
MDS als cartografisch probleem

2.1 DE KAART VAN NEDERLAND

Hoewel de toepassingen die we in het eerste hoofdstuk besproken hebben een psychologische achtergrond hadden, zijn schaaltechnieken niet alleen maar geschikt voor psychologische of andere sociaal-wetenschappelijke gegevens. Dat moge blijken uit het voorbeeld van een MDS-probleem dat we hieronder gedetailleerd zullen bespreken: het maken van een landkaart, de kaart van Nederland.

Stel dat we beschikken over een tabel met afstanden tussen een aantal Nederlandse steden, bijvoorbeeld Amsterdam, Rotterdam, Den Haag, Utrecht, Eindhoven, Arnhem, Zwolle en Groningen. Om die afstanden uit te drukken zijn er verschillende eenheden mogelijk: niet alleen echte afstandsmaten zoals meters, kilometers en mijlen, maar ook functies van afstanden, bijvoorbeeld reistijden (uitgedrukt in dagmarsen, fietsuren, of treinminuten) of reiskosten (treinkaartje, benzinekosten enzovoort). Al deze eenheden zijn bruikbaar, als we maar kunnen aannemen dat zij recht evenredig zijn met de werkelijke afstanden tussen de steden.

Nu kunnen we aan dit soort afstandsgegevens niet zonder meer zien hoe Nederland eruitziet, dat wil zeggen hoe de acht steden ten opzichte van elkaar gelokaliseerd zijn. De afstanden geven alleen maar aan hoever bijvoorbeeld Groningen en Eindhoven van Den Haag verwijderd zijn, maar geven ons geen aanwijzing of vanuit Den Haag gezien Groningen en Eindhoven in dezelfde richting liggen. Daarom zou het prettig zijn als we onze afstandsgegevens konden gebruiken om een landkaart te construeren; er is dus sprake van een cartografisch probleem.

Tabel 2.1 Afstanden in kilometers tussen acht Nederlandse steden

	Adm	Rdm	DnH	Utr	End	Arn	Zwo	Gro
Amsterdam	0							
Rotterdam	73	0						
Den Haag	57	21	0					
Utrecht	37	57	61	0				
Eindhoven	120	111	132	87	0			
Arnhem	98	117	116	62	81	0		
Zwolle	112	151	155	94	150	67	0	
Groningen	202	242	253	195	251	168	102	0

Als we inderdaad over gegevens beschikken die gelijk zijn aan of evenredig met de geografische afstanden tussen de steden (zoals in Tabel 2.1), dan kunnen we de kaart van Nederland construeren door twee van de acht steden op willekeurige plaatsen te tekenen. Hun verbindingslijn kunnen we dan gebruiken als basis van de zes driehoeken die gevormd worden door de gekozen twee en de zes overige steden. Deze driehoeken kunnen we met liniaal en passer construeren. Als we daarbij goed opletten welke steden aan de ene kant en welke steden aan de andere kant van de basislijn moeten komen, krijgen we een perfecte kaart. We moeten ons echter realiseren dat we er bij deze constructiemethode van uitgaan dat de afstanden tussen de steden langs een rechte lijn gemeten zijn. Bovendien gaan we er vanuit dat de kaart van Nederland plat is: anders gezegd, de ruimte waarin de acht steden afgebeeld moeten worden is tweedimensionaal. Zolang dat opgaat is de liniaal-en-passermethode geschikt. In drie dimensies wordt het al veel moeilijker om op deze manier een geschikte afbeelding te construeren. Maar in meer dan drie dimensies is deze methode echt onbruikbaar. Toch bestaan er wel verzamelingen objecten die in meer dan drie onafhankelijke aspecten of dimensies variëren. Daarom zullen we hieronder een algemene methode bespreken die geschikt is voor het analyseren van afstanden tussen punten die in één, twee of (veel) meer dimensies liggen.

2.2 DE YOUNG-HOUSEHOLDERMETHODE

Scalaire producten en coördinaten

Een algemene manier om het meerdimensionale cartografische probleem op te lossen is bedacht door Young en Householder in 1938. Zij maakten daarbij gebruik van de *cosinusregel* die zegt dat in een driehoek met hoekpunten A, B en C en zijden a , b en c geldt dat

$$a^2 = b^2 + c^2 - 2bc \{\cos(\varphi_{BAC})\} \quad [2.1]$$

en dus dat

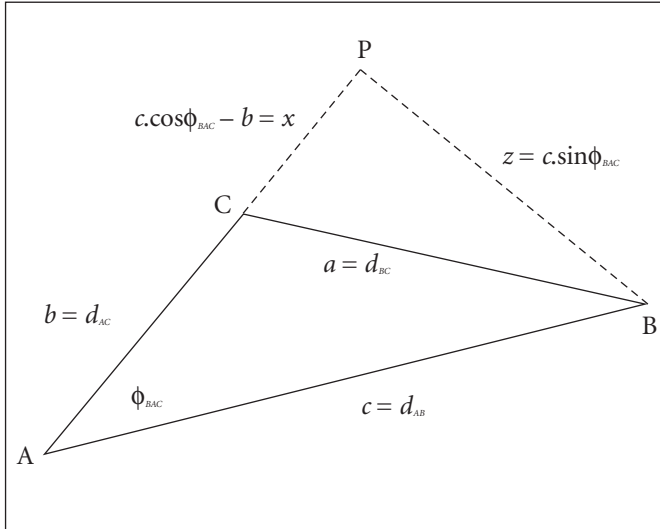
$$bc \{\cos(\varphi_{BAC})\} = (b^2 + c^2 - a^2)/2 \quad [2.2]$$

waarbij φ_{BAC} de hoek is die gevormd wordt door de zijden BA en CA. Een voorbeeld van zo'n driehoek is weergegeven in Figuur 2.1.

De cosinusregel kan gemakkelijk bewezen worden door de lijn AC te verlengen en vanuit punt B de loodlijn op AC neer te laten. Deze loodlijn snijdt AC in het punt P. Passen we in de aldus ontstane rechthoekige driehoek BCP de *Stelling van Pythagoras* toe, dan zien we dat

$$\begin{aligned} a^2 &= x^2 + z^2 \\ &= \{c \cdot \cos(\varphi_{BAC}) - b\}^2 + \{c \cdot \sin(\varphi_{BAC})\}^2 \\ &= c^2 \cdot \cos^2(\varphi_{BAC}) + b^2 - 2bc \cdot \cos(\varphi_{BAC}) + c^2 \cdot \sin^2(\varphi_{BAC}) \\ &= c^2 \{\cos^2(\varphi_{BAC}) + \sin^2(\varphi_{BAC})\} + b^2 - 2bc \cdot \cos(\varphi_{BAC}) \\ &= c^2 + b^2 - 2bc \cdot \cos(\varphi_{BAC}) \end{aligned} \quad [2.3]$$

waarmee de cosinusregel bewezen is (NB: in dit bewijs wordt gebruikgemaakt van het feit dat $\sin^2(\alpha) + \cos^2(\alpha) = 1$).



Figuur 2.1 Bewijs van de cosinusregel

Aangezien a , b en c niets anders zijn dan andere symbolen voor de afstanden d_{BC} , d_{AC} , en d_{AB} , kan men de cosinusregel ook schrijven als

$$d_{AC} d_{AB} \cos(\varphi_{BAC}) = (d_{AC}^2 + d_{AB}^2 - d_{BC}^2)/2. \quad [2.4]$$

Hier maken we op de volgende manier gebruik van.

Gegeven een verzameling van m objecten A, B, C,..., M (met onderlinge afstanden $d_{AB}, d_{AC}, \dots, d_{AM}, \dots, d_{LM}$) die in een meerdimensionale ruimte liggen, kunnen we één willekeurig object (stel A) uitkiezen als het gemeenschappelijke hoekpunt van alle driehoeken die gevormd worden door vanuit dat punt lijnen te trekken naar alle andere punten en deze punten ook onderling met elkaar te verbinden. Aangezien de lengten van alle zijden (de afstanden) bekend zijn, kunnen we in iedere driehoek met behulp van Formule [2.4] de *scalaire producten* $d_{AB}d_{AC}\cos(\varphi_{BAC}), d_{AB}d_{AD}\cos(\varphi_{BAD}), d_{AB}d_{AE}\cos(\varphi_{BAE}),$ enzovoort, uitrekenen. Deze producten, die we respectievelijk b_{BC}, b_{BD} en b_{BE} zullen noemen, verzamelen we in een symmetrische $(m-1) \times (m-1)$ matrix die met het symbool **B** aangeduid wordt. Deze bewerkingen hebben we gedaan met de afstanden uit Tabel 2.1, waarbij we Amsterdam als vast hoekpunt gekozen hebben. Het resultaat staat in Tabel 2.2. Ter verificatie:

$$\begin{aligned} b_{rdm,dnh} &= (d_{adm,rdm}^2 + d_{adm,dnh}^2 - d_{rdm,dnh}^2)/2 \\ &= (73^2 + 57^2 - 21^2)/2 = (5329 + 3249 - 441)/2 \\ &= 8137/2 = 4068.5 \end{aligned}$$

Tabel 2.2 *Scalaire-productenmatrix B voor zeven Nederlandse steden, met Amsterdam als oorsprong*

