HOOFDSTUK 4

Afstandsfuncties en afstandsmodellen

4.1 Afstandsfuncties

In de vorige hoofdstukken zijn we het begrip afstand op twee manieren tegengekomen. In de eerste plaats hebben we gezien dat geobserveerde nabijheidsgegevens uit echte, fysieke afstanden kunnen bestaan. In de tweede plaats hebben we te maken gekregen met afstanden in een MDS-oplossing, dat wil zeggen, met afstanden in een configuratie van punten die door middel van coördinaten op dimensies zijn afgebeeld. Tot nu toe zijn we er steeds stilzwijgend van uitgegaan dat de afstanden in de configuratie zogenaamde *Euclidische* afstanden zijn, afstanden die gemeten worden langs de kortste, rechtlijnige verbinding tussen twee punten. In dit hoofdstuk wordt een aantal andere manieren behandeld waarop afstanden tussen punten berekend kunnen worden.
Om uit de coördinaten van een verzameling punten de afstanden tussen die punten te bepalen, kan men gebruikmaken van verschillende *afstandsfuncties*. Afstandsfuncties zijn wiskundige functies die getallen – afstanden – toekennen aan alle tweetallen uit een verzameling van punten in een ruimte. Een afstand is dan de waarde die wordt aangenomen door een wiskundige functie die de coördinaten van de punten als argument heeft. Voor stimulus i en j in een ruimte met r dimensies en coördinaten $\{x_{is}, x_{js}; s = 1,...,r\}$ kunnen we dus schrijven: $d_{ij} = f(x_{i1},...,x_{ir}; x_{j1},...,x_{jr})$. In matrixnotatie: $\mathbf{D} = f(\mathbf{X})$ waarbij $\mathbf{X}$ de matrix met coördinaten $\{x_{is}; i = 1,...,m; s = 1,...,r\}$ voorstelt.
Er bestaan vele verschillende afstandsfuncties. Echter, voor iedere afstandsfunctie, hoe die er verder ook uit mag zien, moeten de volgende vier eigenschappen gelden.

1 *Niet-negativiteit*: voor de afstand d_{ij} tussen de punten *i* en *j* moet gelden dat $d_{ij} \geq 0$.
2 *Niet-gedegenereerdheid*: enerzijds moet altijd gelden dat $d_{ii} = 0$, anderzijds dat $d_{ij} = 0$ als en slechts dan als $i = j$, dus uitsluitend als *i* en *j* in één punt in de ruimte samenvallen.
3 *Symmetrie*: $d_{ij} = d_{ji}$, dat wil zeggen: de afstand van punt *i* tot punt *j* is gelijk aan de afstand van punt *j* tot punt *i*.
4 *Driehoeksongelijkheid*: $d_{ij} \leq d_{ik} + d_{jk}$, dat wil zeggen: de afstand tussen twee punten, zeg *i* en *j*, is altijd kleiner dan of gelijk aan de som van de afstand tussen *i* en een derde punt, *k*, en de afstand tussen *j* en *k*.

Functies die getallen toekennen aan puntenparen, zijn alleen dan afstandsfuncties als zij bovengenoemde vier eigenschappen hebben; zij worden dan *metrieken* genoemd. Functies die deze eigenschappen niet hebben zijn dus geen afstandsfuncties. Aan de andere kant: iedere functie $f(\mathbf{X})$ die deze eigenschappen wel bezit, is 'automatisch' een afstandsfunctie, een metriek.

Minkowski-afstanden

Beperken we ons tot de ruimte van de reële getallen, dan kunnen we een aantal afstandsfuncties onderscheiden die behoren tot de zogenaamde familie van Minkowski-afstandsfuncties. De algemene formule van een Minkowski-afstand is

$$d_{ij} = [\Sigma_s |x_{is} - x_{js}|^p]^{1/p} \qquad [4.1]$$

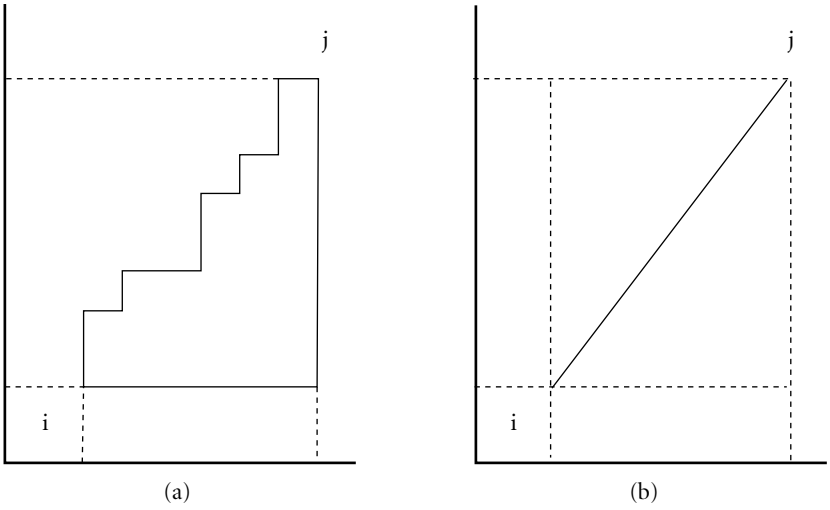
waarin x_{is} de coördinaat is van punt *i* op dimensie *s*. De exponent *p* is de zogenaamde Minkowski-parameter, die alle waarden van één tot en met oneindig aan kan nemen ($1 \leq p \leq \infty$). De afstand tussen twee punten wordt volgens deze formule dus berekend door op elke dimensie het absolute verschil tussen de coördinaten van de punten te nemen, deze verschillen tot de macht *p* te verheffen en bij elkaar op te tellen, en daarna uit dit totaal de wortel met exponent *p* te trekken. Van de familie van Minkowski-afstandsfuncties zijn er enkele van bijzonder belang, namelijk die met $p = 1$, $p = 2$, en $p = \infty$. Deze zullen we hieronder afzonderlijk behandelen.

'City block'-afstanden. Dit zijn afstanden die men krijgt als men voor *p* de waarde 1 substitueert, zodat Formule [4.1] versimpelt tot

$$d_{ij} = \Sigma_s |x_{is} - x_{js}|. \qquad [4.2]$$

Dit betekent dat de afstand tussen punt *i* en punt *j* gelijk is aan het absolute verschil van de coördinaten van *i* en *j* op de eerste dimensie plus het absolute verschil van hun coördinaten op de tweede dimensie, plus de absolute verschillen tussen hun coördinaten op de eventuele derde en volgende dimensies. Van punt *i* naar punt *j* gaande, mag je alleen maar bewegen in richtingen die

evenwijdig zijn aan de assen van de ruimte, zodat je alleen maar loodrechte hoeken maakt (zie Figuur 4.1a). Dit soort afstanden gelden als je in een stad met loodrecht op elkaar staande straten (bijvoorbeeld Manhattan) van de ene plaats naar de andere gaat: vandaar de naam 'City block'-afstanden.



