

CORRESPONDENTIE-ANALYSE ALS METHODE VOOR HET ANALYSEREN
VAN DE STRUCTUUR VAN CONTINGENTIETABELLEN (*)

K. Matthijs (**)
L. Daemen (***)

Samenvatting

Het artikel belicht correspondentie-analyse als een exploratieve data-analyse techniek. Deze techniek kan vooral nuttig zijn voor het verwerven van inzicht in de structuur van grote tabellen. Na de eigenlijke voorstelling van de techniek en de toelichting van eigenschappen wordt het geheel geïllustreerd met een concreet voorbeeld. Vervolgens komen ook een aantal interpretatieve hulpmiddelen aan bod zoals clusteranalyse en het invoeren van supplementaire empirische informatie. Tenslotte wordt dan kort de veralgemening van correspondentie-analyse naar meerdere variabelen aan de orde gesteld.

(*) Paper voorgebracht tijdens de Sociologendagen 1984 (25-26 april - Vrije Universiteit Amsterdam) in het kader van de sessie "Methoden en Technieken", deelsessie 2: "Analysetechnieken". Aangepaste versie.

(**) K. Matthijs is Doctor in de Sociale Wetenschappen en is verbonden aan het Centrum voor Gezinssociologie van het Departement Sociologie (K.U.Leuven). Correspondentie-adres: Van Evenstraat 2c, 3000 Leuven; 016/22.83.43.

(***) L. Daemen is als assistent verbonden aan de Afdeling Sociologische Theorie en Methoden van het Departement Sociologie (K.U.Leuven). Correspondentie-adres: idem.

0. INLEIDING

Het afgelopen decennium hebben sociologen in toenemende mate gebruik gemaakt van loglineaire modellen voor de analyse van de structuur van contingentietabellen. Dit is een prijzenswaardige evolutie, aangezien loglineaire analyse een theoretisch krachtige techniek is, die bovendien vaak aansluit bij de doelstellingen van sociologen.

Toch zijn er ook een aantal moeilijkheden: loglineaire analyse levert vlug praktische en interpretatieve problemen op wanneer de gekruiste variabelen een groot aantal modaliteiten hebben. Voorliggend artikel bedoelt precies een exposé te bieden van een alternatieve analyse-techniek, correspondentie-analyse genaamd, die een deel van de bezwaren opvangt die verbonden zijn aan loglineaire analyse van grote tabellen.

Correspondentie-analyse is een exploratieve ('inductieve', zeggen sommigen) techniek, die in zijn actuele vorm vooral is ontwikkeld binnen de Franse traditie van de "Analyse des Données". Historisch gezien zet de aldaar ontwikkelde visie zich vrij sterk af tegen de Angelsaksische traditie van de zogenaamde parametrische statistiek. Terwijl deze laatste 'school' zich toelegt op het ontwikkelen van formele methodes voor parameter-schattingen en hypothese-toetsing onder allerlei restrictieve verdelingsassumpties, probeert de eerstgenoemde traditie vooral methodes te ontwikkelen die, los van assumpties, primair als bedoeling hebben bepaalde structuren binnen data-sets te detecteren, respectievelijk te vereenvoudigen.

Vandaag de dag heeft deze visie op data-analyse zich vrij sterk verspreid. Vooral geavanceerde statistici uit de VSA en Nederland leveren daarbij baanbrekend werk. In België daarentegen is genoemde techniek zo goed als ongekend. Dit wordt door de auteurs des te meer als een lacune aangevoeld omdat een veelgehoorde klacht onder sociologen precies betrekking heeft op een vermeend tekort (met name met betrekking tot het verwerven van inzicht in de complexe structuur van 'grote' tabellen) in het statistisch analyse-apparaat. Voorliggend artikel bedoelt precies aan deze klacht tegemoet te komen.

Correspondentie-analyse kan vanuit diverse invalshoeken belicht worden. Benzecri (1973) stelt bijvoorbeeld "nos formules sont à un carrefour où l'on arrive par diverses voies". In voorliggend artikel wordt de techniek beschouwd vanuit het oogpunt van de analyse van een kruistabel volgens een afstandsmodel. Er kan aansluiting worden gevonden bij enige andere modellen uit de multidimensionele schalingstechnologie.

1. OPZET VAN CORRESPONDENTIE-ANALYSE

1.1. Uitgangspunt

Het opzet van correspondentie-analyse bestaat in het voorstellen van de categorieën van de variabelen, die in een kruistabel aan elkaar gerelateerd worden, als punten in een laag-dimensionele ruimte. Stel bijvoorbeeld dat men beschikt over n observaties op twee categorische variabelen A en B met respectievelijk I en J modaliteiten. Er kan dan een kruistabel geconstrueerd worden (met frequenties f_{ij}) waarin het gezamenlijk voorkomen wordt weergegeven van modaliteit i van variabele A en modaliteit j van variabele B. Deze bivariate frequentie-verdeling kan voorgesteld worden in een $(I \times J)$ -matrix F met elementen:

$$(F)_{ij} = f_{ij}$$

De marginale verdelingen

$$f_{i.} = \sum_j f_{ij} \quad i = 1 \dots I$$

$$f_{.j} = \sum_i f_{ij} \quad j = 1 \dots J$$

zijn dan te verzamelen in twee diagonale matrices D_A en D_B . Deze marginale verdelingen kunnen ook verzameld worden in twee kolomvectoren m_A en m_B (1) :

$$m_A = D_A 1_I \quad (m_A)_i = f_{i.}$$

$$m_B = D_B 1_J \quad (m_B)_j = f_{.j}$$

De bedoeling van correspondentie-analyse bestaat er dan in elke rij en kolom van de matrix F ruimtelijk voor te stellen als een punt. Hiervoor zijn twee invalshoeken mogelijk:

- a. een representatie van de categorieën van A als I punten in een J -dimensionele ruimte (R^J) door de rijen van F als coördinaten te gebruiken. Aangezien er telkens een verschillend aantal eenheden in het geding is, wordt aan elk punt een bepaalde 'massa' of 'gewicht' toegekend. Als gewichten wordt de relatieve marginale verdeling van A gebruikt, d.w.z.:

$$n^{-1}D_A$$

- b. een representatie van de categorieën van B als J punten in een I -dimensionele ruimte (R^I) door de kolommen van F als coördinaten te gebruiken. De gewichten van de punten zijn dan gegeven op de diagonaal van:

$$n^{-1}D_B$$

(1) 1_I is een $(I \times 1)$ -vector waarvan alle elementen gelijk zijn aan 1, m.a.w. $(1_I)_i = 1$. Analoog voor 1_J .