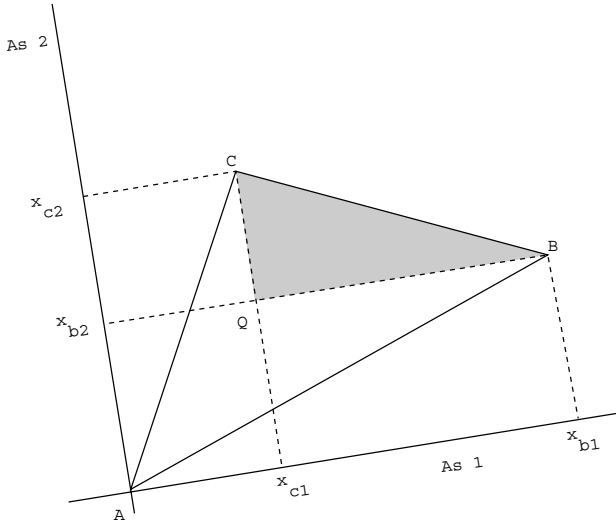
	Rdm	DnH	Utr	End	Arn	Zwo	Gro
Rotterdam	5329.0						
Den Haag	4068.5	3249.0					
Utrecht	1724.5	448.5	1369.0				
Eindhoven	3704.0	112.5	4100.0	14400.0			
Arnhem	622.0	-301.5	3564.5	8721.5	9604.0		
Zwolle	-2464.0	-4116.0	2538.5	2222.0	8829.5	12544.0	
Groningen	-6215.5	-9978.0	2074.0	-3898.5	11092.0	21472.0	40804.0

Uitgaande van de matrix **B** is het mogelijk om voor alle punten coördinaten te berekenen op de loodrecht op elkaar staande assen van de ruimte waarin die punten liggen. De oorsprong van dit assenstelsel wordt dan gevormd door het ene, eerder gekozen, vaste hoekpunt. Bijvoorbeeld: in het tweedimensionale geval dat in Figuur 2.2 is afgebeeld, valt hoekpunt A van driehoek ABC samen met de oorsprong van het assenstelsel. A, B en C hebben respectievelijk de coördinaten $(0, 0)$, (x_{B1}, x_{B2}) en (x_{C1}, x_{C2}) op de twee dimensies. Als we vanuit B en C loodlijnen neerlaten op As 1 en As 2, dan snijden twee van deze loodlijnen elkaar in punt Q. Voor de rechthoekige driehoek BCQ geldt nu dat

$$\begin{aligned}
d_{BC}^2 &= d_{BQ}^2 + d_{CQ}^2 \\
&= (x_{B1} - x_{C1})^2 + (x_{B2} - x_{C2})^2 \\
&= x_{B1}^2 + x_{B2}^2 + x_{C1}^2 + x_{C2}^2 - 2x_{B1}x_{C1} - 2x_{B2}x_{C2} \\
&= \sum_s x_{Bs}^2 + \sum_s x_{Cs}^2 - 2\sum_s x_{Bs}x_{Cs}.
\end{aligned} \tag{2.5}$$

Merk op dat $\sum_s x_{Bs}^2 = d_{AB}^2$ en dat $\sum_s x_{Cs}^2 = d_{AC}^2$. Combineren we dit resultaat met dat van Formule [2.4] dan zien we dat

$$d_{AC}d_{AB}\cos(\varphi_{BAC}) = b_{BC} = \sum_s x_{Bs}x_{Cs}. \tag{2.6}$$



Figuur 2.2 De relatie tussen de cosinusregel en de coördinaten van drie punten in twee dimensies

Het belang van deze formule is, dat er nu een relatie gelegd is tussen de scalaire producten, die we direct uit de bekende afstanden kunnen uitrekenen, en de nog onbekende coördinaten van de punten in een meerdimensionale ruimte. Het precieze aantal dimensies van deze ruimte is ook nog onbekend, we weten alleen dat het maximale aantal dimensies $m - 1$ is. Het MDS-probleem bestaat nu hieruit dat we, gegeven een tabel van scalaire producten, met behulp van Formule [2.6] de onbekende coördinaten moeten zien te vinden. Een oplossing voor dit probleem wordt gegeven door de *lineaire algebra*: de onbekende coördinaten kunnen gevonden worden door een zogenaamde *eigenwaarden-eigenvectorenanalyse* van de matrix van scalaire producten B .

Eigenwaarden-eigenvectorenanalyse van B

De relatie tussen scalaire producten en coördinaten van Formule [2.6] kan als volgt in matrixnotatie¹ worden weergegeven

$$B = XX' \quad [2.7]$$

waarbij X een $(m-1) \times r$ matrix is die de coördinaten van de $m-1$ punten op r assen bevat. Ons cartografische probleem is nu om gegeven de matrix B , die uit de afstanden berekend is, de bijbehorende matrix X te berekenen. Daartoe maken we gebruik van een bekende stelling uit de matrixalgebra die zegt dat iedere symmetrische matrix S van orde $k \times k$ (dat wil zeggen met k rijen en k kolommen) te ontbinden is in het product

$$S = P\Lambda P' \quad [2.8]$$

waarin P een $k \times k$ -matrix is die de *eigenvectoren* van S bevat; Λ is een diagonale $k \times k$ -matrix met de *eigenwaarden* van S (een diagonale matrix is een vierkante matrix met nullen in alle cellen behalve op de diagonaal). De ontbinding van een vierkante, symmetrische matrix in een product van eigenwaarden en eigenvectoren wordt *eigenwaarden-eigenvectorenanalyse* genoemd.

Passen we het resultaat van Formule [2.8] toe op de matrix B , dan kunnen we schrijven $B = Q\Lambda Q'$ en kunnen we X berekenen als $X = Q\Lambda^{1/2}$ ($\Lambda^{1/2}$ is een diagonale matrix met op de diagonaal de wortels uit de eigenwaarden van B).

Door de eigenwaarden-eigenvectorenoplossing beschikken we nu dus over $m-1$ eigenwaarden λ_s met $s = 1, \dots, (m-1)$ en over $m-1$ eigenvectoren die elk weer bestaan uit $m-1$ elementen q_{is} met $i = 1, \dots, (m-1)$. Vermenigvuldiging van de elementen van elke eigenvector met de wortel uit de bijbehorende eigenwaarde levert de getallen x_{is} op die we kunnen opvatten als de coördinaten van de objecten B, C, D, ..., enzovoort op de s -de dimensie. Maximaal zijn er in dit geval $m-1$ eigenwaarden die groter dan nul zijn, zodat we op deze manier de coördinaten berekenen van $m-1$ punten in een maximaal $(m-1)$ -dimensionale ruimte waarvan de oorsprong samenvalt met het m -e punt.

Ordenen we de eigenwaarden van groot naar klein, dan zullen vaak slechts p eigenwaarden groter dan nul zijn. In dat geval kunnen we de objecten perfect afbeelden in een p -dimensionale ruimte. Bij de eigenwaarden die groter dan nul zijn, zien we vaak een groepje van enkele eigenwaarden die relatief groot zijn, naast een aantal eigenwaarden die maar weinig van nul verschillen, of in ieder geval klein zijn vergeleken bij de grote eigenwaarden. In de praktijk kunnen we die kleinere dimensies vaak net zo goed laten vallen. We beelden de objecten dan af in r dimensies met $r < p < (m-1)$. Zo'n afbeelding is dan niet

¹ De lezer die nog niet met matrixalgebra bekend is, kan misschien toch wel de grote lijn van de tekst volgen door de diverse matrixbewerkingen als symbolische bewerkingen (vermenigvuldigen, delen, ontbinden in factoren) te interpreteren. In appendix A van dit boek is een elementaire inleiding in enkele basisbegrippen van de matrixalgebra opgenomen.

langer perfect, maar is meestal wel praktisch bruikbaar; vaak beter bruikbaar dan een perfecte oplossing in een (veel) groter aantal dimensies. Bijvoorbeeld: stel dat de afstanden tussen de steden gemeten waren tussen de hoogste punten van de hoogste gebouwen van elke stad. In die afstand speelt dan ook het hoogteverschil tussen de gebouwen mee. Een perfecte afbeelding zou dus drie dimensies vereisen. Echter, de verschillen in hoogte (enkele tientallen meters) vallen in 't niet bij de verschillen in lengte- en breedterichting (tientallen kilometers). De eigenwaarde die met de hoogtedimensie correspondeert, zal zeer klein zijn, in ieder geval veel kleiner dan de andere twee eigenwaarden, en mag daarom verwaarloosd worden.

De Young-Householderconstructie van de kaart van Nederland

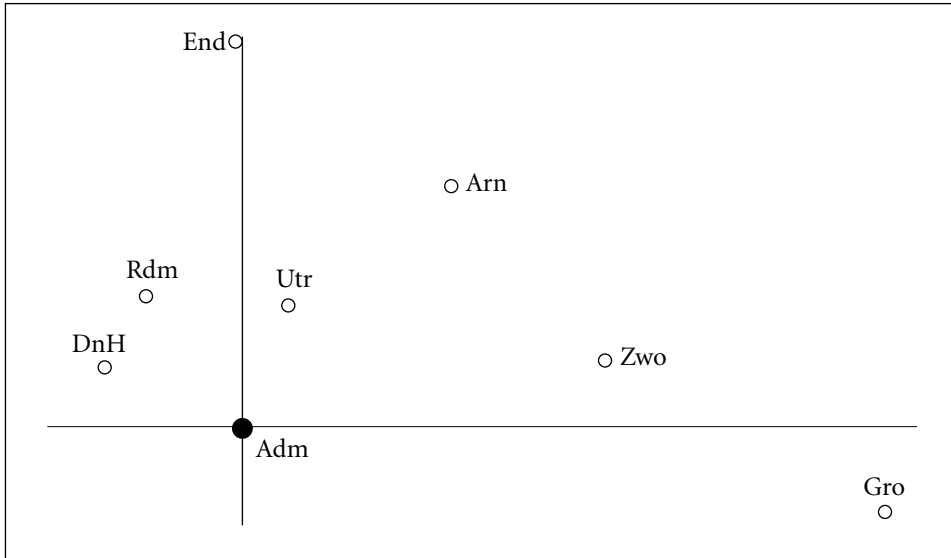
Eerder in dit hoofdstuk is in Tabel 2.2 de matrix **B** weergegeven. Deze bevat de scalaire producten die berekend zijn uit de afstanden tussen de acht steden in Tabel 2.1, waarbij Amsterdam als oorsprong gekozen is. In Tabel 2.3 zijn de eerste twee eigenwaarden² en eigenvectoren van **B** weergegeven, alsmede de coördinaten van de acht steden op twee dimensies. Zetten we deze coördinaten uit in een grafiek, dan levert dit een gedraaide en gespiegelde kaart van Nederland op (zie Figuur 2.3). De kaart van Nederland zoals wij die gewend zijn, krijgen we door de grafiek op een doorzichtig blaadje over te trekken. Draaien we nu dit blaadje tegen de richting van de klok in, zodanig dat Groningen helemaal boven en Eindhoven helemaal beneden ligt en keren we het blaadje om, dan zien we dat de kaart nagenoeg perfect is.