Figuur 4.1 City block-afstanden (a) en Euclidische afstanden (b) in twee dimensies

Euclidische afstanden. In dit geval is $p = 2$. Afstanden tussen punten in de ruimte kunnen nu berekend worden via de vertrouwde *Stelling van Pythagoras* (zie Figuur 4.1b). Omdat alle verschillen tussen de coördinaten gekwadrateerd worden, kunnen de absoluutstrepen uit Formule [4.1] worden weggelaten, zodat

$$d_{ij} = [\Sigma_s (x_{is} - x_{js})^2]^{1/2}. \qquad [4.3]$$

Dit is de Euclidische afstandsmetriek zoals we die in ons dagelijks leven in feite toepassen.
De Euclidische afstandsfunctie kan ook gemakkelijk in vectorvorm weergegeven worden. Stel dat X een matrix is met *m* rijen en *r* kolommen waarin de coördinaten van *m* punten op *r* dimensies zijn opgeslagen. De coördinaten van punt *i* staan dan in de *rijvector* $\mathbf{x'}_i$, die van punt *j* in de rijvector $\mathbf{x'}_j$[1].
Trekken we $\mathbf{x'}_j$ van $\mathbf{x'}_i$ af, dan ontstaat er een nieuwe rijvector met elementen $\{(x_{i1} - x_{j1}), (x_{i2} - x_{j2}),...,...,(x_{ir} - x_{jr})\}$ die we kunnen schrijven als $(\mathbf{x'}_i - \mathbf{x'}_j)$.
Deze vector bevat dus de verschillen tussen de coördinaten van *i* en *j* op de

1 Het accent in $\mathbf{x'}_i$ geeft aan dat het hier de *i*-de *rij* van matrix **X** betreft; $\mathbf{x}_i$ zonder accent duidt de *i*-de *kolom* van **X** aan.

opeenvolgende dimensies. Gebruikmakend van deze notatie ziet de Euclidische afstandsfunctie er als volgt uit:

.

$$d^2_{ij} = (\mathbf{x}_i' - \mathbf{x}'_j)(\mathbf{x}'_i - \mathbf{x}'_j)' . \qquad [4.4]$$

De gekwadrateerde Euclidische afstand tussen *i* en *j* is dus het product van de rijvector $(\mathbf{x}'_i - \mathbf{x}'_j)$ navermenigvuldigd met de getransponeerde van zichzelf.

Dominantie-afstanden (ook wel *Chebychev*-afstanden genoemd). In dit geval is $p = \infty$, waarvoor geldt dat

$$d_{ij} = [\Sigma_s |x_{is} - x_{js}|^{\infty}]^{1/\infty} = \text{MAXIMUM}_s |x_{is} - x_{js}| . \qquad [4.5]$$

In woorden: d_{ij} is gelijk aan de waarde van het absolute verschil tussen de coördinaten van *i* en *j* op die dimensie waarop dat verschil het grootst is. De afstand op de dimensie waarop het verschil het grootst is, *domineert* de afstanden op alle andere dimensies. Dit kunnen we als volgt bewijzen. Als we Formule [4.1] uitschrijven voor het driedimensionale geval, dan krijgen we

$$d_{ij}^p = [|x_{il} - x_{jl}|^p + |x_{i2} - x_{j2}|^p + | x_{i3} - x_{j3}|^p] \qquad [4.6]$$

waaruit volgt dat

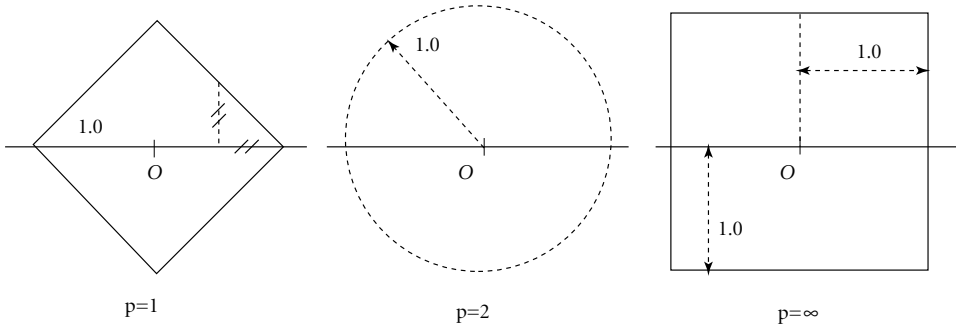
$$d_{ij} = \left(\frac{|x_{i1} - x_{j1}|^{p-1}}{d_{ij}^{p-1}}\right) |x_{i1} - x_{j1}| + \left(\frac{|x_{i2} - x_{j2}|^{p-1}}{d_{ij}^{p-1}}\right) |x_{i2} - x_{j2}| + \left(\frac{|x_{i3} - x_{j3}|^{p-1}}{d_{ij}^{p-1}}\right) |x_{i3} - x_{j3}| \qquad [4.7]$$

De drie termen (voor $s = 1,...,3$) van het rechterlid $|x_{is} - x_{js}|$ worden elk vermenigvuldigd met een factor $(|x_{is} - x_{js}|^{p-1})/d_{ij}^{p-1}$. Wanneer er, zoals hierboven, drie dimensies zijn, dan weten we dat de absolute verschillen tussen *i* en *j* op de twee dimensies waar die verschillen het kleinst zijn in ieder geval kleiner zijn dan d_{ij}. Immers, uit de algemene vorm van de Minkowski-metriek volgt dat d_{ij} groter dan of gelijk moet zijn aan het verschil tussen *i* en *j* op die dimensie (laten we die *g* noemen) waarop dat verschil het grootst is. Als *p* nu naar oneindig gaat, dan naderen de factoren $(|x_{is} - x_{js}|^{p-1})/d_{ij}^{p-1}$ waarmee de verschillen op de twee dimensies met de kleinste verschillen vermenigvuldigd worden, al gauw tot nul. Daaruit volgt dat $d_{ij} = (|x_{ig} - x_{jg}|^p)/d_{ij}^{p-1} + 0$, zodat $d_{ij}^p = |x_{ig} - x_{jg}|^p$ en dus dat $d_{ij} = |x_{ig} - x_{jg}| = \text{MAXIMUM}_s |x_{is} - x_{js}|$.
Stel dat het verschil tussen *i* en *j* het grootst is op dimensie drie, dan wordt $d_{ij} = 0 \times |x_{i1} - x_{j1}| + 0 \times |x_{i2} - x_{j2}| + 1 \times |x_{i3} - x_{j3}| = |x_{i3} - x_{j3}|$.
Bovenstaande redenering geldt uiteraard ook als er slechts twee of juist meer dan drie dimensies zijn.

Overigens laat Formule [4.7] zien dat er een interessante interpretatie van de Minkowski-parameter p te geven is (zie Cross, 1965). De absolute verschillen tussen i en j op alle dimensies zijn de componenten waaruit de afstand tussen i en j is opgebouwd. Elk van deze componenten wordt *gewogen* met een factor $(|x_{i1} - x_{j1}|^{p-1})/d_{ij}^{p-1}$. Als $p = 1$, dan worden alle componenten even zwaar gewogen, maar naarmate p groter wordt, worden de grotere componenten steeds zwaarder gewogen, totdat bij $p = \infty$ alleen de grootste component nog meetelt.[2]

Eigenschappen van Minkowski-metrieken

Om enig inzicht in de eigenschappen van de genoemde afstandsfuncties te geven, zijn in Figuur 4.2 drie zogenaamde 'eenheidscirkels' getekend. Dit zijn de meetkundige plaatsen van alle punten in twee dimensies die allemaal op afstand 1 van de oorsprong O verwijderd liggen. In de Euclidische ruimte is dit een echte cirkel; in de 'city block'- en in de dominantiemetriek hebben deze meetkundige plaatsen de vorm van vierkanten. In de dominantiemetriek geldt voor ieder punt op de verticale zijden dat het verschil op de horizontale dimensie (dat gelijk aan 1 is) het verschil op de verticale dimensie domineert. Voor alle punten op de horizontale zijden geldt het omgekeerde.