Strikt genomen kan de nabijheid van twee punten in de voorstelling van A geïnterpreteerd worden als de mate van gelijkheid tussen de verdelingen van B binnen de categorieën van A, en omgekeerd. Aangezien er gesteund wordt op de absolute frequenties, wordt de spreiding van de punten sterk bepaald door de verschillen in de marginale verdeling van de gerepresenteerde variabele. Dit probleem kan opgelost worden door gebruik te maken van de relatieve conditionele verdelingen. Vandaar een nieuw representatie-voorstel:

- a. voor de voorstelling van de categorieën van A in R^J wordt gebruik gemaakt van de rijen van

$$P_A = D_A^{-1}F$$

Deze matrix bevat de relatieve verdeling van B conditioneel op de modaliteiten van A:

$$(P_A)_{ij} = \frac{f_{ij}}{f_{i.}} = p_{j/i}$$

De massa's van de A-punten blijven dezelfde als deze uit het eerste representatie-voorstel.

- b. voor de representatie van de categorieën van B in R^I wordt gesteund op de kolommen van

$$P_B = F D_B^{-1}.$$

De gewichten blijven opnieuw identiek aan deze uit het eerste voorstel.

De afstanden tussen de punten zijn een belangrijk hulpmiddel bij de interpretatie van een kruistabel. Als rechtstreeks wordt uitgegaan van de rijen van P_A als coördinaten in een R^J , dan is de euclidische afstand tussen twee punten a en b (d_{ab}) als volgt gedefinieerd:

$$\begin{aligned} d_{ab}^2 &= (P_{Aa} - P_{Ab})'(P_{Aa} - P_{Ab}) \\ &= \sum_j (p_{j/a} - p_{j/b})^2 \end{aligned}$$

waarbij p'_{Aa} en p'_{Ab} rijvectoren van P_I zijn. Het heeft echter verscheidene voordelen om gebruik te maken van een andere afstandsmaat. Door elke term in d_{ab}^2 te wegen met de inverse van de marginale proportie eenheden in de categorieën van B, wordt een krachtiger afstandsmaat (δ_{ab}) bekomen:

$$\delta_{ab}^2 = n (p_{Aa} - p_{Ab})' D_B^{-1} (p_{Aa} - p_{Ab}) \quad [1a]$$

$$= \sum_j \frac{1}{p_{.j}} (p_{j/a} - p_{j/b})^2 \quad [1b]$$

Deze afstandsmaat is om een drie-tal redenen interessant. Ten eerste zal een gegeven verschil ($p_{j/a} - p_{j/b}$) sterker geprononceerd worden als het betrekking heeft op een categorie van B met een klein effectief, dan als het betrekking heeft op een categorie van B met een groot effectief. Ten tweede wordt ook de relatie tussen δ_{ab}^2 en de klassieke χ^2 -statistiek duidelijk. Het gaat hier telkens om een maat voor de afstand tussen twee kansverdelingen, beschouwd vanuit het standpunt van een derde verdeling: in δ_{ab}^2 wordt de afstand tussen twee conditionele verdelingen beschouwd vanuit de marginale verdeling, en in χ^2/n wordt de afstand tussen de geobserveerde en de verwachte verdeling bekeken vanuit het standpunt van de verwachte verdeling. Onder meer daarom wordt δ_{ab}^2 soms χ^2 -afstand tussen de punten a en b genoemd (Caillez & Pages, 1975: 444-452). Vanuit deze logica kan ook een relatie worden gelegd tussen correspondentie-analyse en meer traditionele analysetechnieken, zoals χ^2 -, correlatie- en variantie-analyse. Ten derde wordt door het gebruik van de δ_{ab}^2 -afstandsmaat een zekere onafhankelijkheid bekomen van de categorie-indeling. Als bijvoorbeeld de categorieën van B met een gelijke relatieve conditionele verdeling worden samengevoegd, dan wijzigen de afstanden tussen de categorieën van A niet. Deze eigenschap wordt ook wel het 'principe d'équivalence distributionnelle' genoemd (zie o.m. Lebart, Morineau & Tabard, 1977: 50-51; Lebart, Morineau & Fênelon, 1979: 309-310; Caillez & Pages, 1975: 460-462).

De vraag die zich vervolgens opdringt is hoe de coördinaten van de punten moeten voorgesteld worden opdat de (euclidische) afstanden tussen de punten zouden overeenkomen met de χ^2 -afstanden. De te representeren matrix voor de punten van A wordt dan:

$$H_A = \sqrt{n} P_A D_B^{-1/2} \quad [2a]$$

met als elementen:

$$(H_A)_{ij} = \frac{p_{j/i}}{\sqrt{p_{\cdot j}}} = \frac{f_{ij}}{f_{i\cdot}} \sqrt{\frac{n}{f_{\cdot j}}} = h_{Aij} \quad [2b]$$

(dit is eenvoudig te verifiëren door de matrix met weegfactoren nD^{-1} op te splitsen in $D^{-1/2} \sqrt{n} \sqrt{n} D^{-1/2}$ en dit uit te werken tussen de haakjes in [1]).