Tabel 2.3 Eigenvectoren en eigenwaarden van de matrix uit Tabel 2.2 en de bijbehorende coördinaten van zeven Nederlandse steden^a

	Eigenvectoren		Coördinaten	
Rotterdam	-.124	.259	-30.21	39.92
Den Haag	-.188	.114	-45.91	17.55
Utrecht	.060	.237	14.59	36.54
Eindhoven	-.003	.762	-.65	117.51
Arnhem	.265	.481	64.69	74.24
Zwolle	.450	.138	109.91	21.29
Groningen	.820	-.181	200.17	-27.91
Eigenwaarden	59568.5	23788.5		

^a Amsterdam heeft coördinaten (0, 0).

2 Voor de scalaire producten van Tabel 2.2 vinden we van groot naar klein geordend zeven eigenwaarden: 59568.5, 23788.5, 5245.0, 1534.4, -33.2, -417.2 en -2387.0. De eerste twee zijn aanmerkelijk groter dan de overige. De (kleine) negatieve eigenwaarden geven aan dat er inconsistenties in de data zitten waardoor zij eigenlijk niet in een Euclidische ruimte passen. Oorzaak is in dit geval dat de afstanden wegkilometers zijn en niet langs rechte lijnen gemeten zijn.



Figuur 2.3 Acht steden en hun positie op de kaart van Nederland

De fit van een schaaloplossing

Omdat wij al weten hoe de werkelijke kaart van Nederland eruitziet, kunnen we aan Figuur 2.3 meteen zien dat de gevonden afbeelding een behoorlijk grote overeenkomst met de werkelijkheid heeft. Zouden we – wat meestal het geval is – de echte configuratie van de punten niet al van tevoren kennen, dan kunnen we aan de uiteindelijke afbeelding niet zonder meer zien of de oplossing een goede overeenstemming (*fit*) met de oorspronkelijke, geobserveerde gegevens heeft. Om uit te maken in hoeverre oplossing en data overeenkomen, kunnen we één of meer van onderstaande methoden toepassen.

- 1 Stel dat de matrix X de coördinaten x_{is} bevat op de dimensies die als oplossing gekozen zijn. Uit deze coördinaten kunnen we nieuwe scalaire producten \hat{b}_{ij} berekenen volgens $\hat{b}_{ij} = \sum_s x_{is} x_{js}$ [Formule 2.6] of $\hat{B} = XX'$ [Formule 2.7]. Vervolgens maken we een grafiek waarin we de oorspronkelijke scalaire producten uitzetten tegen de scalaire producten die we uit de oplossing hebben terugberekend. In het ideale geval vallen de punten in die grafiek op een rechte lijn door het nulpunt. In Tabel 2.4 staan de scalaire producten die berekend zijn op basis van de coördinaten op de twee dimensies uit Tabel 2.3. In Figuur 2.4 zijn de geobserveerde scalaire producten uitgezet tegen de scalaire producten uit de oplossing. Het blijkt dat de afgebeelde punten nagenoeg een rechte lijn vormen.
- 2 In plaats van scalaire producten, kunnen we uit de coördinaten van de oplossing ook de bijbehorende afstanden \hat{d}_{ij} berekenen en deze vergelijken met de oorspronkelijk geobserveerde afstanden (op basis waarvan de scalaire producten voor de analyse werden berekend). Om de afstanden-uit-de-oplossing terug te rekenen gebruiken we de Euclidische afstandsfunctie (zie ook Hoofd-

stuk 4) aangezien ook de berekening van de scalaire producten van deze functie uitgaat. In twee dimensies is dus $\hat{d}_{ij} = \{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2\}^{1/2}$. Vervolgens maken we weer een grafiek waarin we de geobserveerde afstanden d_{ij} uitzetten tegen de afstanden \hat{d}_{ij} die zijn terugberekend. Idealiter liggen ook in deze grafiek de punten op een rechte lijn door de oorsprong. In Tabel 2.5 staan de afstanden tussen de steden die terugberekend zijn op basis van de coördinaten op twee dimensies; in Figuur 2.5 zijn deze afstanden uitgezet tegen de geobserveerde afstanden. Ook deze grafiek laat slechts kleine afwijkingen van een rechte lijn zien.

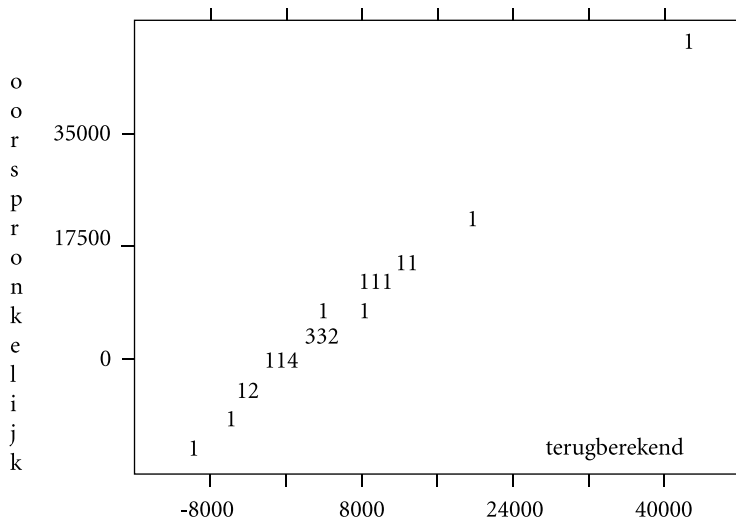
Tabel 2.4 *Scalaire producten terugberekend uit de coördinaten op twee dimensies*

	Rdm	DnH	Utr	End	Arn	Zwo	Gro
Rotterdam	2506*						
Den Haag	2088	2416					
Utrecht	1018	-29	1548				
Eindhoven	4710	2092	4285	13809			
Arnhem	1009	-1667	3657	8682	9696		
Zwolle	-2471	-4672	2382	2430	8690	12533	
Groningen	-7162	-9680	1901	-3409	10877	21407	
40849							

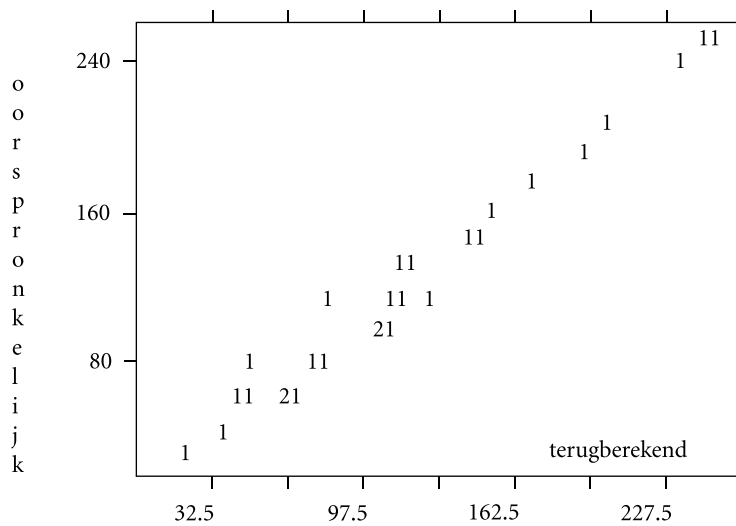
* De scalaire producten zijn afgerond op gehele getallen.

Tabel 2.5 *Afstanden terugberekend uit de coördinaten op twee dimensies*

	Adm	Rdm	DnH	Utr	End	Arn	Zwo
Rotterdam	50.1						
Den Haag	49.2	27.3					
Utrecht	39.3	44.9	63.4				
Eindhoven	117.5	83.0	109.7	82.4			
Arnhem	98.5	100.9	124.3	62.7	78.4		
Zwolle	112.0	141.4	155.9	96.5	146.6	69.6	
Groningen	202.1	240.2	250.3	196.5	248.0	169.7	102.8



Figuur 2.4 Oorspronkelijke scalaire producten (verticale as) uitgezet tegen scalaire producten uit de oplossing (horizontale as)



Figuur 2.5 Geobserveerde afstanden (verticale as) uitgezet tegen afstanden in de oplossing (horizontale as)

- 3 Een derde manier om de kwaliteit van een oplossing te bepalen, is door middel van coëfficiënten, zogenaamde *fit*-maten. Het ligt voor de hand om daarbij uit te gaan van de som van de gekwadraterde *residuen* $\sum_i \sum_j (b_{ij} - \hat{b}_{ij})^2$ en $\sum_i \sum_j (d_{ij} - \hat{d}_{ij})^2$. Delen we deze sommen respectievelijk door $\sum_i \sum_j b_{ij}^2$ en $\sum_i \sum_j d_{ij}^2$ en trekken we deze quotiënten van 1 af, dan krijgen we twee *goodness-of-fit*-maten die we hier BFIT en DFIT zullen noemen:

$$\text{BFIT} = 1 - \frac{\sum_i \sum_j (b_{ij} - \hat{b}_{ij})^2}{\sum_i \sum_j b_{ij}^2} \quad [2.9]$$

en

$$\text{DFIT} = 1 - \frac{\sum_i \sum_j (d_{ij} - \hat{d}_{ij})^2}{\sum_i \sum_j d_{ij}^2}. \quad [2.10]$$

Op deze coëfficiënten zullen we later in dit hoofdstuk nog terugkomen.