Figuur 4.2 Eenheidscirkels in verschillende Minkowski-metrieken

Beals, Krantz en Tversky (1968) hebben laten zien dat Minkowski-afstanden, naast de eigenschappen niet-negativiteit, niet-gedegenereerdheid, symmetrie en driehoeksongelijkheid, nog enkele extra eigenschappen hebben: *intradimensionale subtractiviteit*, *interdimensionale additiviteit* en *segmentsgewijze*

2 Aan de mogelijke waarden van p zou men een inhoudelijke psychologische interpretatie kunnen geven. Stel dat de dimensies eigenschappen zijn waarop stimuli kunnen variëren en dat d_{ij} de mate is waarin proefpersonen tussen deze stimuli kunnen differentiëren. De waarde $p = 1$ impliceert dan dat alle verschillen tussen de stimuli bij elkaar worden opgeteld en in gelijke mate meedoen bij de discriminatie van de stimuli. De waarde $p = \infty$ impliceert dat alleen het grootste verschil een rol speelt (zie Coombs, Dawes & Tversky, 1970).

additiviteit. Intradimensionale subtractiviteit slaat op de eigenschap dat de afstand tussen twee punten een functie is van hun (absolute) verschillen op de afzonderlijke dimensies. Bijvoorbeeld: in de Euclidische afstandsfunctie is afstand gelijk aan de wortel uit de som van de gekwadrateerde absolute *verschillen*. Interdimensionale additiviteit betreft de eigenschap dat de afstand tussen twee punten een *optelling* is over de dimensies van bepaalde waarden per dimensie van de desbetreffende punten. Bijvoorbeeld: in de 'city block'-metriek worden de absolute verschillen tussen de punten over de dimensies opgeteld. Wanneer een metriek aan beide eigenschappen voldoet, dan impliceert dit dat de dimensies waarop punten *niet* verschillen *niet* bijdragen aan de afstand tussen die punten. Ook impliceert dit dat grotere verschillen op de ene dimensie in zekere mate gecompenseerd worden door kleinere verschillen op een of meer andere dimensies. Zoals aan Formule [4.1] is te zien, zijn Minkowski-metrieken in wezen additieve-verschilmetrieken, en dus hebben zij de eigenschappen van intradimensionale subtractiviteit en interdimensionale additiviteit. De derde eigenschap, segmentsgewijze additiviteit, komt erop neer dat voor elk punt k dat op de kortste verbindingslijn tussen twee punten i en j ligt, geldt dat $d_{ik} + d_{kj} = d_{ij}$.
Een uitgebreidere behandeling van de drie extra eigenschappen van Minkowski-metrieken en hun implicaties is te vinden in Beals, Krantz en Tversky (1968) en in Bezembinder (1970). Beals, Krantz en Tversky hebben op basis van deze eigenschappen procedures voorgesteld om te bepalen of een verzameling empirische afstandsdata in een Minkowski-metriek weergegeven kan worden. De praktische bruikbaarheid van deze procedures is echter gering.[3]

Varianten van Minkowski-afstanden

Naast de Minkowski-afstandsfuncties, die de drie hierboven genoemde extra eigenschappen hebben, bestaan er vele andere afstandsfuncties die deze eigenschappen niet of niet alledrie hebben. Carroll en Wish (1974) bespreken twee varianten van de Minkowski-metriek. De eerste daarvan is een familie van afstandsfuncties die zij de *uitgebreide Minkowski-metriek* noemen. Deze heeft de algemene vorm

$$d_{ij} = [\Sigma_s |x_{is} - x_{js}|^p] \qquad [4.8]$$

waarin de exponent $1/p$ ontbreekt en p waarden tussen 0 en 1 aanneemt. De afstandsfuncties die tot deze familie behoren, zijn wel echte metrieken, in de zin dat ze de vier essentiële eigenschappen (niet-negatief, niet-gedegenereerd,

3 De door Beals, Krantz en Tversky (1968) voorgestelde procedures zijn alleen uitvoerbaar als men wil aannemen dat de data vrij zijn van toevalsfouten en via het discrete proces gemeten zijn. Voor sommige procedures is het nodig dat de dimensies waarop de objecten van elkaar verschillen en de bijbehorende (volgorde van de) coördinaten van tevoren bekend zijn.

symmetrie, driehoeksongelijkheid) en twee van de drie extra eigenschappen bezitten. Wat ontbreekt, is de segmentsgewijze additiviteit.
Een tweede variant van de Minkowski-metriek ontstaat als men toestaat dat p in de originele Minkowski-afstandsfunctie van Formule [4.1] ook negatieve waarden aan mag nemen. De afstand tussen twee objecten is in deze variant een *inverse* functie van hun absolute verschillen op de diverse dimensies. Met andere woorden, hoe groter hun absolute verschillen op de dimensies, hoe kleiner hun afstand. Carroll en Wish (1974) noemen deze variant een *semi-metriek* omdat de afstanden die hierbij horen niet meer aan de driehoeksongelijkheid voldoen. Een speciaal geval van deze semi-metriek is de functie die men krijgt door p gelijk aan $-\infty$ te maken. Hierdoor ontstaat de *'minimum-afstandsmetriek'*, de tegenhanger van de dominantie-afstand. Volgens de minimum-afstandsmetriek is d_{ij} gelijk aan de onderlinge afstand van punt i en j op die dimensie waarop hun afstand het kleinst is.
Een derde variant van de Minkowski-afstandsfunctie ontstaat als men toestaat dat p ook waarden tussen -1 en +1 aanneemt. Deze variant wordt voor het tweedimensionale geval gedetailleerd beschreven door Van de Geer (1995), als speciaal geval van een *gegeneraliseerde Minkowski-afstand*, waarin p elke willekeurige waarde tussen min oneindig en plus oneindig kan aannemen. Een bijzondere situatie doet zich voor als p tot nul nadert. Als p een minimaal klein beetje groter dan nul is (laten we zeggen $p = \Delta$), dan is elke afstand gelijk aan plus oneindig (behalve wanneer twee objecten slechts op één dimensie van elkaar verschillen; in dat geval is de afstand zonder meer gelijk aan het betreffende verschil). Is p daarentegen een minimaal klein beetje kleiner dan nul ($p = -\Delta$), dan wordt elke afstand gelijk aan nul (ook weer: behoudens het geval dat twee objecten slechts op één dimensie van elkaar verschillen). Van de Geer (1995) geeft een interessante psychologische interpretatie aan deze twee gevallen; deze zal in Blok 4.2 nader besproken worden.