Voor de afstanden tussen de punten van B kan een analoge redenering worden gevolgd. De punten worden dan gerepresenteerd door de kolommen van:

$$H_B = \sqrt{n} D_A^{-1/2} P_B \quad [3]$$

VOORBEELD

Het voorbeeld (zie Dobbelaere & Billiet, 1975: 65) dat verder in de tekst zal geanalyseerd worden is ontleend aan een onderzoek dat in 1973 werd uitgevoerd bij 449 moeders die in het genoemde jaar een kind in het eerste jaar van het middelbaar onderwijs hadden te Mechelen (een middelgrote stad in Vlaams-Brabant in België). In dit onderzoek werden onder meer metingen gedaan rond kerkelijke integratie en de uitgedrukte voorkeur voor een bepaalde politieke partij. De variabele 'kerkelijke integratie' had vier modaliteiten: vrijzinnig, marginaal katholiek, onregelmatig praktizerend en kerks. De variabele uitgedrukte partijvoorkeur had vijf modaliteiten: CVP (christen-democratisch), VU (Vlaams-nationalistische partij), BSP (socialistisch), PVV (liberaal) en 'geen voorkeur'. De bivariate frequentieverdeling van deze twee variabelen is te verzamelen in matrix F:

	CVP	VU	BSP	PVV	'geen'
F =	1	5	35	7	14
	15	9	43	9	42
	26	18	20	7	30
	109	29	7	2	21

vrijzinnig
 marg. katholiek
 onregel. prak.
 kerks

De marginale verdelingen kunnen voorgesteld worden in volgende vectoren :

$$\begin{aligned}
 m'_A &= (62 \quad 118 \quad 101 \quad 168) \\
 m'_B &= (151 \quad 61 \quad 105 \quad 25 \quad 107)
 \end{aligned}$$

De matrix met de χ^2 -afstanden tussen de punten van kerkelijke integratie is de volgende :

$$\begin{bmatrix}
 0. & .303 & .873 & 2.647 \\
 .303 & 0. & .261 & 1.622 \\
 .873 & .261 & 0. & .744 \\
 2.647 & 1.622 & .744 & 0.
 \end{bmatrix}$$

Er blijkt dat, in χ^2 -afstanden gemeten, de meest gelijkende verdelingen van partij-voorkeur deze van de onregelmatig praktizerenden en de marginaal katholieken zijn. Hetzelfde geldt voor de vrijzinnigen en de marginaal katholieken. De kerksen stellen zich duidelijk anders op dan de vrijzinnigen en de marginaal katholieken.

Voor de voorstelling van de partijvoorkeur verkrijgt men de volgende χ^2 -afstanden:

$$\begin{bmatrix}
 0. & .279 & 2.288 & 1.952 & 1.229 \\
 .279 & 0. & 1.214 & .875 & .455 \\
 2.288 & 1.214 & 0. & .066 & .379 \\
 1.952 & .875 & .066 & 0. & .201 \\
 1.229 & .455 & .379 & .201 & 0.
 \end{bmatrix}$$

De 'kerkelijke integratie' van PVV- en BSP-stemmers lijkt het meest op elkaar, deze van CVP- en BSP-voorstanders lijkt het minst op elkaar. Ook de CVP- en VU-kiezers hebben een gelijksoortige kerkelijkheid. Tussen de PVV- en BSP-groep enerzijds en de CVP- en VU-groep anderzijds situeert zich de groep van personen zonder politieke voorkeur. Nochtans sluit deze groep naar 'kerkelijke integratie' iets meer aan bij de BSP- en PVV-groep dan bij de CVP- en VU-stemmers.

1.2. Eigenschappen

De matrices H_A en H_B definiëren twee puntenwolken (respectievelijk in R^J en R^I) die een aantal interessante eigenschappen bezitten. Een eerste eigenschap betreft de positie van een punt ten opzichte van het zwaartepunt van die wolk. Dit zwaartepunt is gedefinieerd als het gewogen gemiddelde van de punten op de respectievelijke assen. Voor de representatie van A in R^J is dit derhalve

$$g_A = n^{-1} H_A' m_A$$

met

$$(g_A)_j = \sum_i p_i \cdot h_{Aij}$$

of uitgewerkt

$$(g_A)_j = \sqrt{p_{\cdot j}} = \sqrt{\frac{f_{\cdot j}}{n}} \quad [4]$$

De afstand van een punt a tot het zwaartepunt is dan:

$$\delta_{ga}^2 = (h_{Aa} - g_A)' (h_{Aa} - g_A)$$

of uitgewerkt

$$\delta_{ga}^2 = \sum_j \frac{1}{p_{\cdot j}} (p_{j/a} - p_{\cdot j})^2 \quad [5]$$

Indien een conditionele verdeling op een bepaalde categorie van A gelijk is aan de marginale verdeling van B , dan valt dit punt samen met het zwaartepunt. In het geval van volledige statistische onafhankelijkheid in tabel F , vallen alle punten samen met het zwaartepunt. Dit laat een verband vermoeden met de χ^2 -statistiek. De gewogen som van de afstanden van de punten tot het zwaartepunt is inderdaad direct gerelateerd aan de χ^2 -statistiek:

$$\begin{aligned}\sum_i p_i \cdot \delta_{gi}^2 &= \sum \sum \frac{1}{f_i \cdot f_j} (f_{ij} - \frac{f_i \cdot f_j}{n}) \\ &= \frac{\chi^2}{n}\end{aligned}\quad [6]$$

Elke afstand tot het zwaartepunt is dus een bijdrage tot de χ^2 -statistiek. De bijdrage van elke categorie aan de totale associatie in de tabel kan ook proportioneel uitgedrukt worden:

$$\pi_{Aa} = \frac{f_a \cdot \delta_{ga}^2}{\chi^2} \quad [7]$$

χ^2/n is dus een aanduiding van de spreiding (dispersie) van de puntenwolk rond het zwaartepunt. Dit wordt ook wel de 'inertie in het zwaartepunt' genoemd. Met de notie 'inertie in een punt' wordt de afstand aangeduid van een punt tot een puntenwolk. In het zwaartepunt is de inertie minimaal. Deze minimale inertie wordt soms aangeduid als 'totale inertie'. (1)

Een tweede belangrijk element is de (cosinus van de) hoek tussen twee punten bekeken vanuit het zwaartepunt. Deze maat is vergelijkbaar met een correlatie-coëfficiënt. Het geeft een uitdrukking van de mate waarin de afwijkingen van het zwaartepunt samen covariëren over de categorieën van de niet-gerepresenteerde variabele (dus B). Als de cosinus van de hoek gelijk is aan nul (d.w.z. dat de punten een hoek van 90° vormen) dan covariëren de afwijkingen niet met elkaar; als de cosinus positief is (dus een hoek tussen 0° en 90°) dan zijn er afwijkingen in dezelfde zin; is de cosinus negatief (een hoek tussen 90° en 180°) dan is er sprake van covariatie in omgekeerde zin. De cosinusmaat geeft derhalve richting aan de χ^2 -afstand. Bovendien is deze maat onafhankelijk van de afstand van de punten tot het zwaartepunt.