De oorsprong van de configuratie

In de hierboven gevolgde aanpak om scalaire producten te berekenen moet men een van de objecten laten samenvallen met de oorsprong van de afbeeldingsruimte. Hoewel het er in principe niet toe doet welk object men als oorsprong kiest, zou het in de praktijk het handigst zijn als men dát object zou kiezen dat in het centrum van de configuratie ligt. In ons voorbeeld hadden we in plaats van Amsterdam dus beter Utrecht als oorsprong kunnen kiezen. In sommige gevallen heeft men van tevoren geen enkel idee welke stimulus in het midden komt te liggen. Gelukkig is het niet nodig om de oorsprong van de ruimte altijd met een van de objecten te laten samenvallen. De relaties tussen afstanden, scalaire producten en coördinaten gelden ook wanneer de oorsprong in een willekeurig punt van de ruimte geplaatst wordt. Een handige keuze is om de oorsprong te laten samenvallen met het *zwaartepunt*, de *centroïde*, van de configuratie van objecten. De centroïde van een verzameling punten is dat punt waarvoor geldt dat de som van de gekwadraterde afstanden van dat punt tot alle andere punten zo klein mogelijk is. Daardoor ligt de centroïde op elke dimensie in het gemiddelde van de coördinaten van de overige punten.

Om ook dan de Young-Householdermethode toe te kunnen passen, hebben we scalaire producten nodig van de punten in een ruimte waarvan de oorsprong in de centroïde van die punten ligt. Deze scalaire producten, die we weer b_{ij} zullen noemen, kunnen we berekenen uit

$$b_{ij} = -(1/2)(d_{ij}^2 - \sum_j d_{ij}^2/m - \sum_i d_{ij}^2/m + \sum_i \sum_j d_{ij}^2/m^2). \quad [2.11]$$

Hieronder zullen we zien dat we de scalaire producten van alle punten tegelijkertijd kunnen berekenen door de tabel met gekwadrateerde afstanden tussen de punten *dubbel te centreren*. Deze door Torgerson (1958) bedachte methode wordt gedetailleerd beschreven in Blok 2.1. Zijn de nieuwe scalaire producten weer in een matrix B verzameld, dan kunnen via eigenwaarden-eigenvectoren-analyse van die matrix de coördinaten x_{is} gevonden worden.

BLOK 2.1 SCALAIRE PRODUCTEN MET DE CENTROÏDE ALS OORSPRONG

Bij het berekenen van scalaire producten hoeven we niet *per se* een van de stimuluspunten als oorsprong van de ruimte kiezen. In plaats daarvan kunnen we de oorsprong in de centroïde van de stimulusconfiguratie plaatsen. Als we dat doen, dan geldt dat de som van de coördinaten van de punten op elke dimensie gelijk aan nul is. Dus, als x_{is} de coördinaat is van punt i op dimensie s , dan is $\sum_i x_{is} = 0$ voor alle dimensies s , waarbij $s = 1, \dots, r$.

Laten we de centroïde aanduiden met de letter Z. Ieder tweetal punten vormt met Z een driehoek. Bijvoorbeeld: A en B vormen met Z de driehoek ABZ. Voor deze driehoek ABZ geldt natuurlijk weer de cosinusregel:

$$d_{AB}^2 = d_{AZ}^2 + d_{BZ}^2 - 2d_{AZ}d_{BZ}\cos(\phi_{AZB}). \quad [2.12]$$

Is er nog een derde punt C, dan geldt eveneens dat

$$d_{AC}^2 = d_{AZ}^2 + d_{CZ}^2 - 2d_{AZ}d_{CZ}\cos(\phi_{AZC}) \quad [2.13]$$

en

$$d_{BC}^2 = d_{BZ}^2 + d_{CZ}^2 - 2d_{BZ}d_{CZ}\cos(\phi_{BZC}). \quad [2.14]$$

Merk op dat we de gekwadrateerde afstanden van de punten A, B en C tot de oorsprong Z kunnen schrijven in termen van de coördinaten van A, B en C. Zo is

$$d_{AZ}^2 = \sum_s x_{as}^2 \quad [2.15]$$

waarbij x_{as} de coördinaat is van stimulus A op dimensie s van de in totaal r -dimensionale ruimte. Analooch zijn $d_{BZ}^2 = \sum_s x_{bs}^2$ en $d_{CZ}^2 = \sum_s x_{cs}^2$. Gebruik makend van Formule [2.6] weten we dat

$$d_{AZ}^2 d_{BZ}^2 \cos(\phi_{AZB}) = \sum_s^r x_{as}^2 x_{bs}^2 \quad [2.16]$$

zodat Formule [2.12] herschreven kan worden als

$$d_{AB}^2 = \sum_s^r x_{as}^2 + \sum_s^r x_{bs}^2 - 2 \sum_s^r x_{as} x_{bs}. \quad [2.17]$$

Analoog zijn $d_{AC}^2 = \sum_s^r x_{as}^2 + \sum_s^r x_{cs}^2 - 2 \sum_s^r x_{as} x_{cs}$ en $d_{BC}^2 = \sum_s^r x_{bs}^2 + \sum_s^r x_{cs}^2 - 2 \sum_s^r x_{bs} x_{cs}$. Is er nog een vierde punt, D, dan krijgen we $d_{AD}^2 = \sum_s^r x_{as}^2 + \sum_s^r x_{ds}^2 - 2 \sum_s^r x_{as} x_{ds}$, $d_{BD}^2 = \sum_s^r x_{bs}^2 + \sum_s^r x_{ds}^2 - 2 \sum_s^r x_{bs} x_{ds}$ en $d_{CD}^2 = \sum_s^r x_{cs}^2 + \sum_s^r x_{ds}^2 - 2 \sum_s^r x_{cs} x_{ds}$.

Stel dat we vier punten A, B, C en D hebben waarvan de afstanden d_{AB} , d_{AC} , d_{AD} , d_{BC} , d_{BD} en d_{CD} in de afstandsmatrix D staan. De kwadraten van de getallen in D zijn stuk voor stuk opgebouwd uit de hierboven genoemde componenten (zie Formule [2.17] en verder). Deze componenten zijn uitgeschreven in Tabel 2.6. Merk op dat ook de cellen op de diagonaal (die de waarde nul hebben) in componenten zijn opgesplitst: $d_{AA}^2 = d_{BB}^2 = d_{CC}^2 = d_{DD}^2 = 0$ zijn vervangen door respectievelijk $(2 \sum_s^r x_{as}^2 - 2 \sum_s^r x_{as} x_{as})$, $(2 \sum_s^r x_{bs}^2 - 2 \sum_s^r x_{bs} x_{bs})$, $(2 \sum_s^r x_{cs}^2 - 2 \sum_s^r x_{cs} x_{cs})$ en $(2 \sum_s^r x_{ds}^2 - 2 \sum_s^r x_{ds} x_{ds})$.

Tabel 2.6 Componenten van de gekwadrateerde afstanden uit D

Stimulus					
Stimulus	A	B	C	D	$(1/4) \sum_j^m$
A	$\sum_s^r x_{as}^2 + \sum_s^r x_{as}^2$ $- 2 \sum_s^r x_{as} x_{as}$	$\sum_s^r x_{as}^2 + \sum_s^r x_{bs}^2$ $- 2 \sum_s^r x_{as} x_{bs}$	$\sum_s^r x_{as}^2 + \sum_s^r x_{cs}^2$ $- 2 \sum_s^r x_{as} x_{cs}$	$\sum_s^r x_{as}^2 + \sum_s^r x_{ds}^2$ $- 2 \sum_s^r x_{as} x_{ds}$	$\sum_s^r x_{as}^2 +$ $(1/4) \sum_j^m \sum_s^r x_{js}^2$
B	$\sum_s^r x_{as}^2 + \sum_s^r x_{bs}^2$ $- 2 \sum_s^r x_{as} x_{bs}$	$\sum_s^r x_{bs}^2 + \sum_s^r x_{bs}^2$ $- 2 \sum_s^r x_{bs} x_{bs}$	$\sum_s^r x_{bs}^2 + \sum_s^r x_{cs}^2$ $- 2 \sum_s^r x_{bs} x_{cs}$	$\sum_s^r x_{bs}^2 + \sum_s^r x_{ds}^2$ $- 2 \sum_s^r x_{bs} x_{ds}$	$\sum_s^r x_{bs}^2 +$ $(1/4) \sum_j^m \sum_s^r x_{js}^2$
C	$\sum_s^r x_{as}^2 + \sum_s^r x_{cs}^2$ $- 2 \sum_s^r x_{as} x_{cs}$	$\sum_s^r x_{bs}^2 + \sum_s^r x_{cs}^2$ $- 2 \sum_s^r x_{bs} x_{cs}$	$\sum_s^r x_{cs}^2 + \sum_s^r x_{cs}^2$ $- 2 \sum_s^r x_{cs} x_{cs}$	$\sum_s^r x_{cs}^2 + \sum_s^r x_{ds}^2$ $- 2 \sum_s^r x_{cs} x_{ds}$	$\sum_s^r x_{cs}^2 +$ $(1/4) \sum_j^m \sum_s^r x_{js}^2$
D	$\sum_s^r x_{as}^2 + \sum_s^r x_{ds}^2$ $- 2 \sum_s^r x_{as} x_{ds}$	$\sum_s^r x_{bs}^2 + \sum_s^r x_{ds}^2$ $- 2 \sum_s^r x_{bs} x_{ds}$	$\sum_s^r x_{cs}^2 + \sum_s^r x_{ds}^2$ $- 2 \sum_s^r x_{cs} x_{ds}$	$\sum_s^r x_{ds}^2 + \sum_s^r x_{ds}^2$ $- 2 \sum_s^r x_{ds} x_{ds}$	$\sum_s^r x_{ds}^2 +$ $(1/4) \sum_j^m \sum_s^r x_{js}^2$
$(1/4) \sum_j^m$:	$\sum_s^r x_{as}^2 +$ $(1/4) \sum_i^m \sum_s^r x_{is}^2$	$\sum_s^r x_{bs}^2 +$ $(1/4) \sum_i^m \sum_s^r x_{is}^2$	$\sum_s^r x_{cs}^2 +$ $(1/4) \sum_i^m \sum_s^r x_{is}^2$	$\sum_s^r x_{ds}^2 +$ $(1/4) \sum_i^m \sum_s^r x_{is}^2$	