Merk op dat zowel de oorspronkelijke Minkowski-metriek als bovengenoemde varianten daarvan zijn op te vatten als speciale gevallen van een *algemene machtsmetriek* (power metric) met de formule

$$d_{ij} = [\Sigma_s |x_{is} - x_{js}|^p]^{1/r} \qquad [4.9]$$

De gewogen Euclidische afstandsfunctie
Een variant van de Euclidische afstandsfunctie, de *gewogen Euclidische afstandsfunctie*, ziet er als volgt uit:

$$d_{ij} = [\Sigma_s w_s (x_{is} - x_{js})^2]^{1/2} = [\Sigma_s (w_s^{1/2} x_{is} - w_s^{1/2} x_{js})^2]^{1/2} \qquad [4.10]$$

Deze afstandsfunctie houdt in dat alle coördinaten op dimensie s met het gewicht $\sqrt{w_s}$ vermenigvuldigd (gewogen!) worden. Als er in totaal r dimensies zijn, dan zijn er r van zulke gewichten, waarvan sommige gelijk aan nul

kunnen zijn. Merk op dat de gewichten w_s nooit negatief mogen zijn omdat $\sqrt{w_s}$ dan geen reële waarde heeft. Deze gewogen Euclidische afstandsfunctie wordt vaak als afstandsmodel gebruikt als men meerdere nabijheidsmatrices tegelijkertijd wil analyseren. Deze toepassing komt uitgebreid aan de orde in Hoofdstuk 9.

Andere metrieken

Twee semi-metrieke afstandsfuncties die niet eenvoudig als varianten van de algemene machtsmetriek te schrijven zijn, zijn de zogenaamde *nulmetriek* en de *triviale afstand* (zie Van der Ven, 1977). In de nulmetriek is de afstand tussen twee punten gelijk aan het aantal dimensies waarop beide punten verschillende coördinaten hebben. Hiermee verwant is de triviale afstand. Deze komt erop neer dat alle punten die niet samenvallen dezelfde afstand (bijvoorbeeld 1) tot elkaar krijgen, terwijl alle punten die wel samenvallen uiteraard de afstand 0 hebben.

4.2 PROFIELAFSTANDEN

Hierboven is een aantal afstandsfuncties behandeld waarin afstanden berekend worden als een additieve-verschilfunctie van de coördinaten van punten in een ruimte. Zijn van een verzameling punten de desbetreffende coördinaten bekend, dan kan men eenvoudig de onderlinge afstanden van die punten uitrekenen. Daarbij gaan we er steeds van uit dat de verschillende assen waarop de punten coördinaten hebben loodrecht op elkaar staan (dat wil zeggen onderling onafhankelijk zijn).

Nu kan het zich voordoen dat men de afstanden wil berekenen tussen objecten waarvan men wel een aantal eigenschappen kent, maar waarvan men niet weet wat nu precies de onderliggende, onderling onafhankelijke dimensies zijn, en nog minder, wat hun precieze coördinaten zijn op die dimensies. Bijvoorbeeld: van een aantal personen uit een steekproef weten we (uitgedrukt in centimeters) hun totale lengte, de lengte van hun armen, de lengte van hun benen, en de omvang van hals, borst, taille en heupen. We kunnen nu natuurlijk doen alsof elk van deze zeven afmetingen een aparte dimensie vertegenwoordigt en de desbetreffende aantallen centimeters opvatten als coördinaten op deze dimensies. Vervolgens kunnen we één van de behandelde afstandsfuncties kiezen en voor elk paar personen de onderlinge afstand berekenen.[4] Dit soort afstanden heten *profielafstanden*. Immers, elke persoon wordt gekenmerkt door een profiel van afmetingen, en de afstanden die we berekenen geven ver-

4 Zulke afstanden kunnen zinnige informatie opleveren voor een ontwerper van confectiekleding. Deze kan de afstanden gebruiken om *clusters* personen met min of meer dezelfde lichaamsafmetingen op te sporen en de maten van de kleding daarbij aan te passen.

schillen tussen die profielen weer. Merk op dat het voor zulke profielafstanden uitermate belangrijk is in welke eenheden de kenmerken van het profiel worden uitgedrukt. Merk ook op dat de zeven lichaamsafmetingen onderling gecorreleerd zijn. Er zijn bijvoorbeeld hoge, positieve correlaties te verwachten tussen totale lengte en lengte van armen en benen enerzijds, en tussen hals-, borst-, taille- en heupomvang anderzijds. De drie lengtevariabelen zullen stuk voor stuk ook positief maar minder sterk correleren met de vier omvangsvariabelen. Merk ten slotte op dat het in dit voorbeeld onbekend is welke afstandsfunctie men 'het beste' (in welk opzicht overigens?) kan nemen. Vaak wordt in dit soort gevallen de Euclidische afstandsfunctie gekozen, omdat die het beste aansluit bij onze intuïtieve kennis van het begrip afstand.

Mahalanobis-afstanden

Stel dat men voor de hierboven genoemde lichaamsafmetingen de Euclidische afstandsfunctie kiest om profielafstanden tussen individuen te berekenen. In dat geval doen de verschillen tussen de personen in de hoog correlerende lengterichtingen als het ware drie keer mee in de profielafstand, terwijl verschillen in de eveneens hoog correlerende breedtematen als het ware vier keer meetellen. Om te corrigeren voor de vertekeningen die de correlaties tussen de profielkenmerken op de berekende afstanden teweegbrengen, kan men zogenaamde *Mahalanobis-afstanden* berekenen.

Mahalanobis-afstanden zijn Euclidische afstanden die men berekent door de coördinaten van de punten in het oorspronkelijke assenstelsel zodanig te transformeren dat ze als het ware onafhankelijk van elkaar worden en allemaal even belangrijk worden. Dit is het eenvoudigst weer te geven met behulp van matrixnotatie. De oorspronkelijke matrix met coördinaten $\mathbf{X}$ wordt navermenigvuldigd[5] met een transformatiematrix $\mathbf{T} = \mathbf{S}^{-1/2}$. De matrix $\mathbf{S}^{-1/2}$ berekent men uit de variantie-covariantiematrix $\mathbf{S} = (\mathbf{X} - \overline{\mathbf{X}})'(\mathbf{X} - \overline{\mathbf{X}})/(n-1)$ van de m profielkenmerken over n personen ($\overline{\mathbf{X}}$ is de matrix met n identieke rijen die allemaal de m kolomgemiddelden van $\mathbf{X}$ bevatten). De Mahalanobis-afstand tussen i en j is gelijk aan de wortel uit

$$\begin{aligned} d^2_{ij} &= (\boldsymbol{x}'_i\boldsymbol{S}^{-1/2} - \boldsymbol{x}'_j\boldsymbol{S}^{-1/2})\,(\boldsymbol{x}'_i\boldsymbol{S}^{-1/2} - \boldsymbol{x}'_j\boldsymbol{S}^{-1/2})' \\ &= (\boldsymbol{x}_i' - \boldsymbol{x}_j')\boldsymbol{S}^{-1/2}\boldsymbol{S}^{-1/2}(\boldsymbol{x}_i' - \boldsymbol{x}_j')' \\ &= (\boldsymbol{x}_i' - \boldsymbol{x}_j')\boldsymbol{S}^{-1}(\boldsymbol{x}_i' - \boldsymbol{x}_j')' \end{aligned} \qquad [4.11]$$

5 Waar het op neerkomt is dat er voor ieder punt i en iedere as s nieuwe coördinaten x^*_{is} gemaakt worden die een *lineaire combinatie* zijn van de oorspronkelijke coördinaten. In het tweedimensionale geval geldt: $x^*_{i1} = t_{11}x_{i1} + t_{21}x_{i2}$ en $x^*_{i2} = t_{12}x_{i1} + t_{22}x_{i2}$. De coëfficiënten t_{11}, t_{12}, t_{21} en t_{22} zijn gewichten die aangeven hoe een coördinaat op een 'oude' as (aangeduid door het eerste subscript) gewogen wordt om een coördinaat te krijgen op een 'nieuwe' as (aangeduid met het tweede subscript). Zie ook Hoofdstuk 5.