(1) Zwaartepunt en inertie zijn generalisering in R^q van de begrippen gemiddelde en variantie in R. Voor verdere uitleg rond deze termen, zie o.m. Caillez & Pages, 1975: 207-208, 226-237 en Bouroche & Saporta, 1983: 30-32.

De cosinus van de hoek tussen de punten a en b van variabele A kan als volgt worden geschreven (Van De Geer, 1971: 19):

$$\cos(a,b) = \frac{1}{\sigma_{ga}\sigma_{gb}} (h_{Aa} - g_A)'(h_{Ab} - g_A) \quad [8a]$$

of, na uitwerking

$$\cos(a,b) = \frac{1}{\sigma_{ga}\sigma_{gb}} \sum_j \frac{1}{p_{.j}} (p_{j/a} - p_{.j})(p_{j/b} - p_{.j}) \quad [8b]$$

Een analoge redenering voor de verhouding van de punten t.o.v. het zwaartepunt kan worden herhaald voor de punten B in R^I .

Uit het voorgaande blijkt dat het zinvol is om de respectievelijke puntenwolken te bekijken ten opzichte van hun zwaartepunt. In een ruimtelijke voorstelling kan dit gerealiseerd worden door de oorsprong van de ruimtes te laten samenvallen met de respectievelijke zwaartepunten. Dit gebeurt door de volgende matrices te representeren:

$$H_{A*} = H_A - 1_I g_A' \quad [9]$$

$$H_{B*} = H_B - g_B 1_J' \quad [10]$$

Boven zijn ruimtes geconstrueerd waarin de categorieën van de betrokken variabelen volgens een afstandsmodel worden voorgesteld als punten. De afstand tussen 2 punten drukt de gelijkheid tussen categorieën uit met betrekking tot hun samenstelling uit de groepen die door de andere variabele gedefinieerd zijn. In deze zin is correspondentie-analyse een multi-dimensionele schaaltechniek waarin gelijkheid-data volgens een afstandsmodel gerepresenteerd worden (Heiser & Meulman, 1983).

VOORBEELD

De afstanden van de punten van kerkelijke integratie tot het zwaartepunt zijn de volgende:

vrijzinnig	marg. kath.	onregelm. prakt.	kerks
(.854	.295	.055	.547)

Hiermee stemmen de volgende proportionele bijdragen tot χ^2
(= 185,132) overeen:

(.286 .188 .030 .494)

De vrijzinnigen hebben de sterkste afwijkingen van de marginale verdeling, vervolgens de kerksen, de marginaal katholieken en tenslotte de onregelmatig praktizerenden. De orde tussen de vrijzinnigen en de kerksen is voor wat betreft de bijdrage aan de inertie precies omgekeerd als voor wat betreft de afwijkingen van de marginale verdeling. Dit is te wijten aan de grotere omvang van de kerksen (zie hoger).

De matrix met de cosini tussen de rijpunten is :

1.	.843	.084	-.912
.843	1.	.349	-.971
.084	.349	1.	-.410
-.912	-.971	-.410	1.

De afwijkingen die de kerksen naar partijvoorkeur vertonen ten opzichte van de marginale verdeling zijn dus tegengesteld aan deze die betrekking hebben op de minder kerkelijk geïntegreerden. Deze laatste groepen daarentegen vertonen afwijkingen in dezelfde zin.

De afstanden van uitgedrukte partijvoorkeur tot het zwaartepunt zijn de volgende:

CVP	VU	BSP	PVV	'geen'
(.562	.123	.616	.426	.163)

Hiermee stemmen volgende bijdragen tot de totale inertie overeen:

(.458 .041 .349 .006 .094)

De verdeling van de CVP-, BSP-, en in mindere mate ook de PVV-kiezers, verschilt het sterkst van de algemene verdeling. De bijdrage aan de totale inertie van de PVV-groep is relatief klein (.006) omwille van de kleine omvang van deze groep. Hetzelfde geldt in omgekeerde zin voor de CVP- en BSP-kiezers (zie hoger).

De matrix met de cosini tussen de kolompunten is:

$$\begin{bmatrix} 1. & .772 & -.943 & -.984 & -.834 \\ .772 & 1. & -.865 & -.712 & -.601 \\ -.943 & -.865 & 1. & .953 & .633 \\ -.984 & -.712 & .953 & 1. & .731 \\ -.834 & -.601 & .631 & .736 & 1. \end{bmatrix}$$

Er treden duidelijk twee onderscheiden blokken naar voor. Enerzijds een groep geconstitueerd door de CVP en VU, anderzijds een groep samengesteld uit de BSP, de PVV en de 'geen voorkeur'-subgroep. Binnen deze groepen zijn er gelijklopende afwijkingen van de totale marginale verdeling van kerkelijke integratie. Tussen deze groepen zijn de afwijkingen tegengesteld aan mekaar.

2. REPRESENTATIE IN EEN RUIMTE VAN LAGERE DIMENSIONALITEIT

In het vorige punt is aangetoond dat het mogelijk is de categorieën van A als punten in een R^J (evenals de categorieën van B in een R^I) zodanig grafisch voor te stellen dat er een aanduiding gegeven wordt van de structuur van een kruistabel. Als het aantal categorieën van de variabelen groter wordt dan drie (en het is vooral voor dergelijke tabellen dat correspondentie-analyse werd ontwikkeld), dan is zulk een voorstelling een weinigzeggende abstractie. Vandaar dat de assen zó zullen gerooteerd worden dat zoveel mogelijk informatie in de puntenwolk wordt gevat uit een subruimte (een hypervlak) die wordt bepaald door een beperkt aantal betekenisvolle assen. De assen worden zo verdraaid dat de punten op een aantal van deze assen zeer lage projecties hebben (in absolute waarde), zodat deze kunnen verwaarloosd worden.