Tellen we nu voor elke rij van Tabel 2.6 de getallen over de kolommen bij elkaar op en delen we deze som vervolgens door m (het aantal stimuli; hier vier!) dan bestaan die rijgemiddelden uit de componenten die in de meest rechtse kolom van de tabel zijn weergegeven. Bijvoorbeeld:

$$\begin{aligned} (1/4)\sum_j^m d_{ij}^2 &= (4/4)\sum_j^r x_{is}^2 + (1/4)\sum_j^r x_{js}^2 - (2/4)\sum_j^r x_{is}x_{js} \\ &= \sum_s^r x_{is}^2 + (1/4)\sum_j^r x_{js}^2 \end{aligned} \quad [2.18]$$

omdat $\sum_i \sum_s^r x_{is}x_{js} = \sum_s^r x_{is}x_{js} = \sum_s^r x_{is} \sum_j x_{js} = 0$ aangezien $\sum_j x_{js} = 0$. De gemiddelden van alle kolommen, berekend over de rijen, staan in de onderste rij van Tabel 2.6. Door de rij- of kolomgemiddelden op hun beurt zelf weer te middelen, berekenen we in feite de gemiddelde gekwadrateerde afstand van alle punten. Deze waarde is opgebouwd uit de volgende elementen:

$$\begin{aligned} (1/4)\sum_i (\sum_j d_{ij}^2 / 4) &= (1/4)\sum_i \sum_s^r x_{is}^2 - (1/4)(4)(1/4)\sum_j^r x_{js}^2 \\ &= (1/4)\sum_i \sum_s^r x_{is}^2 + (1/4)\sum_j^r x_{js}^2 \end{aligned} \quad [2.19]$$

Gebruikmakend van de Formules [2.17], [2.18] en [2.19] kunnen we zien wat er gebeurt als we de matrix met de gekwadrateerde waarden van **D** dubbel centreren. Dubbel centreren wil zeggen dat we van iedere cel het rijgemiddelde en het kolomgemiddelde aftrekken en het totaalgemiddelde erbij optellen. Dus: d_{ij}^2 wordt na dubbel centreren gelijk aan onderstaande optelling:

$(d_{ij}^2):$	$\sum_s^r x_{is}^2 + \sum_s^r x_{js}^2 - 2\sum_s^r x_{is}x_{js}$	
– (gemiddelde rij i):	$-\sum_s^r x_{is}^2 +$	$-(1/4)\sum_j^r x_{js}^2$
– (gemiddelde kolom j):	$-\sum_s^r x_{js}^2 +$	$-(1/4)\sum_i \sum_s^r x_{is}^2$
+ (totaalgemiddelde):		$+(1/4)\sum_i \sum_s^r x_{is}^2 + (1/4)\sum_j^r x_{js}^2$
(na optellen):		
	$0 + 0 - 2\sum_s^r x_{is}x_{js} + 0$	$+ 0$

wat dus -2 keer het scalaire product is van de coördinaten van punt i en punt j in een ruimte met als oorsprong Z . Noemen we deze scalaire producten weer b_{ij} , dan kunnen we b_{ij} berekenen uit

$$b_{ij} = - (1/2) (d_{ij}^2 - \sum_j d_{ij}^2 / m - \sum_i d_{ij}^2 / m + \sum_i \sum_j d_{ij}^2 / m) \quad [2.20]$$

Verzamelen we de scalaire producten weer in een matrix **B**, dan kunnen via eigenwaarden-eigenvectorenanalyse van **B** de coördinaten x_{is} gevonden worden in een ruimte waarvan de oorsprong in de centroide van de punten ligt.

Passen we Formule [2.11] of [2.20] toe op de gekwadrateerde afstanden tussen de acht Nederlandse steden uit ons voorbeeld, dan vinden we de scalaire producten die in Tabel 2.7 zijn weergegeven. Analyse hiervan levert de eigenvectoren, eigenwaarden en coördinaten op die in Tabel 2.8 zijn weergegeven. De bijbehorende landkaart is afgebeeld in Figuur 2.6. Uitgaande van de gevonden coördinaten zijn weer de bijbehorende scalaire producten en afstanden-in-de-oplossing terugberekend. Deze staan respectievelijk in Tabel 2.9 en Tabel 2.10. In de Figuren 2.7 en 2.8 zijn de oorspronkelijke scalaire producten en afstanden uitgezet tegen de scalaire producten en afstanden uit de oplossing.

Tabel 2.7 Scalaire producten voor acht Nederlandse steden met hun centroide als oorsprong

	A d m	R d m	D n H	U t r	E n d	A r n	Z w o	G r o
Amsterdam	2874.1							
Rotterdam	2028.0	6510.9						
Den Haag	3688.7	6911.1	7752.3					
Utrecht	896.7	1775.1	2159.8	288.3				
Eindhoven	-796.1	2061.8	131.0	1326.5	9933.7			
Arnhem	-2392.5	-2616.5	-1879.3	-805.3	2658.9	1945.1		
Zwolle	-2254.2	-5564.3	-5555.6	-1693.1	-3702.4	1308.8	5161.6	
Groningen	-4044.7	-11106.3	-13208.1	-3948.1	-11613.4	1780.8	12299.1	29840.6

Tabel 2.8 Eerste twee eigenvectoren, eigenwaarden en coördinaten van de dubbelgecentreerde gekwadrateerde afstanden

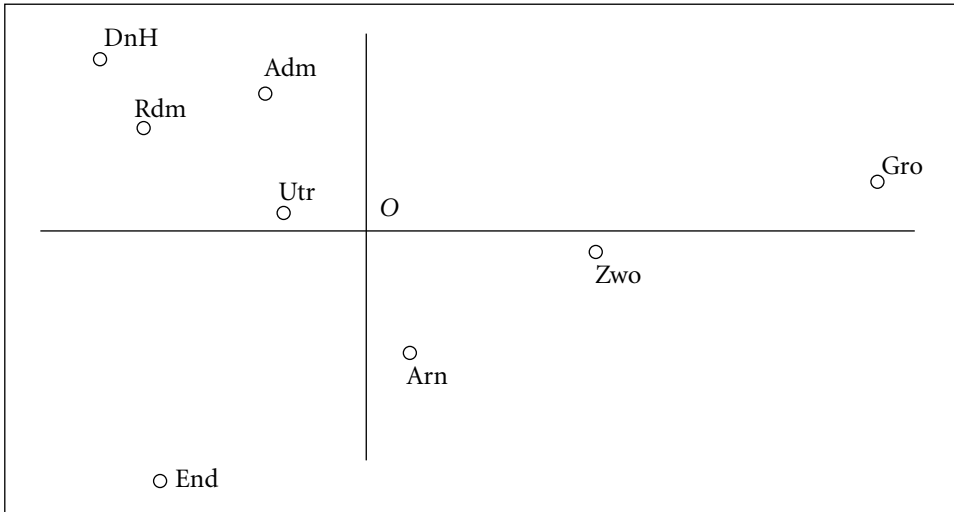
	Eigenvectoren		Coördinaten	
Amsterdam	-.120	.306	-27.0	34.6
Rotterdam	-.312	.259	-70.3	29.2
Den Haag	-.345	.399	-77.7	45.1
Utrecht	-.106	.037	-23.9	4.1
Eindhoven	-.261	-.723	-58.8	-81.7
Arnhem	.060	-.363	13.5	-41.0
Zwolle	.319	-.057	71.9	-6.4
Groningen	.765	.144	172.3	16.3
Eigenwaarden	50745.3	12752.4		

Tabel 2.9 Scalaire producten (afgerond) terugberekend uit de coördinaten van de oplossing