Het transformeren van **X** door navermenigvuldiging met $\mathbf{S}^{-1/2}$ heeft tot gevolg dat alle assen na transformatie even belangrijk zijn. Daar alle variabelen onder andere gewogen worden met de inverse van hun standaarddeviaties, spelen variabelen met grote varianties (dimensies met grote verschillen tussen de objecten) nu een even grote rol als dimensies met kleine varianties.[6]
Een eenvoudig voorbeeld van een Mahalanobis-afstand is het volgende geval. Stel dat van een (groot) aantal personen de lengte (gemeten in centimeters) en het gewicht (gemeten in kilogrammen) bekend is. In deze steekproef blijken de standaarddeviaties van lengte en gewicht respectievelijk 6,4 centimeter en 11,3 kilogram te zijn. De correlatiecoëfficiënt blijkt .36 te bedragen. Stel dat twee personen, A en B, 165 en 183 centimeter lang zijn en respectievelijk 65 en 78 kilogram wegen. De (Euclidische) profielafstand tussen deze twee personen is dan gelijk aan

$$d_{AB} = [(165 - 183)^2 + (65 - 78)^2]^{1/2} = [324 + 169]^{1/2} = 22.20$$

In deze berekening is geen rekening gehouden met het feit dat lengte en gewicht positief gecorreleerd zijn en verschillende varianties hebben. Willen we nu alleen corrigeren voor het verschil in variantie, dan kunnen we definiëren:

$$d_{AB} = \left(\frac{(165 - 183)^2}{6.4^2} + \frac{(65 - 78)^2}{11.3^2}\right)^{1/2}$$
$$= [(324/41.0) + (169/127.7)]^{1/2} = 3.04$$

In de Mahalanobis-afstandsformule wordt niet alleen gecorrigeerd voor de verschillen in variantie, maar ook voor de correlatie tussen de variabelen. De Mahalanobis-afstand tussen dezelfde twee personen is dan

$$d_{AB} = \left(\frac{1}{1 - .36^2}\right)^{1/2} \bullet \left[\frac{(165 - 183)^2}{6.4^2} + \frac{(65 - 78)^2}{11.3^2} - \frac{2(\bullet 36)(165 - 183)(65 - 78)}{6.4 \times 11.3}\right]^{1/2}$$
$$= 1.0719 \times [(324/41.0) + (169/127.7) - (168.48/72.32)]^{1/2} = 2.82$$

Bij meer dan twee variabelen worden de berekeningen een stuk ingewikkelder.

6 Soms kan $\mathbf{S}^{-1/2}$ niet zonder meer berekend worden, omdat er geen inverse van **S** bestaat (sommige eigenwaarden van **S** zijn dan gelijk aan nul). In die gevallen is het mogelijk om $\mathbf{S}^{-1/2}$ te benaderen uit een benadering van **S**, namelijk door alleen die eigenvectoren van **S** te gebruiken die de grootste bijbehorende eigenwaarden hebben. Immers, **S** is te ontbinden in een product van eigenvectoren **P** en eigenwaarden Λ: $\mathbf{S} = \mathbf{P}\Lambda\mathbf{P}'$ zodat $\mathbf{S}^{-1/2} = \mathbf{P}\Lambda^{-1/2}$. Kiezen we uit **P** de r eigenvectoren ($r < m$) met de grootste eigenwaarden, dan wordt $\mathbf{S}^{-1/2}$ benaderd door een $m \times r$ matrix die de oorspronkelijke m-dimensionale coördinaten uit **X** projecteert op een ruimte met minder dimensies. Op die manier gebruikt men alleen de belangrijkste, onafhankelijke dimensies om de afstanden tussen de punten uit te rekenen. Zo'n benadering van **S** heet een benadering met gereduceerde rang.

Aspectafstanden

Als laatste afstandsfunctie behandelen we hier de zogenaamde *aspectafstanden* die men kan gebruiken als de punten waartussen men de afstanden wil berekenen uit verzamelingen bestaan. Stel dat 'object' *i* gelijk is aan de verzameling {A, B, C} en dat *j* de verzameling {B, C, D, E} aanduidt. De aspectafstand d_{ij} is nu gedefinieerd als

$$d_{ij} = m(i) + m(j) - 2m\,(i \cap j) = m(i \cup j) - m(i \cap j) \qquad [4.12]$$

waarin $m(i)$ en $m(j)$ de aantallen elementen in, respectievelijk, de verzamelingen *i* en *j* aanduiden; $m(i \cup j)$ is het aantal elementen in de *vereniging* van de verzamelingen *i* en *j* en $m(i \cap j)$ is het aantal elementen in de *doorsnede* van *i* en *j*. In woorden: de afstand tussen twee verzamelingen is de som van het aantal elementen dat zij niet gezamenlijk bezitten. Voor de hierboven genoemde verzamelingen *i* en *j* is de afstand dus gelijk aan $d_{ij} = 3 + 4 - (2 \times 2) = 3$.
Ook deze afstandsfunctie is op te vatten als één van de Minkowski-metrieken, namelijk het 'city block'-model (wat in dit geval equivalent is aan de nulmetriek). Vatten we namelijk alle verschillende elementen op als afzonderlijke dimensies, dan is iedere verzameling te karakteriseren door middel van een vector van enen en nullen: een één als het betreffende element tot de desbetreffende verzameling behoort, en een nul als dat niet zo is. In ons voorbeeld zijn er vijf dimensies (A, B, C, D en E); *i* en *j* zijn dus respectievelijk {1, 1, 1, 0, 0} en {0, 1, 1, 1, 1}. De enen en nullen zijn op te vatten als de coördinaten van de verzamelingen op de dimensies. De 'city block'-afstand tussen *i* en *j* is nu gelijk aan de som van de absolute verschillen op de dimensies. Deze verschillen kunnen de waarden 1 of 0 aannemen, zodat d_{ij} domweg gelijk is aan het aantal dimensies waarop *i* en *j* van elkaar verschillen.