2.1. Basisnoties

Als de gewogen kwadratensom genomen wordt als maat voor de hoogte van de projecties, dan is het gestelde probleem analoog aan een klassiek principale componenten-probleem (Van De Geer, 1971: 128-140). De eigenvectoren U_k en V_k van de respectievelijke 'cross-product'-matrices

$$S_A = n^{-1} H_{A*}' D_A H_{A*} \quad [11]$$

$$S_B = n^{-1} H_{B*}' D_B H_{B*} \quad [12]$$

definiëren de nieuwe referentie-assen in de respectievelijke representaties van A en B, en de eigenwaarden λ_{Ak} en λ_{Bk} geven de inertie aan van de projecties van de punten op deze nieuwe assen. Deze projecties (de principale componenten) zijn dan gedefinieerd als:

$$X = H_{A*} U \quad [13]$$

$$Y = H_{B*}' V \quad [14]$$

Het optimale hypervlak R^q wordt bepaald door de eerste q principale assen (u_q en v_q) zodat de matrices met de eerste q principale componenten kunnen aangeduid worden als X_q en Y_q .

Er stellen zich derhalve twee aparte problemen (i.e. één m.b.t. de representatie van A, één m.b.t. die van B). Deze kunnen echter herleid worden tot één geïntegreerd probleem via een singuliere waarde decompositie (SVD) van de matrix.

$$\begin{aligned} K &= n^{-1/2} D_A^{1/2} H_{A*} = H_{B*}' D_B^{1/2} n^{-1/2} \\ &= D_A^{-1/2} F D_B^{-1/2} - g_B g_A' \end{aligned} \quad [15]$$

omdat geldt dat

$$K'K = S_A$$

$$KK' = S_B$$

zodat voor K geldt

$$K = V \Lambda^{1/2} U' \quad V'V = I; U'U = I \quad [16]$$

Er blijkt dus dat de eigenwaarden van S_A en S_B gelijk zijn, zodat in ieder van de twee representaties geldt dat de overeenkomstige componenten eenzelfde inertie hebben.

De samenhang tussen de twee geformuleerde problemen blijkt verder ook uit het feit dat

$$X = H_{A*} U = n^{1/2} D_A^{-1/2} V \Lambda^{1/2}$$

$$Y = H_{B*} V = n^{1/2} D_B^{-1/2} U \Lambda^{1/2}$$

of na substitutie

$$X = P_A Y \Lambda^{-1/2} \quad [17]$$

$$Y = P_B' X \Lambda^{-1/2} \quad [18]$$

Op de factor met de eigenwaarden na blijken de punten die de categorieën van A representeren het zwaartepunt te zijn van de punten die de categorieën van B representeren, mits deze gewogen worden met de relatieve verdeling van B binnen de betreffende categorieën van A. Als x'_a en $\tilde{y}_j$ rijvectoren zijn van respectievelijk X en $\tilde{Y} = Y \Lambda^{-1/2}$, dan is

$$x_a = \sum_j p_{j/a} \tilde{y}_j \quad [19]$$

Het omgekeerde geldt natuurlijk ook voor de punten van B.

Dit is een belangrijk gegeven omdat hier voor het eerst duidelijk de relatie tussen de modaliteiten van de variabelen naar voor komt. Uitgaande van dit gegeven kan men zich een beeld vormen van de wijze waarop de relaties tussen beide variabelen zich voltrekken. De factor met de eigenwaarden fungeert dan als factor die aangeeft in welke mate dit principe (ook wel het 'principe barycentrique' genoemd) verstoord wordt op de respectievelijke componenten. Beide puntenwolken kunnen immers in het algemeen niet tegelijkertijd zwaartepunt zijn van elkaar, zodat een

versoepeling via zulk een factor zich opdringt. Het is mogelijk correspondentie-analyse vanuit deze eigenschap af te leiden (Lebart, Morineau & F  nelon, 1977: 57-59). Het zijn genoemde relaties die een gezamenlijke voorstelling van beide representaties motiveren door deze gewoon over elkaar te leggen. Hier vind men ook aansluiting bij de multidimensionele schaaltechnieken, nl. bij het zgn. 'reciprocal gravity model'. Dit model is verwant met het ontvouwingsmodel, en representeert nabijheids-data volgens een afstandsmodel (Heiser, 1981).

2.2. Verdere uitwerkingen met consequenties

2.2.1. Het aantal assen dat maximaal nodig is om een perfecte representatie te krijgen is gelijk aan $t = \min(I - 1, J - 1)$. Stel $I < J$. De representatie van de I punten van A in R^J kan altijd gevat worden in een hypervlak R^I . Zo kunnen bijvoorbeeld twee punten in R^5 altijd gevat worden in een vlak, als de oorsprong mee opgenomen wordt. Overigens 'bespaart' men in zekere zin nog een as aangezien de marginale verdelingen vast liggen zodat er een lineaire afhankelijkheid in de gegevens bestaat.

2.2.2. Bij de gezamenlijke voorstelling van beide representaties dient erop gelet dat de factor $\Lambda^{-1/2}$ het 'principe barycentrique' verstoort. Vandaar dat er vanuit een bepaald perspectief kan gepleit worden voor het doorbreken van de symmetrie door de punten van de   ne variabele werkelijk als zwaartepunt van de punten van de andere variabelen voor te stellen. Dit kan op een tweetal wijzen gebeuren (bijv. A als zwaartepunt van B) :

- a. de locatie van de punten van A wijzigen door A te representeren als $\tilde{X} = X \Lambda^{1/2}$ zodat $\tilde{X} = P_A Y$ (Bouroche & Saporta, 1983: 95; Caillez & Pages, 1975: 456-459; Escoufier, 1982: 65-67);
- b. de locatie van de punten van B wijzigen door B te representeren door $\tilde{Y} = Y \Lambda^{-1/2}$, zodat $X = P_A \tilde{Y}$ zoals hoger reeds werd gesuggereerd (Gifi, 1981: 134-136).