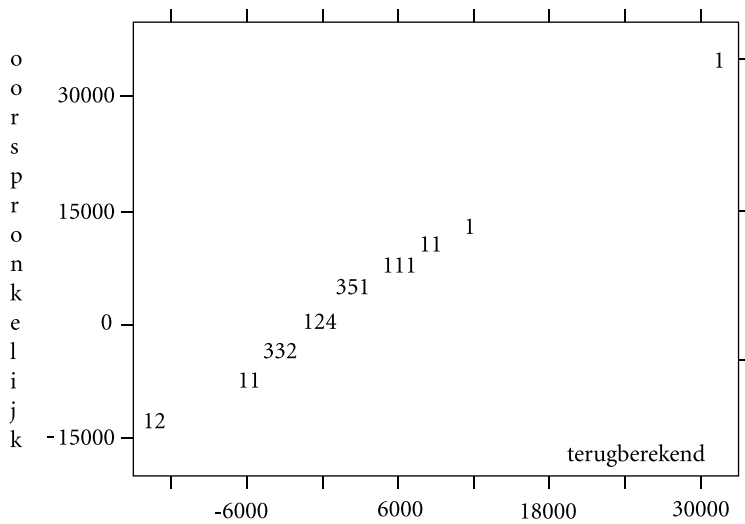
	Adm	Rdm	DnH	Utr	End	Arn	Zwo	Gro
Amsterdam	1922							
Rotterdam	2906	5778						
Den Haag	3658	6774	8077					
Utrecht	789	1800	2047	589				
Eindhoven	-1231	1733	887	1068	10120			
Arnhem	-1781	-2146	-2898	-492	2561	1866		
Zwolle	-2166	-5233	-5881	-1747	-3698	1233	5211	
Groningen	-4098	-11613	-12665	-4054	-11440	1657	12281	29933

Tabel 2.10 Afstanden terugberekend uit de coördinaten van de oplossing

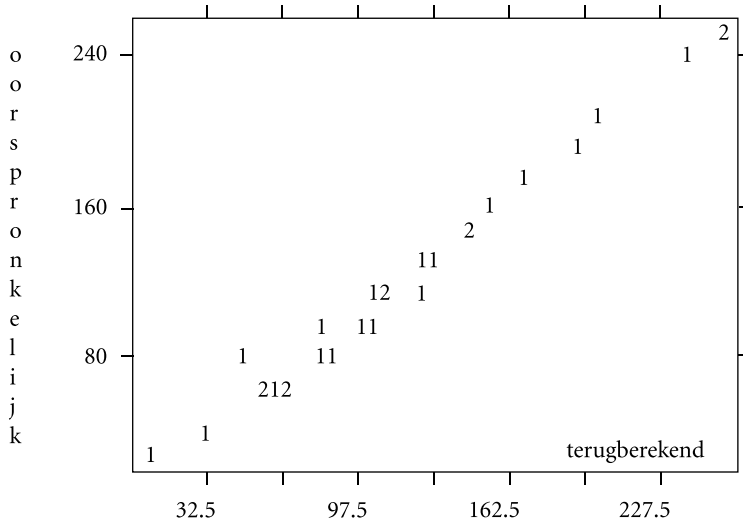
	Adm	Rdm	DnH	Utr	End	Arn	Zwo
Rotterdam	43.5						
Den Haag	51.8	17.5					
Utrecht	30.5	52.6	67.6				
Eindhoven	120.4	111.5	128.1	92.6			
Arnhem	85.7	109.3	125.5	58.6	82.9		
Zwolle	107.1	146.5	158.3	96.4	150.8	67.9	
Groningen	200.1	242.8	250.9	196.5	250.9	168.8	102.9



Figuur 2.6 Afbeelding van acht Nederlandse steden met hun centroïde in de oorsprong



Figuur 2.7 Oorspronkelijke scalaire producten uitgezet tegen terugberekende scalaire producten uit de oplossing met de centroïde als oorsprong



Figuur 2.8 Oorspronkelijke afstanden uitgezet tegen de afstanden in de oplossing

Nogmaals: de FIT van de oplossing

Eerder in dit hoofdstuk hebben we gezien dat de kwaliteit van een oplossing niet alleen met behulp van grafieken maar ook door middel van coëfficiënten vastgesteld kan worden. In Formule [2.9] en [2.10] definieerden we BFIT en DFIT die gebaseerd zijn op de residuen $(b_{ij} - \hat{b}_{ij})$ en $(d_{ij} - \hat{d}_{ij})$. Van BFIT kan bewezen worden dat deze coëfficiënt gelijk is aan

$$\text{BFIT} = \frac{\sum_i \sum_j \hat{b}_{ij}^2}{\sum_i \sum_j b_{ij}^2} = \frac{\sum_{s=1}^r \lambda_s^2}{\sum_{s=1}^m \lambda_s^2} \quad [2.21]$$

waarbij m zoals steeds het aantal objecten is en r het aantal dimensies is dat men voor de oplossing gekozen heeft. BFIT geeft aan hoe goed de scalaire producten van de m punten, die in een ruimte met maximaal $m - 1$ dimensies liggen, gereproduceerd kunnen worden door ze af te beelden (te projecteren) in een deelruimte met minder dimensies. Door juist die r dimensies te kiezen die de grootste eigenwaarden hebben, zorgen we ervoor dat de afbeelding *optimaal* is. Dat wil zeggen: iedere andere r -dimensionale afbeelding is minder goed dan de gekozen oplossing of is hooguit even goed (wat alleen kan als sommige eigenwaarden aan elkaar gelijk zijn).

Opgemerkt moet worden dat de precieze waarde van BFIT afhangt van het punt dat men als oorsprong van de ruimte kiest. Naarmate de oorsprong verder van de centroide afligt zal BFIT over het algemeen groter zijn dan wanneer de oorsprong in het midden van de puntenconfiguratie ligt. Alleen bij een perfecte r -dimensionale oplossing maakt het niet uit waar de oorsprong ligt; BFIT heeft dan altijd de waarde 1.0. In Tabel 2.11 zien we dat de tweedimensionale afbeel-

ding met Amsterdam als oorsprong een B_{FIT} heeft die maar weinig van 1.00 afwijkt en nagenoeg gelijk is aan de B_{FIT} van de dubbelgecentreerde oplossing. Naast B_{FIT} en D_{FIT} is er nog een derde, veelgebruikte *goodness-of-fit-maat* die we hier Λ_{FIT} zullen noemen:

$$\Lambda_{FIT} = \frac{\sum_{s=1}^r \lambda_s}{\sum_{s=1}^m \lambda_s}. \quad [2.22]$$

Uitsluitend en alleen als het hierbij gaat om de eigenwaarden van een dubbelgecentreerde B -matrix, dus alleen wanneer de oorsprong van de ruimte in de centroïde van de punten ligt, geldt er dat

$$\Lambda_{FIT} = \frac{\sum_i \sum_j \hat{d}_{ij}^2}{\sum_i \sum_j d_{ij}^2}. \quad [2.23]$$

Hieruit blijkt dat de eigenwaarden van een dubbelgecentreerde B -matrix een interessante interpretatie hebben, namelijk als de proportie van de gekwadrateerde afstanden die door een bepaalde dimensie gereproduceerd of ‘verklaard’ worden.

Merk op dat Λ_{FIT} grote waarden aan kan nemen als er een of meer niet te verwaarlozen, negatieve eigenwaarden zijn. Het voorkomen van relatief veel of relatief belangrijke negatieve eigenwaarden duidt erop dat de geobserveerde afstanden niet goed in een Euclidische ruimte kunnen worden weergegeven, terwijl de hoge Λ_{FIT} juist suggereert dat we wel een goede oplossing hebben. Daarom worden er in de literatuur (Cox & Cox, 1994) twee modificaties van Λ_{FIT} voorgesteld:

$$\Lambda_{FIT}^* = \frac{\sum_{s=1}^r \lambda_s}{\sum_{s=1}^m |\lambda_s|}. \quad [2.24]$$

en

$$\Lambda_{FIT}^{**} = \frac{\sum_{s=1}^r \lambda_s}{\sum_{s=1}^p (\lambda_s | \lambda_s > 0)} = \frac{\sum_{s=1}^r \lambda_s}{\sum (\text{alle positieve eigenwaarden})} \quad [2.25]$$

Tabel 2.11 *Fit-maten voor de twee oplossingen van de kaart van Nederland*

Coëfficiënt	Oorsprong in Amsterdam	Oorsprong in centroïde
λ_1	59568.5	50745.0
λ_2	23788.5	12752.0
λ_3	5245.0	1831.0
λ_4	1534.4	1529.0
λ_5	- 33.2	0.0
λ_6	- 417.2	- 31.0
λ_7	- 2387.0	- 395.0
λ_8	—	- 2124.0
$\sum_s^2 \lambda_s$ (a)	83357.0	63497.0
$\sum_s^m \lambda_s$ (b)	87299.0	64307.0
Λ_{FIT} (a/b)	.955	.987
$\sum_s^m \lambda_s $ (c)	92974.0	69410.0
Λ_{FIT}^* (a/c)	.892	.915
$\sum_s^2 \lambda_s^2$ (d)	4.1143×10^9	2.7377×10^9
$\sum_s^m \lambda_s^2$ (e)	4.1500×10^9	2.7481×10^9
B_{FIT} (d/e)	.996	.991
$\sum_i \sum_j d_{ij}^2 / (2m)$	1028904.0/16 = 64306.5	64306.5
$\sum_i \sum_j \hat{d}_{ij}^2 / (2m)$	61403.0	63498.0

In Tabel 2.11 worden voor de twee oplossingen van het kaart-van-Nederland-probleem verschillende *fit*-coëfficiënten weergegeven, samen met de waarden van een aantal grootheden waaruit zij berekend zijn. De conclusie is dat de kwaliteit van beide oplossingen goed of zelfs bijna perfect te noemen is, en dat de *fit*-maten elkaar dus niet of nauwelijks ontlopen.