4.3 AFSTANDSFUNCTIES ALS MODEL

In het eerste deel van dit hoofdstuk zijn verschillende (families van) afstandsfuncties aan de orde gekomen. Het belang van deze functies en metrieken is dat men ze als *model* kan gebruiken voor empirische, geobserveerde nabijheidsdata. Deze afstandsfuncties kunnen dus fungeren als *afstandsmodellen*, dat wil zeggen, zij worden geacht te beschrijven via welk proces empirische afstandsdata gegenereerd worden. Bij het gebruik van afstandsfuncties als model voor empirische data kan men ruwweg twee toepassingen onderscheiden. In de eerste toepassing gaat het om het berekenen van afstanden tussen objecten waarvan bekend is op welke dimensies zij variëren en welke coördinaten zij op deze dimensies hebben. Bijvoorbeeld: van een aantal kleurenstimuli is de helderheid, de verzadiging en de golflengte van het licht bekend. Verzamelen we nu gelijkenisgegevens voor alle kleurenparen, dan kunnen we nagaan via welke afstandsfunctie de empirische gelijkenisgegevens uit de bekende

coördinaten en dimensies zijn af te leiden. We moeten dan nagaan welke afstandsmodellen een goede overeenstemming (*fit*) tussen afstanden en observaties opleveren, en welk afstandsmodel daarvan de beste voorspellingen geeft. Vinden we een afstandsfunctie die een goede *fit* laat zien tussen de geobserveerde gelijkenissen en de via de functie berekende afstanden, dan is deze afstandsfunctie voorlopig een bruikbaar model voor het (cognitieve) proces dat aan de gelijkenisgegevens ten grondslag ligt. Voorbeelden van onderzoek op dit gebied worden beschreven in Blok 4.1 en 4.2.
De tweede manier waarop men afstandsfuncties als model kan gebruiken, houdt in dat men *aanneemt* dat een verzameling gelijkenisdata via één bepaald model gegenereerd is, en dat men vervolgens de bijbehorende dimensies en coördinaten van de objecten tracht te bepalen. Dit is de manier waarop afstandsmodellen in MDS worden toegepast. Strikt genomen vergt dit type toepassing veel meer vertrouwen in het gekozen afstandsmodel. Er volgt namelijk geen expliciete toets van de houdbaarheid van het model.[7]

In de meeste MDS-toepassingen wordt als afstandsmodel de Euclidische metriek gebruikt. Daarvoor zijn twee oorzaken aan te wijzen: allereerst onze intuïtieve vertrouwdheid met de Euclidische ruimte en het gebrek aan *a priori*-argumenten om een ander model te veronderstellen. Ten tweede bestaan er slechts enkele computerprogramma's die het mogelijk maken om andere afstandsmodellen toe te passen. De voornaamste computerprogramma's hebben niet de mogelijkheid om een andere afstandsfunctie dan de Euclidische te gebruiken[8].

7 Weliswaar zou men in principe de data ook op basis van één of meer andere afstandsmodellen kunnen analyseren en de verschillende *fit*-waarden met elkaar vergelijken, maar dat gebeurt niet vaak. Een van de redenen daarvoor is dat zo'n procedure niet altijd tot duidelijke conclusies zal hoeven leiden. Immers, andere modellen leveren ook optimale, zij het andere, coördinaten van de objecten op, mogelijk op een ander aantal dimensies. Daarom hoeft de optimaliteit (de *fit*) van verschillende oplossingen niet erg veel variatie te vertonen. Met name niet omdat geobserveerde data meestal ook meetfouten bevatten. Om uit te maken wat het 'correcte' model is, is men dus veel meer op inhoudelijke aspecten van de oplossing aangewezen.

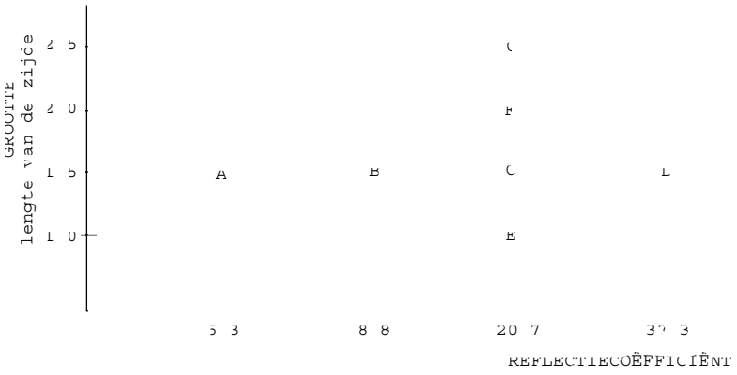
8 Veel computerprogramma's hebben wel de mogelijkheid een variant van het Euclidische afstandsmodel toe te passen: het zogenaamde *gewogen Euclidische afstandsmodel*. Dit model wordt verderop in dit hoofdstuk besproken. Toepassingen ervan vindt men in Hoofdstuk 8 en 9.

Blok 4.1 Afstandsmodellen als onderwerp van onderzoek

In het eerste hoofdstuk van dit boek is er kort stilgestaan bij een onderwerp uit de psychofysica: het onderzoek naar psychofysische krommen die de relatie beschrijven tussen fysieke stimulusintensiteit en intensiteit van de psychologische gewaarwording. Bijvoorbeeld: men laat van een aantal piepjes met een vaste toonhoogte de geluidssterkte variëren en vraagt proefpersonen de geluidsintensiteit van deze piepjes te beoordelen. Uit de beoordelingen kan men schaalwaarden voor de psychologische gewaarwordingen afleiden en die kan men uitzetten tegen de fysieke geluidsintensiteit van de stimuli. Zo krijgt men een psychofysische kromme, die volgens Fechner de vorm van een logaritmische functie heeft.
Een soortgelijk psychofysisch experiment kan men ook uitvoeren voor de *waargenomen gelijkenis* tussen stimuli die op een objectieve, bekende manier van elkaar verschillen, zodat men een objectieve maat voor de gelijkenis tussen de stimuli kan definiëren. De psychofysische kromme is dan de functie die de relatie tussen de objectieve en de waargenomen gelijkenis beschrijft. Doel van zo'n onderzoek is niet alleen het vaststellen van de precieze vorm van die functie. Het gaat vooral om de vraag hoe men de objectieve waarden van de verschillen tussen de stimuli moet combineren om de waarden van de psychologische gelijkenissen zo goed mogelijk te kunnen voorspellen. Met andere woorden: welke afstandsfunctie geeft, bij toepassing op de objectieve eigenschappen van de stimuli, de beste benadering van de waargenomen gelijkenis tussen die stimuli? De kernvraag is dus: waardoor komt het dat dingen op elkaar lijken of van elkaar schijnen te verschillen?
Een klassieke serie experimenten waarin deze vraag onderzocht werd, is in 1950 uitgevoerd door Attneave.[9] In zijn tweede experiment gebruikte hij zeven grijze, papieren vierkantjes papier die varieerden qua grootte en reflectiecoëfficiënt, dat wil zeggen, de mate waarin zij licht reflecteerden.

De zijden van een vierkant konden 1.0, 1.5, 2.0 en 2.5 inches lang zijn; de reflectiecoëfficiënten bedroegen respectievelijk 5.3, 8.8, 20.7 en 37.3 procent. Het design van dit experiment is afgebeeld in Figuur 4.3.

9 Attneave begon zijn artikel met: 'The question "What makes things seem alike or seem different?" is one so fundamental to psychology that very few psychologists have been naïve enough to ask it' (p. 516).