Hierdoor kan de interpretatie van de afstanden en hoeken tussen de verplaatste punten niet meer gehandhaafd blijven.

2.2.3. De elementen van K zijn gelijk aan

$$k_{ij} = \frac{1}{\sqrt{n}} \frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}} \quad \text{met } e_{ij} = \frac{f_{i.} \cdot f_{.j}}{n}$$

Er wordt dus in feite een SVD uitgevoerd op de gestandaardiseerde residuen van het onafhankelijkheidsmodel. Als K wordt uitgedrukt in termen van X en Y, is

$$K = n^{-1} D_A^{1/2} X \Lambda^{-1/2} Y' D_B^{1/2}$$

$$(K)_{ij} = \frac{1}{n} \sqrt{f_{i.} \cdot f_{.j}} \sum_k \frac{1}{\sqrt{\lambda_k}} x_{ik} y_{jk}$$

waarna uitdrukking in termen f_{ij} geeft

$$F = n^{-1} D_A (X \Lambda^{-1/2} Y' + 1_I 1_J') D_B \quad [20a]$$

$$f_{ij} = e_{ij} \left(1 + \sum_k \frac{1}{\sqrt{\lambda_k}} x_{ik} y_{jk} \right) \quad [20b]$$

De principale componenten 'verklaren' dus waarom de geobserveerde frequenties afwijken van de frequenties die op basis van de marginalen kunnen verwacht worden als de kenmerken onafhankelijk zijn. Als de tabel F gerepresenteerd wordt in een R^Q -hypervlak dan kan steunend op [20] een benaderde tabel $\hat{F}$ gereproduceerd worden:

$$\hat{F}_q = n^{-1} D_A (X_q \Lambda_q^{-1/2} Y_q' + 1_I 1_J') D_B.$$

Het is eenvoudig aantoonbaar dat de marginale frequenties van $\hat{F}$ gelijk zijn aan die van F. De marginalen blijven dus gelijk. $(F - \hat{F})$ geeft dan het absoluut aandeel van de frequenties die niet in R^Q kunnen gerepresenteerd worden, of m.a.w. de residuele frequenties.

2.2.4. De SVD heeft kleinste kwadraten-eigenschappen met betrekking tot de relatie tussen K en $\hat{K}_q$, gegeven dat $\hat{K}$ perfect kan gerepresenteerd worden in q dimensies. Dit betekent dat $\hat{F}_q$ de tabel is die in kleinste kwadratenzin het meest gelijk op F , maar die toch in q dimensies kan weergegeven worden (Eckart & Young, 1936).

2.2.5. De relaties tussen A en B kunnen nog op een andere wijze grafisch voorgesteld worden:

$$\begin{aligned} H_{A*} &= n^{1/2} D_A^{-1/2} K \\ &= n^{-1/2} X \Lambda^{-1/2} Y' D_B^{1/2} \\ &= n^{-1/2} X \tilde{Y}' D_B^{1/2} \\ n P_A D_B^{-1} - 1_I 1_I' &= X \tilde{Y}' \end{aligned}$$

Als dus de punten van de categorieën van A in de asymmetrische voorstelling met behulp van methode b (zie punt 2.2.2) geprojecteerd worden op de lijn die de oorsprong verbindt met een punt b van B , dan zijn deze

projecties gelijk aan $p_{b/i} \frac{1}{p_{.b}} - 1 = \frac{f_{ij} - e_{ij}}{e_{ij}}$. In een gereduceerde

ruimte R^q geldt hierbij de gewogen (met e_{ij}/n) kleinste kwadrateneigenschap. Het vergelijken van de projecties van de punten van A op één zulke a of b geeft dus een directe aanduiding van de verschillen in de conditionele kansen van de $p_{b/i}$. Correspondentie-analyse sluit dus aan bij de scalair-produkt-modellen uit de multi-dimensionele schalingstechnologie (bv. Carroll, 1972).

2.2.6. De inertie (spreiding) die wordt opgenomen door elke a is gelijk aan λ_k . Hieraan kan onmiddellijk de vraag worden gekoppeld wat de opgenomen inertie is in het R^q hypervlak. Deze is gewoon gelijk aan de som (Caillez & Pages, 1976: 228-243):

$$\sum_{k=1}^q \lambda_k.$$

De totale inertie is dan gelijk aan

$$\text{tr}(\Lambda) = \frac{\chi^2}{n} = \sum_{k=1}^q \lambda_k + \sum_{k=q+1}^t \lambda_k$$

Dit kan ook proportioneel uitgedrukt worden. τ_k is dan de 'proportie verklaarde inertie'. Het gaat hier dus om de proportie verklaarde χ^2 door een component k .

$$\tau_k = \frac{\lambda_k}{\text{tr}(\Lambda)}$$

De proportie verklaarde inertie in een R^q kan dan ook aangeduid worden als:

$$\tau_{1,q} = \sum_{k=1}^q \tau_k$$

Deze laatste maat geeft aan in welke mate de totale spreiding in de puntenwolken kan weergegeven worden in een R^q . Hoe groot deze waarde moet zijn voor een adequate benadering van de relatiestructuur wordt zelden vermeld. Sikkel (1980: 18) vermeldt het o.i. vrij hoge .95.

2.3. De notie 'absolute' en 'relatieve bijdrage'

Een vraag die zich in het kader van het gegeven exposé opdringt is welke de bijdrage is van een bepaalde categorie a aan de inertie die door een as wordt opgenomen. De door een component opgenomen inertie is gelijk aan

$$\lambda_k = \sum_{i=1}^I p_i \cdot x_{ik}^2.$$

De proportionele bijdrage van een categorie a tot die component is dan

$$\theta_{ak} = \frac{p_a \cdot x_{ak}^2}{\lambda_k}$$

Deze maat drukt het relatieve belang uit die een categorie heeft bij de constructie van een as. Dit is vooral interessant als hulpmiddel bij de interpretatie van de assen: de θ_{ik} kunnen een licht werpen op de relevante categorieën voor een as of dimensie. Soms wordt deze maat ook 'contribution absolue' genoemd (Lebart, Morineau & F  nelon, 1977: 60-61).