Beperkingen van de Young-Householdermethode

Hierboven hebben we gezien dat de Young-Householdermethode een geschikte methode is om m punten op grond van hun onderlinge afstanden perfect of bijna perfect in een meerdimensionale ruimte af te beelden. In het stedenvoorbeeld gingen we ervan uit dat de gegevens die geanalyseerd moesten worden *perfect* overeenstemmen, dat wil zeggen, recht evenredig zijn met ‘echte’, fysieke, geografische, langs rechte lijnen gemeten afstanden tussen de objecten. In

de meeste praktijkgevallen beschikken we niet over zulke perfecte afstandsgegevens. De 'afstanden' die op de een of andere manier 'gemeten' zijn, zullen vaak niet zonder meer evenredig zijn met de 'echte' afstanden. Dit kan gebeuren (a) doordat er toevallige fouten gemaakt zijn bij het vaststellen van de afstanden (meetfouten; onbetrouwbaarheid), (b) doordat de 'gemeten' afstanden een of andere transformatie zijn van de 'echte' afstanden of (c) doordat beide oorzaken van toepassing zijn.

In het eerste geval, 'echte' afstanden met meetfouten, kunnen we nog steeds de Young-Householdermethode gebruiken. Het enige dat er gebeurt is dat er meer eigenwaarden ongelijk nul gevonden worden dan voor een goede afbeelding van de objecten nodig is. Die extra eigenwaarden kunnen positief of negatief zijn. Als de meetfouten niet al te groot zijn, dan zijn de kleinste eigenwaarden qua absolute waarde (zeer) klein vergeleken bij de grootste eigenwaarden. De coördinaten die berekend worden uit de eigenvectoren met de grootste eigenwaarden zijn dan waarschijnlijk voldoende om de objecten adequaat weer te geven. De eigenvectoren met kleinere eigenwaarden nemen in dat geval de meetfouten voor hun rekening. In het voorbeeld van de acht steden ontstaan zulke meetfouten doordat er een verschil bestaat tussen wegkilometers en afstanden in vogelvlucht. Deze meetfouten worden genegeerd doordat men de derde en latere dimensies laat vallen. In de gecentreerde afbeelding van de acht steden namen de eigenwaarden van de eerste twee dimensies samen ruim 91 procent van de som van de absolute eigenwaarden voor hun rekening ($\Delta_{FIT}^* = .915$; de derde tot en met zevende dimensie 'verklaarden' gezamenlijk dus nog geen tien procent van de scalaire producten!). Pas als de meetfouten groot worden vergeleken met de 'echte' afstanden ontstaat er een probleem: de positieve eigenwaarden zijn dan klein en verschillen niet duidelijk genoeg meer van elkaar om met zekerheid te kunnen zeggen welke dimensies belangrijke informatie geven en welke niet. Bovendien kunnen grote meetfouten meer en belangrijkere negatieve eigenwaarden veroorzaken.

Naast meetfouten in de afstanden kan het ook voorkomen dat de geobserveerde afstandsgegevens niet zonder meer evenredig zijn met 'echte' afstanden. Een voorbeeld is de prijs (in strippen) van een bus- of tramrit. In principe geldt dat hoe verder men reist, hoe meer strippen men moet laten afstempelen; het aantal strippen is dus een maat voor de gereisde afstand. Echter, onafhankelijk van de afstand die men gaat reizen, moet men één extra strip afstempelen. Het aantal strippen s is dus een lineaire functie van de afstand d (namelijk: $s = a \cdot d + 1$) en dus is ook afstand een lineaire functie van het aantal strippen ($d = s/a - 1/a = h \cdot s + c$). Voordat we op geobserveerde afstandsgegevens *à la* s de Young-Householdermethode toe mogen passen, moeten we eigenlijk eerst de 'echte' afstanden d berekenen door de waarde van c , de zogenaamde *additieve constante*, te schatten. Daarvoor zijn verschillende methoden ontwikkeld (zie Blok 2.2).

Wanneer de afstanden niet een simpele lineaire functie van de afstandsgegevens zijn, maar een of andere meer gecompliceerde, onbekende functie daar-

van, is de Young-Householdermethode strikt genomen niet meer toepasbaar. In dat geval moet men zogenaamde niet-metrische MDS-technieken gaan gebruiken. Die technieken komen in de rest van dit boek uitgebreid aan de orde.

BLOK 2.2 HET PROBLEEM VAN DE ADDITIEVE CONSTATE

In de metrische MDS van Young-Householder wordt in principe aangenomen dat de geobserveerde afstandsgegevens, die we voortaan als $\{o_{ij}\}$ zullen aanduiden, recht evenredig zijn met de 'echte' afstanden $\{d_{ij}\}$ tussen de objecten. Er geldt dus dat $o_{ij} = a \cdot d_{ij}$ of, analoog, dat $d_{ij} = b \cdot o_{ij}$. Omdat de factor b alleen maar invloed heeft op de grootte van de afbeelding en niet op de vorm ervan (b is de *schaal*, de afbeeldingsmaatstaf van de oplossing), kunnen we deze factor net zo goed weglaten, zodat we kunnen schrijven $d_{ij} = o_{ij}$. Vaak is het realistischer om aan te nemen dat

$$d_{ij} = o_{ij} + c \quad [2.26]$$

wat impliceert dat $o_{ij} = d_{ij} - c$. Met andere woorden, de geobserveerde nabijheden zijn afstanden-op-een-intervalschaal. De constante c uit bovenstaande formules wordt de *additieve constante* genoemd; door c bij de geobserveerde waarden op te tellen worden van de nabijheidsgegevens echte afstanden gemaakt.

Wanneer we in de praktijk een verzameling geobserveerde nabijheden analyseren alsof het echte afstanden waren, dus zonder rekening te houden met de waarde van c , kan dat vervelende gevolgen hebben. Bijvoorbeeld: stel dat alle geobserveerde afstanden tussen de Nederlandse steden twintig kilometer kleiner waren dan de afstanden uit Tabel 2.1. De onderlinge nabijheden van Arnhem, Zwolle en Groningen zouden dan zijn 47, 82 en 148 (respectievelijk voor Arnhem-Zwolle, Zwolle-Groningen en Arnhem-Groningen). Zouden we nu met behulp van de liniaal-en-passermethode de driehoek Arnhem-Zwolle-Groningen willen construeren, dan blijkt dat dat niet gaat. Als we een lijn van 148 millimeter met Arnhem en Groningen als uiteinden nemen, dan kunnen we geen punt voor Zwolle vinden, omdat de twee cirkels vanuit de uiteinden met stralen 47 en 82 millimeter elkaar niet snijden. De geobserveerde nabijheden 47, 82 en 148 blijken niet te voldoen aan de zogenaamde *driehoeksongelijkheid*: de afstand tussen twee punten moet altijd kleiner zijn dan of gelijk zijn aan de som van de afstanden van die twee punten tot een derde punt. Als dat niet zo is, kunnen we geen afbeelding van de objecten vinden!

Laten we vervolgens eens veronderstellen dat de geobserveerde afstanden tussen de Nederlandse steden allemaal 20 kilometer groter waren dan de

waarden in Tabel 2.1. De afstanden Arnhem-Zwolle, Zwolle-Groningen en Arnhem-Groningen zouden dan 87, 122 en 188 kilometer bedragen. Nu kunnen we met de passer-methode wel een afbeelding van deze drie steden maken: we zien echter dat Arnhem, Zwolle en Groningen een uitgesproken driehoek vormen, terwijl ze in werkelijkheid (zie Figuur 2.3) nagenoeg op een rechte lijn liggen (immers: $d_{arn,gro} = 168$, $d_{arn,zwo} = 67$, $d_{zwo,gro} = 102$ en $168 \approx 67 + 102$).

Deze voorbeelden laten zien dat het consequenties voor de oplossing kan hebben als de waarde van de additieve constante onbekend is. Hieronder worden drie methoden besproken waarmee men dit probleem kan aanpakken. Twee van deze methoden zijn procedures om de additieve constante afzonderlijk te schatten (zodat men de nabijheidsgegevens kan aanpassen om vervolgens de Young-Householdermethode te gebruiken). In de derde methode wordt de additieve constante tegelijkertijd met de uiteindelijke MDS-afbeelding van de objecten verkregen.

Het schatten van c met behulp van de driehoeksongelijkheid

Voor ieder drietal $\{i, j, k\}$ geldt de driehoeksongelijkheid, dat wil zeggen dat

$$d_{ij} + d_{jk} \geq d_{ik}. \quad [2.27]$$

Substitutie van Formule [2.26] in [2.27] geeft

$$o_{ij} + c + o_{jk} + c \geq o_{ik} + c \quad [2.28]$$

zodat

$$c \geq o_{ik} - o_{ij} - o_{jk}. \quad [2.29]$$

Dus: voor een willekeurig drietal $\{i, j, k\}$ is de beste schatting voor c gelijk aan $\hat{c}_{\max(ijk)}$, dat wil zeggen, aan de maximumwaarde van

$$\begin{aligned} \hat{c}_{ikj} &= o_{ik} - o_{ij} - o_{jk} \\ \hat{c}_{ijk} &= o_{ij} - o_{ik} - o_{jk} \\ \hat{c}_{jki} &= o_{jk} - o_{ij} - o_{ik}. \end{aligned}$$

Voor alle drietallen gezamenlijk is de beste schatting van c daarom gelijk aan het maximum van de maximale schattingen van c in de afzonderlijke driehoeken. In formulevorm:

$$\hat{c} = \max_{(i,j,k)} (o_{jk} - o_{ij} - o_{ik}) \quad [2.30]$$

waarbij i, j en k elk de waarden 1 tot en met m kunnen aannemen.

