) or
   (A=3 and I=2 and O=1) or (A=2 and I=3 and O=1) PREF123=3.
```

```
if ((E=1 and I=2 and O=3) or (E=1 and I=3 and O=2) or
   (B=2 and I=1 and O=3) or (B=3 and I=1 and O=2) or
   (B=3 and I=2 and O=1) or (B=2 and I=3 and O=1) PREF123=4.
*
*  Omdat er maar vier stimuli zijn, hadden de regels voor PREF123
*  vervangen kunnen worden door:
*    if (A=4) PREF123=4.
*    if (E=4) PREF123=3.
*    if (I=4) PREF123=2.
*    if (O=4) PREF123=1.
*
list /variables persoon PREF1 to PREF1234.
```

Bovenstaand gedeelte uit het SPSS-programma levert onderstaande datamatrix op, die met HOMALS geanalyseerd gaat worden.

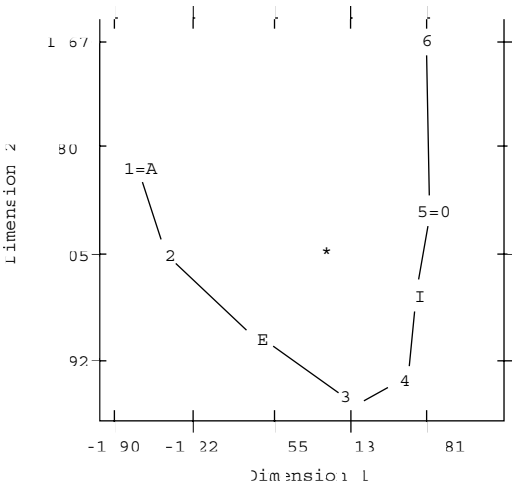
```
PERSOON     PREF1    PREF12   PREF123

    1     1.00     1.00     1.00
    2     2.00     1.00     1.00
    3     2.00     4.00     4.00
    4     3.00     4.00     4.00
    5     3.00     6.00     4.00
    6     4.00     6.00     4.00
```

De rest van het SPSS-programma luidt:

```
HOMALS variables PREF1 (4) PREF12 (6) PREF123 (4)
 /dimensions 2
 /print default object
 /plot discrim object quant.
```

De belangrijkste uitvoer van dit programma is:

```
THE OBJECT SCORES ARE:
======================
OBJECT *     DIMENSION
               1      2
   1 *     -1.58    .64
   2 *     -1.17   -.28
   3 *      .27   -1.32
   4 *      .72   -1.02
   5 *      .86    .36
   6 *      .90   1.61
- - - - - - - - - - - - - -

CATEGORY QUANTIFICATIONS
=========================================
 VARIABLE: PREF1
 - - - - -
 CATEGORY   DIMENSIONS
 - - - -     - - - - -
               1      2
     1    -1.58    .64
     2     -.45   -.80
     3      .79   -.33
     4      .90   1.61
=========================================
```



Deze figuur laat weer het bekende hoefijzer zien, met de stimuli in volgorde AEIO.

Een *quick and dirty*-methode

Een voor de hand liggende mogelijkheid om voorkeursrangordeningen te analyseren is door deze gegevens tot keuzedata te herleiden. Bijvoorbeeld: als iemand vijf stimuli A, B, C, D en E naar preferentie moet rangordenen en A en C als meest en op-één-na meest geprefereerde stimuli aanwijst, dan mogen we aannemen dat deze persoon dezelfde stimuli zou hebben aangewezen, als wij hem of haar gevraagd hadden de twee meest geprefereerde stimuli *uit te kiezen*. Dus door de opdracht **recode** `A B C D E (1,2=1) (3,4,5=0)` maken we van een matrix met voorkeursrangordeningen een matrix met keuzedata. Deze kunnen we met HOMALS analyseren, op de manier die in Hoofdstuk 10 is toegelicht.

Opgemerkt moet worden dat de onderzoeker hier zelf moet bepalen welke rangordecijfers tot enen en welke tot nullen moeten worden gehercodeerd. Nemen we bij vijf stimuli alleen de eerste twee stimuli of nemen we ook de derde erbij? Nemen we bij negen of meer stimuli alleen de eerste twee, de eerste drie of misschien wel de eerste vijf rangordecijfers in de analyse op? Voor het beantwoorden van deze vragen moet men op twee zaken letten. Nemen we alleen de eerste twee keuzen, dan kan er zich een aantal (soms zelfs: veel) unieke antwoordpatronen voordoen. De betreffende personen gaan dan als uitbijters fungeren; zij kunnen de oplossing gaan domineren en een gedegenereerde oplossing veroorzaken. Veranderen we veel meer rangordecijfers in enen (bijvoorbeeld met negen stimuli de eerste vier of eerste vijf), dan wordt de overlap tussen de keuzen van verschillende personen groter. Daardoor kan het gebeuren dat (mogelijk belangrijke) verschillen tussen de personen niet meer in de resultaten zijn terug te vinden. In de praktijk kan men verschillende hercoderingen uitproberen om uit te maken wat de best interpreteerbare oplossing geeft.

In Blok 11.9 behandelen we de aansturing en de uitvoer van een HOMALS-analyse, op alleen de eerste twee keuzen van ons voorbeeld.

BLOK 11.9 HOMALS OP TOT KEUZEDATA HERLEIDE RANGORDENINGEN

In eerste instantie moeten de rangordeningen tot 1/0-data gehercodeerd worden. Dat kan als volgt:

```
recode A E I O (1,2=1) (3,4=0).
list /variables A E I O.
```

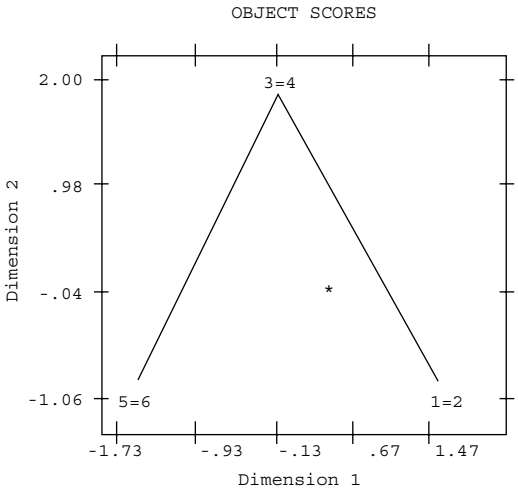
De hercodering levert onderstaande matrix met keuzedata op.

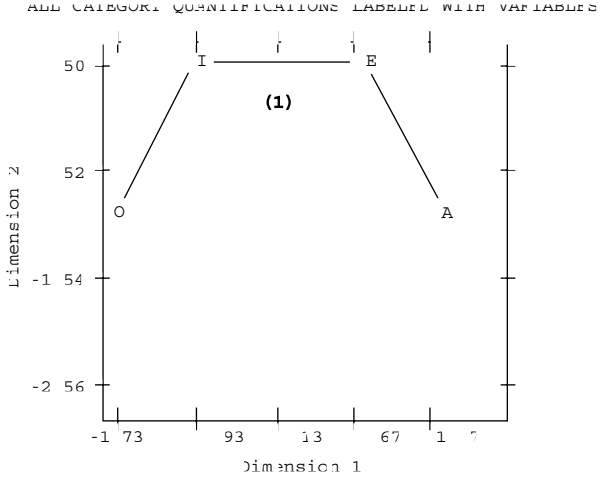
```
A E I O
1 1 0 0
1 1 0 0
0 1 1 0
0 1 1 0
0 0 1 1
0 0 1 1
```

Deze worden vervolgens met HOMALS geanalyseerd.

```
HOMALS variables A E I O (2)
  /dimensions 2
  /print default object
  /plot discrim object quant.

 THE OBJECT SCORES ARE:
 ======================
OBJECT *     DIMENSION
           1      2
   1 *    1.73   -1.00
   2 *    1.73   -1.00
   3 *     .00    2.00
   4 *     .00    2.00
   5 *   -1.73   -1.00
   6 *   -1.73   -1.00
- - - - - - - - - - - - - - - -

CATEGORY QUANTIFICATIONS
============================
 VARIABLE: A
 - - - - -
 CATEGORY   DIMENSIONS
 - - - -     - - - - -
           1      2
    1    1.73   -1.00
    2     .00    .00
============================
```

```
VARIABLE: E
- - - - -
CATEGORY   DIMENSIONS
- - - -     - - - - -
          1     2
    1     .87    .50
    2     .00    .00
==============================
VARIABLE: I
- - - - -
CATEGORY   DIMENSIONS
- - - -     - - - - -
          1     2
    1    -.87    .50
    2     .00    .00
==============================
VARIABLE: O
- - - - -
CATEGORY   DIMENSIONS
- - - -     - - - - -
          1     2
    1   -1.73   -1.00
    2     .00    .00
```





We zien hierboven dus dat er een simplificatie heeft plaatsgevonden. De personen zijn samengeklonterd tot drie groepjes (1 en 2, 3 en 4, en 5 en 6). De ordening van de stimuli is echter wel bewaard gebleven. We zouden de configuratie van de stimuli nu weer kunnen gebruiken om er de personen volgens hun voorkeurs*rangordeningen* op af te beelden. Dat wil zeggen, we kunnen de oorspronkelijke rangordecijfers gebruiken om de groepjes personen weer uiteen te rafelen. Er zijn verschillende manieren waarop dat zou kunnen: (a) we laten de personen samenvallen met hun eerste keuzen, of (b) we gebruiken de coördinaten van de stimuli en de rangordeningen van de proefpersonen om met behulp van *externe unfolding* (zie Hoofdstuk 7) ideaalpunten voor de personen te bepalen. Overigens: in een dataset met meer stimuli en meer personen zal het meestal niet zo erg zijn als er (enige) simplificatie optreedt.