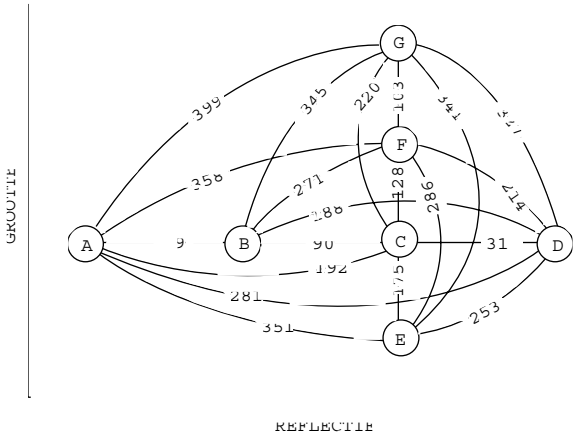


Figuur 4.3 Posities van de zeven vierkantjes op de dimensies grootte en reflectie

Alle vierkantjes werden paarsgewijs met alle andere gecombineerd en op stukken zwart karton geplakt. De 21 paren werden voorgelegd aan 90 proefpersonen, aan wie gevraagd werd de 'overall similarity' van elk paar te beoordelen op een schaal met zeven categorieën die van 'uitermate gelijk' naar 'uiterst verschillend' liep. Deze gelijkenisbeoordelingen werden omgezet[10] in metingen op intervalniveau van de psychologische afstanden tussen de stimuli. Daarna berekende Attneave een geschikte additieve constante; hij koos het getal dat ervoor zorgde dat de volgende vergelijkingen bij benadering juist waren:
$d_{AC} = d_{AB} + d_{BC}$, $d_{AD} = d_{AB} + d_{BD}$, $d_{AD} = d_{AC} + d_{CD}$, $d_{BD} = d_{BC} + d_{CD}$, $d_{EF} = d_{EC} + d_{CF}$, $d_{EG} = d_{EC} + d_{CG}$, $d_{EG} = d_{EF} + d_{FG}$, en $d_{CG} = d_{CF} + d_{FG}$.
De resulterende afstandsschattingen zijn weergegeven in Figuur 4.4.

In de configuratie van vierkantjes vormen de drietallen ACG, ACF, BCG, BCF, ACE, BCE, DCG, DCF, en DCE allemaal rechthoekige driehoeken met C als hoekpunt van de rechte hoek. In elk van deze driehoeken kunnen we de lengte van de hypotenusa voorspellen door de lengten van de rechthoekszijden (dus: de verschillen op de dimensies) in de diverse Minkowski-metrieken te substitueren. Bijvoorbeeld: volgens het Euclidische model is $d_{AG} = (139^2 + 231^2)^{1/2} = 269.6$, volgens het 'city block'-model is $d_{AG} = 139 + 231 = 370$, en volgens de dominantie-metriek is $d_{AG} = 231$.

10 Attneave gebruikte daarvoor zijn *method of graded dichotomies* (1949), een variant van Thurstones *Law of categorical judgment* (vgl. Torgerson, 1958).



Figuur 4.4 Geschatte psychologische afstanden tussen zeven stimuli

De geobserveerde waarde voor d_{AG} is 399, wat het meest overeenkomt met de 'city block'-afstand. Hetzelfde blijkt voor de andere hypotenusa's te gelden. In Attneaves grafiek (1950, p. 533) van de geobserveerde afstanden tegen de 'city block'-voorspellingen liggen de negen punten zeer dicht tegen een rechte lijn aan.

Het bovenstaande resultaat geeft antwoord op de vraag hoe in dit geval (dat wil zeggen: met deze stimuli) psychologische afstanden op verschillende dimensies gecombineerd worden tot een totaalindruk van gelijkenis. Namelijk door te werken volgens het 'city block'-model, dat wil zeggen, door simpele optellingen van de (absolute) verschillen per dimensie. Hiermee weten we nog niet hoe zulke psychologische afstanden samenhangen met grootte en reflectie, de objectieve kenmerken van de stimuli. Attneave heeft die vraag als volgt opgelost:

De psychologische afstanden tussen de vier stimuli met gelijke grootte (A, B, C, D) zette hij uit tegen de verschillen van de logaritmen van hun reflectie. Dus d_{AB} (= 49) werd uitgezet tegen {log(8.8) – log (5.3)}. Analoog zette hij de afstanden tussen C, E, F en G (de stimuli met gelijke reflectie) uit tegen de verschillen van de logaritmen van hun oppervlakte. In beide gevallen leverde dit (nagenoeg) rechte lijnen op, en dus een ondersteuning van de wet van Weber-Fechner. Vervolgens voerde Attneave een multipele regressie-analyse uit, met als afhankelijke variabele de psychologische afstanden tussen de stimuli en de verschillen tussen de logaritmen van reflectie en oppervlakte als onafhankelijke variabelen. Dit leverde de volgende regressieformule op (*R* is reflectie; *O* is oppervlakte):

$$d_{ij} = 3.32 \mid \log R_i - \log R_j \mid + 4.97 \mid \log O_i - \log O_j \mid + .08$$

met bijbehorende multipele correlatiecoëfficiënt van .967 en dus 93% verklaarde variantie. Op grond van deze resultaten trok Attneave (1950) de volgende conclusie.

> *We have seen that ... the psychological distances between [the stimuli] may be conceptualized as distances in a non-Euclidean space; that this space has an axis system which is psychologically fixed, and hence not subject to rotation in the treatment of data*[11]*; and that a multidimensional distance in this space does not differ greatly from the sum of its projections on the axes (p. 551).*

Over deze conclusie valt overigens wel het een en ander op te merken. In de eerste plaats zijn de resultaten van Attneaves andere experimenten minder fraai dan de hierboven gepresenteerde. In de tweede plaats kan men zich afvragen of een monotone transformatie van de psychologische afstanden niet even mooie resultaten in een Euclidische ruimte zou opleveren. In ieder geval zijn er zolang er MDS-technieken bestaan, vele studies verricht waarin de resultaten juist uitstekend met behulp van een Euclidische ruimte verklaard konden worden. Zoals Attneave zelf al opmerkte (1950; zie ook Torgerson 1958), kunnen zijn resultaten in de hand zijn gewerkt door de door hem gekozen dimensies en stimuli. De twee dimensies waarop de stimuli varieerden waren duidelijk van elkaar te onderscheiden en vielen direct op. De geselecteerde stimuli varieerden of op de ene of op de andere dimensie, nooit op beide tegelijkertijd. Shepard (1964) noemde Attneaves stimuli daarom *analyzable* stimuli, in tegenstelling tot *unitary* stimuli, dat wil zeggen stimuli die een eenheid vormen waarin de verschillende dimensies niet gemakkelijk los van elkaar te zien zijn. Een voorbeeld van zulke stimuli zijn kleuren; daarvan kunnen de verschillende aspecten kleursoort, verzadiging en helderheid niet gemakkelijk geïsoleerd worden. Analyseerbare stimuli zouden beter in het 'city block'-model afgebeeld kunnen worden, unitaire stimuli beter in het Euclidische model. In een aantal eigen experimenten vond ook Shepard dat reacties op analyseerbare stimuli beter met het 'city block'-model dan met het Euclidische model verklaard konden worden. Carroll en Wish (1974) maakten enig voorbehoud bij deze conclusie en noemden een mogelijke alternatieve verklaring.