Aan deze schattingsmethode kleven twee problemen. Ten eerste: voor de aldus gevonden waarde \hat{c} geldt dat hij kleiner dan of gelijk is aan de werkelijke waarde van c . Hoewel we hopen dat de 'ware' c in de buurt ligt van \hat{c} , zou c vele malen groter kunnen zijn. Weliswaar zorgt deze schattingsmethode ervoor dat alle afstanden na correctie voldoen aan de driehoeksongelijkheid, maar voor het overige is er nog veel onzekerheid over de precieze waarde van c . Een tweede probleem doet zich voor als er (veel) *error* in de geobserveerde nabijheden zit! Eén waarde van o_{ij} met een grote (al dan niet toevallige) meetfout kan de waarde van \hat{c} ernstig vertekenen.

De Messick-en-Abelsonmethode

Messick en Abelson beschreven in 1956 onderstaande methode om de additieve constante te schatten. Deze methode gaat uit van Formule [2.11], die identiek is aan Formule [2.20] en het verband aangeeft tussen de 'ware' scalaire producten en de 'echte', dubbelgecentreerde afstanden. Deze formule geven we hier opnieuw weer als Formule [2.31]:

$$b_{ij} = -\frac{1}{2} \left(d_{ij}^2 - \frac{\sum_j d_{ij}^2}{m} - \frac{\sum_i d_{ij}^2}{m} + \frac{\sum_i \sum_j d_{ij}^2}{m^2} \right). \quad [2.31]$$

Door substitutie van Formule [2.26] in [2.31] ontstaat [2.32]; daarin worden de 'ware' scalaire producten uitgedrukt als functie van de geobserveerde nabijheden en de onbekende additieve constante. Voor de *buitendiagonale elementen* van **B** (dus voor alle b_{ij} met $i \neq j$) geldt

$$b_{ij} = -\frac{1}{2} \left((o_{ij} + c)^2 - \frac{\sum_j (o_{ij} + c)^2}{m} - \frac{\sum_i (o_{ij} + c)^2}{m} + \frac{\sum_i \sum_j (o_{ij} + c)^2}{m^2} \right). \quad [2.32]$$

Aangezien de diagonale elementen van **O** uit nullen bestaan ($o_{11} = o_{22} = \dots = o_{mm} = 0$) zijn deze elementen exact gelijk aan de afstanden $d_{11} = d_{22} = \dots = d_{mm} = 0$. Hieruit leiden we af dat

$$\begin{aligned} b_{ij} = & -\frac{1}{2} \left(o_{ij}^2 - \frac{\sum_j o_{ij}^2}{m} - \frac{\sum_i o_{ij}^2}{m} + \frac{\sum_i \sum_j o_{ij}^2}{m^2} \right) \\ & - \frac{1}{2} (2c) \left(o_{ij} - \frac{\sum_j o_{ij}}{m} - \frac{\sum_i o_{ij}}{m} + \frac{\sum_i \sum_j o_{ij}}{m^2} \right) \\ & - \frac{1}{2} (c^2) \left(1 - \frac{m-1}{m} - \frac{m-1}{m} + \frac{m(m-1)}{m^2} \right). \end{aligned} \quad [2.33]$$

$$\begin{aligned}
&= -\frac{1}{2} \left(o_{ij}^2 - \frac{\sum_j o_{ij}^2}{m} - \frac{\sum_i o_{ij}^2}{m} + \frac{\sum_i \sum_j o_{ij}^2}{m^2} \right) \\
&- c \left(o_{ij} - \frac{\sum_j o_{ij}}{m} - \frac{\sum_i o_{ij}}{m} + \frac{\sum_i \sum_j o_{ij}}{m^2} \right) - \frac{1}{2} \left(\frac{c^2}{m} \right) \\
&= -\frac{1}{2} \left([o_{ij}^2 + 2co_{ij}] - \frac{\sum_j o_{ij}^2}{m} - \frac{\sum_i o_{ij}^2}{m} + \frac{\sum_i \sum_j o_{ij}^2}{m^2} \right) \\
&- c \left(-\frac{\sum_j o_{ij}}{m} - \frac{\sum_i o_{ij}}{m} + \frac{\sum_i \sum_j o_{ij}}{m^2} \right) - \frac{1}{2} \left(\frac{c^2}{m} \right)
\end{aligned}$$

Voor de diagonaalelementen van \mathbf{B} , de scalaire producten b_{ii} , geldt dat in Formule [2.33] de term tussen vierkante haken in de laatste regel wegvalt en dat er éénmaal de term c^2 te veel in voorkomt. Dat dit zo is komt doordat de elementen op de diagonaal van \mathbf{O} gelijk aan nul zijn. Daardoor en omdat de matrix \mathbf{O} symmetrisch is, wordt

$$\begin{aligned}
b_{ii} &= -\frac{1}{2} \left(-\frac{\sum_j o_{ij}^2}{m} - \frac{\sum_i o_{ij}^2}{m} + \frac{\sum_i \sum_j o_{ij}^2}{m^2} \right) \quad [2.34] \\
&- c \left(-\frac{\sum_j o_{ij}}{m} - \frac{\sum_i o_{ij}}{m} + \frac{\sum_i \sum_j o_{ij}}{m^2} \right) - \frac{c^2}{2} \left(\frac{1-m}{m} \right)
\end{aligned}$$

Tellen we de diagonaalelementen van \mathbf{B} bij elkaar op, dan krijgen we

$$\begin{aligned}
\sum_i b_{ii} &= -\frac{1}{2} \left(-2 \frac{\sum_i \sum_j o_{ij}^2}{m} + m \frac{\sum_i \sum_j o_{ij}^2}{m^2} \right) \quad [2.35] \\
&+ (2c) \frac{\sum_i \sum_j o_{ij}}{m} - mc \frac{\sum_i \sum_j o_{ij}}{m^2} - \frac{c^2}{2} (m-1)
\end{aligned}$$

wat gelijk is aan

$$\sum_i b_{ii} = \frac{1}{2} \left(\frac{\sum_i \sum_j o_{ij}^2}{m^2} \right) + c \left(\frac{\sum_i \sum_j o_{ij}}{m^2} \right) + \frac{1}{2} c^2 (m-1). \quad [2.36]$$

Formule [2.36] is een vierkantsvergelijking van c , waaruit we c zouden kunnen oplossen als we de waarde van $\sum_i b_{ii}$ zouden kennen. Messick en Abelson nu, stelden een procedure voor om deze waarde van $\sum_i b_{ii}$ te schatten. Deze gaat als volgt:

- Stap 1: kies een willekeurige, maar liever niet te kleine waarde voor c ;
- Stap 2: tel deze 'schatting' van c op bij de geobserveerde afstanden o_{ij} ;
- Stap 3: bereken uit deze waarden ($o_{ij} + c$) scalaire producten via Formule [2.19];
- Stap 4: bereken de eigenwaarden van de matrix met deze scalaire producten;
- Stap 5: kies die eigenwaarden die duidelijk groter zijn dan de andere. De som van deze eigenwaarden is een goede schatting van de onbekende waarde $\sum_i b_{ii}$;
- Stap 6: vul in Formule [2.36] deze schatting van $\sum_i b_{ii}$ in en los c op. Neem deze waarde van c als nieuwe schatting en ga naar Stap 2 (NB: Stap 6 is bij Messick en Abelson in feite iets ingewikkelder dan hier beschreven is).

Herhaal deze procedure vervolgens net zolang tot de schatting van c *geconvergeerd* is. Dat wil zeggen tot de nieuwste schatting van c (nagenoeg) gelijk is aan de waarde van de voorafgaande schatting. Opgemerkt kan worden dat deze procedure tamelijk ingewikkeld is, omdat er steeds opnieuw scalaire producten en eigenwaarden berekend moeten worden. Zeker in 1956, toen de computer nog maar net bestond, was dit een hele opgave. Volgens Messick en Abelson echter blijkt dit algoritme doorgaans snel te convergeren.

Zoals Saito (1978) laat zien, komt deze procedure neer op het schatten van c zodanig dat $\Lambda_{FIT} = \sum_s^r \lambda_s / \sum_s^m \lambda_s$ gemaximaliseerd wordt met zo min mogelijk dimensies. Saito zelf stelt een andere procedure voor waarin $BFIT = \sum_s^r \lambda_s^2 / \sum_s^m \lambda_s^2$ gemaximaliseerd wordt.

De directe methode

In deze derde methode wordt de additieve constante tegelijkertijd met de uiteindelijke MDS-afbeelding van de objecten verkregen. Net als bij Messick en Abelson gaat dit via een *iteratieve* procedure. In zo'n procedure worden bij herhaling dezelfde stappen doorlopen, waardoor de oorspronkelijke schatting van een of andere grootte succesievelijk verbeterd

(*geoptimaliseerd*) wordt. In dit geval begint men echter niet met een schatting voor c , maar met een 'schatting' van de MDS-configuratie; men noemt dit de *initiële configuratie*. In deze configuratie berekent men de afstanden tussen de objecten en gaat men na of er een (nagenoeg) perfect lineair verband tussen deze afstanden en de geobserveerde nabijheden bestaat. Als dat zo is, dan heeft men meteen al een goede MDS-oplossing, dat wil zeggen, een oplossing die bij de data past. In het algemeen zullen de afstanden in de initiële configuratie geen perfect lineair verband met de observaties hebben. Daarom wordt vervolgens de configuratie veranderd (aangepast, verbeterd) waarna er opnieuw afstanden-in-de-configuratie berekend worden. Daarna wordt opnieuw nagegaan of deze afstanden een (nagenoeg) perfect lineair verband met de geobserveerde nabijheden hebben. Is dat niet zo, dan wordt de configuratie weer aangepast, enzovoort. Deze procedure herhaalt men net zo lang tot men tevreden is over de lineaire *fit* tussen de data en de afstanden. Dit is de aanpak die in alle moderne meerdimensionale schaaltechnieken gevolgd wordt. Deze procedure wordt uitgebreid behandeld in Hoofdstuk 6.