Boven is met $\tau_{1,q}$ aangegeven hoe goed een puntenwolk kan gerepresenteerd worden in een R^q -hypervlak. Dezelfde vraag kan ook gesteld ten aanzien van de categorie  n op zichzelf. Een categorie a levert een bepaalde bijdrage tot de totale inertie, nI .

$$p_a \cdot \sum_{j=1}^J h_{A*aj}^2 = p_a \cdot \sum_{k=1}^t x_{ak}^2 = p_a \cdot \delta_{ga}^2.$$

Uitdrukking van de bijdrage van een categorie a in een proportie geeft dan

$$\omega_{ak} = \frac{p_a \cdot x_{ak}^2}{p_a \cdot \delta_{ga}^2} = \frac{x_{ak}^2}{\delta_{ga}^2}.$$

Deze maat geeft het relatief belang aan van de respectievelijke assen voor de door een bepaalde categorie aangebrachte inertie. De som

$$\sum_{k=1}^q \omega_{ak}$$

geeft de proportie inertie aan die door R^q gerepresenteerd wordt. Als dit voor een bepaalde categorie zeer laag is, ongeacht een hoge $\tau_{1,q}$, dan verliest men voor deze categorie een groot deel informatie, ondanks een eventueel adequaat gerepresenteerde wolk. Deze mate wordt soms ook 'contribution relative' genoemd (Lebart, Morineau & F  nelon, 1977: 61-62).

VOORBEELD

In het voorbeeld waarbij 4 types van kerkelijke integratie uitgezet worden tegenover de uitgedrukte partijvoorkeur wordt de volgende set met eigenwaarden bekomen:

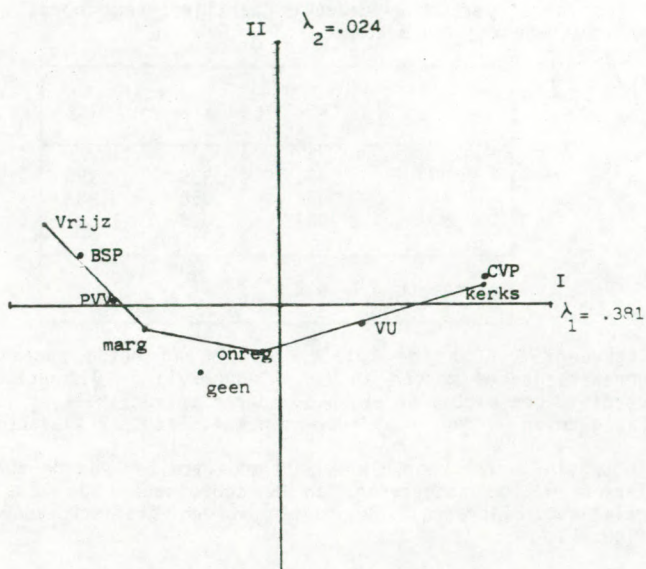
k	λ_k	τ_k	$\tau_{1.k}$
1	.381	.925	.925
2	.024	.058	.983
3	.007	.017	1.
$\chi^2/n = \text{tr}(\Lambda) .412$			

Ongeveer 98 % van de totale ingrtie van beide puntenwolken kan gerepresenteerd worden in een R^2 -hypervlak. Dit wil zeggen dat reeds in een plat vlak een quasi-perfecte voorstelling van de tabel kan gegeven worden. De derde as neemt slechts 2 % associate op.

In tabel 1 wordt voor elk van de modaliteiten van de twee variabelen de waarde aangegeven van de coördinaten, de absolute en de relatieve bijdragen. De punten worden grafisch voorgesteld in figuur 1.

De voorstelling in figuur 1 geeft een concreet voorbeeld van de in het eerste deel besproken indices: afstanden tot de oorsprong, afstanden tussen punten van één variabele en hun hoeken. De plot geeft dus een direct visueel en grafisch beeld van de interne structuur van een tabel. Overigens zal een dergelijke plot vooral dienstig zijn als de tabel een omvangrijk aantal modaliteiten heeft.

Figuur 1. Plot van de relatie in de modaliteiten van 'kerkelijke integratie' en 'politieke voorkeur' (mogelijkheid 1).



De eerste as uit figuur 1 heeft betrekking op een dimensie die vrijwel alle associatie verklaart. Deze as plot enerzijds de vrijzinnigen tegenover de kerksen, met tussenliggend de marginale katholieken en de onregelmatig praktizerenden, en anderzijds de BSP- en PVV-voorkeur tegenover de CVP-stemmers met tussenliggend geen- en VU-voorkeur. Deze dimensie kan dus geduid worden als een kerke- lijks- of confessionaliteitsdimensie.