11 Attneave bedoelt hier dat rotatie van de assen niet is toegestaan. Deze kwestie wordt in Hoofdstuk 5 verduidelijkt.

Blok 4.2 Psychologische interpretaties van Minkowski's p

In het begin van dit hoofdstuk hebben we laten zien dat de Minkowski-parameter aangeeft in welke mate de dimensies met grote verschillen tussen twee objecten benadrukt worden ten opzichte van de dimensies waarop de objecten maar weinig van elkaar verschillen. Bij $p = 1$ worden de absolute verschillen langs iedere dimensie zonder meer bij elkaar opgeteld, bij $p = 2$ worden grotere verschillen (wat) zwaarder gewogen, terwijl bij $p = +\infty$ alleen de dimensie met het grootste verschil de afstand bepaalt. Bij deze wiskundige eigenschappen van p kunnen we ons psychologische parallellen voorstellen. Als de afstandsdata die door een proefpersoon gegenereerd worden, overeen zouden komen met Minkowski-afstanden met een bepaalde p-parameter, dan kunnen we ons voorstellen dat die proefpersoon op overeenkomstige manier de nadruk legt op verschillen tussen de objecten.

Van de Geer (1995) beschrijft drie processen, drie intuïtieve regels, die personen zouden kunnen volgen bij het beoordelen van de (on)gelijkenis tussen objecten. Stel dat het om twee objecten in een tweedimensionale ruimte gaat, en dat het verschil tussen beide objecten op de ene as x is en op de tweede as y bedraagt. *Regel 1* luidt dan: als x gelijk of nagenoeg gelijk aan y is, dan kunnen we verwachten dat de verschillen x en y even zwaar (of nagenoeg even zwaar) meetellen bij de beoordeling van de afstand tussen de objecten. In dit geval verwachten we dat een Minkowski-afstandsmodel met $p = 1$ het gedrag van een proefpersoon het beste beschrijft. In het geval dat x (veel) groter is dan y of y (veel) groter is dan x verwachten we dat *Regel 2* opgaat, namelijk dat de waargenomen ongelijkheid van de objecten voornamelijk bepaald wordt door het grootste verschil. Zulk gedrag correspondeert met (zeer) grote waarden van p, in het extreme geval met $p = +\infty$.
Als proefpersonen de gelijkenis tussen objecten moeten beoordelen, dan zitten er paren van objecten bij die op alle dimensies ongeveer evenveel van elkaar verschillen, naast objecten die op de ene dimensie grote verschillen en op de andere dimensie juist kleine verschillen vertonen. Voor sommige objectparen zal Regel 1 het beoordelingsproces het best beschrijven, op andere paren zal Regel 2 meer van toepassing zijn. Als we de geobserveerde gelijkenisbeoordelingen dan willen voorspellen uit de coördinaten van de objecten (even aangenomen dat we daarover beschikken) dan zou een Minkowski afstandsfunctie met een tussenliggende p een optimale *fit* kunnen hebben:

> *The best match might be found by taking n = 2. However, this does not necessarily mean that the subject has such a mathematical model with n = 2 in mind (as if this model 'explains' his responses). What the subject has in mind is just some compromise between Rules I and II (Van de Geer, 1995, p. 27).*

Naast Regel 1 en Regel 2 formuleerde Van de Geer (1995) nog een derde regel. Vrij vertaald komt die op het volgende neer. Als twee objecten op twee of meer dimensies van elkaar verschillen, zal een proefpersoon soms vinden dat zulke objecten uniek zijn en niet met elkaar vergeleken kunnen worden. De afstand tussen zulke objecten is dan oneindig groot. Een vergelijking tussen twee objecten heeft alleen dan zin als ze slechts op één dimensie van elkaar verschillen, of als men slechts op één dimensie zijn aandacht richt en alle andere dimensies buiten beschouwing laat. Deze regel correspondeert met een p die vanaf de positieve kant naar nul nadert ($p \rightarrow +0$). Regel 3 heeft ook nog een tegenhanger: proefpersonen zouden kunnen vinden dat objecten die op meer dan één dimensie van elkaar verschillen, toch aan elkaar gelijk zijn, tenzij ze slechts één dimensie in beschouwing nemen. Deze regel komt overeen met $p \rightarrow -0$. Van de Geers conclusie (p. 31) is dat het beoordelen van (on)gelijkenissen door proefpersonen wel eens op een aantal intuïtieve regels zou kunnen berusten. Drie mogelijke regels zijn de bovengenoemde Regel 1, 2 en 3. De gelijkenisbeoordelingen door individuele proefpersonen zouden uit een compromis tussen deze regels kunnen bestaan.

4.4 RECAPITULATIE

In dit hoofdstuk hebben we gezien dat er verschillende families van afstandsfuncties bestaan die zelf weer een groot aantal varianten bevatten. Deze functies kunnen gebruikt worden om afstanden te berekenen tussen de rijen of de kolommen van iedere willekeurige matrix. In het eerste geval vatten we de getallen in de matrix op als coördinaten van de rij-objecten op assen die met de kolommen van de matrix corresponderen. In het tweede geval behandelen we de elementen van de matrix als coördinaten van de kolomobjecten op assen die door de rijen worden voorgesteld. In Kruskals (1977) termen (zie Hoofdstuk 3) beschouwen we de betreffende matrix dus als een verzameling *multivariate data*. We hebben gezien dat zulke data uit *observaties* kunnen bestaan die nog geanalyseerd moeten worden, of juist het *resultaat* van een of andere analyse kunnen zijn.
Is de matrix met multivariate data klein, dan kunnen we vaak wel 'met de hand' de gewenste afstanden berekenen. Maar bij grote matrices en ingewikkelde afstandsfuncties is het een stuk efficiënter om de computer te gebruiken. In veel pakketten met statistische computerprogramma's zijn speciale modules opgenomen die afstanden kunnen berekenen volgens allerlei verschillende afstandsfuncties. Zo'n module is `PROXIMITIES` uit het SPSS-pakket (zie bijvoorbeeld Norušis, 1994). Met de SPSS-commando's

```
DATA LIST TABLE /STAD 1-9 (A) DIM1 16-20 DIM2 23-26.
BEGIN DATA.
Amsterdam       -27.0    34.6
Rotterdam       -70.3    29.2
Den Haag        -77.7    45.1
Utrecht         -23.9     4.1
Eindhoven       -58.8   -81.7
Arnhem           13.5   -41.0
Zwolle           71.9    -6.4
Groningen       172.3    16.3
END DATA.
PROXIMITIES DIM1 DIM2 /VIEW=CASE
 /MEASURE=EUCLID.
```

kunnen we uit de coördinaten van acht steden (zie Tabel 2.8) hun onderlinge afstanden berekenen (zie Tabel 2.10). In plaats van `MEASURE=EUCLID` hadden we hier ook `MEASURE=MINKOWSKI(2)` of `MEASURE=POWER(2,2)` kunnen gebruiken. De specificatie `VIEW=CASE` zorgt ervoor dat de afstanden tussen de rijen berekend worden; met `VIEW=VARIABLE` krijgen we afstanden tussen de kolommen van de matrix met coördinaten.