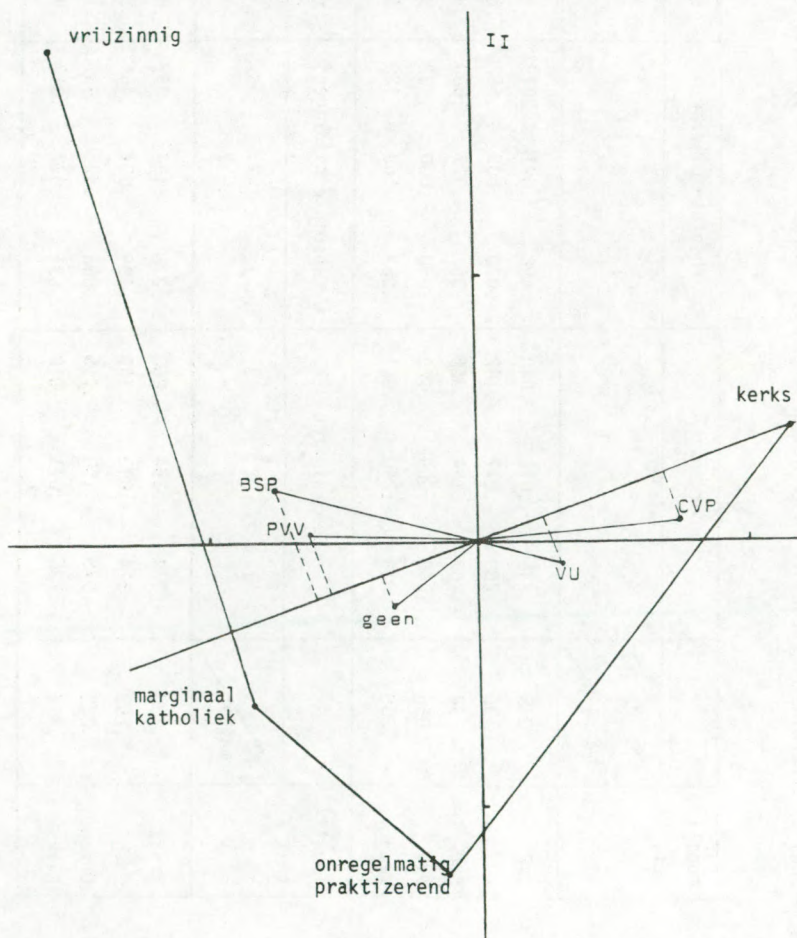
Tabel 1. Coördinaten, absolute en relatieve bijdragen voor elk van de punten.

modali- teit 1		coördinaten			absolute bijdragen			relatieve bijdragen		
	k massa	1	2	3	1	2	3	1	2	3
CVP	.336	.745	.072	.040	.490	.071	.079	.988	.009	.003
VU	.136	.289	-.077	-.182	.030	.033	.652	.682	.048	.270
BSP	.234	-.763	.184	.016	.357	.327	.009	.945	.055	.000
PVV	.056	-.640	.009	-.128	.060	.000	.132	.962	.000	.038
geen	.238	-.319	-.240	.061	.063	.568	.129	.624	.353	.023
modali- teit 2		coördinaten			absolute bijdragen			relatieve bijdragen		
	k massa	1	2	3	1	2	3	1	2	3
Vrijz.	.138	-.874	.294	-.068	.276	.494	.092	.894	.101	.005
Marg.	.263	-.524	-.093	.109	.189	.094	.454	.930	.029	.040
Onreg.	.225	-.076	-.190	-.116	.003	.335	.437	.104	.653	.243
Kerks	.374	.736	.071	.018	.531	.077	.017	.990	.009	.001

Opmerkelijk is wel de positie van het punt 'geen voorkeur'. Dit punt neemt een tussenpositie in, met een lichte neiging naar het niet-confessionele.

De tweede as contrasteert de middenposities van de eerste as met haar extremen. Dit is de zgn. hoefijzerstructuur (zie Schriever, 1983). Wellicht kan dit geïnterpreteerd worden als een intensiteitsdimensie (Guttman, 1950, 1955). Op dit probleem zal hier niet dieper ingegaan worden.

Figuur 2. Plot van de relatie in de modaliteiten van 'kerkelijke integratie' en 'politieke voorkeur' (mogelijkheid 2).



Deze interpretatie kan ook duidelijk afgelezen worden uit de absolute bijdragen (zie tabel 1). Uit de relatieve bijdragen blijkt dat vrijwel alle punten goed gerepresenteerd worden in een R^2 . De uitzonderingen hierop zijn minimaal.

Tot dusver is de interpretatie van de plot beperkt gebleven tot de variabelen afzonderlijk. Via het 'principe barycentrique' kunnen daarnaast ook specifieke relaties tussen de categorieën van beide variabelen afgeleid worden. Zo blijkt CVP-voorkeur vooral aanwezig te zijn bij de kerksen, PVV- en BSP-voorkeur bij de vrijzinnigen en de marginaal katholieken, en VU-voorkeur vooral bij de onregelmatig praktizerenden en de kerksen. Voor detail-analyses zijn deze relaties gemakkelijker te ontdekken door de punten van één variabele zodanig te verschuiven dat deze rechtstreeks als zwaartepunt van de punten van de andere variabele kunnen beschouwd worden. In het voorbeeld (figuur 2) zijn de punten voor kerkelijke integratie zodanig verschoven dat de punten van partijvoorkeur het zwaartepunt zijn van de kerkelijkheidspunten (i.c. optie b uit punt 2.2.2.: representatie van kerkelijkheid als $\tilde{Y} = Y \Lambda^{-1/2}$). De relaties tussen de variabelen zijn hier duidelijk visueel weergegeven. Bovendien kan de partijvoorkeur geprojecteerd worden op de as die een kerkelijkheidspunt verbindt met de oorsprong. De verschillen tussen deze projecties geven de verschillen aan in de kans dat iemand een bepaalde 'kerkelijke integratie' heeft, gegeven een bepaalde uitgedrukte politieke voorkeur. Dit geeft dezelfde interpretatie als wanneer gesteund wordt op het 'principe barycentrique'. Wel wordt iets meer detail-informatie geboden.

3. INTERPRETATIEVE HULPMIDDELEN

In dit punt zal dieper ingegaan worden op een aantal hulpmiddelen die kunnen aangewend worden ter interpretatie van de bekomen resultaten. Daarvoor zal een onderscheid gemaakt worden tussen interne en externe criteria.

3.1. Interne procedure: clusteranalyse

Bij de situering van correspondentie-analyse is uitgegaan van een grafische voorstelling om een inzicht te krijgen in de χ^2 -afstanden tussen de categorieën van de respectievelijke variabelen. Een alternatieve manier

om deze doelstelling te realiseren is clusteranalyse. Dit analyse-procédé bedoelt groepen van punten te detecteren die intern homogeen en extern heterogeen zijn. Er is dus een zeker parallelisme tussen clusteranalyse en correspondentie-analyse, zodat het uitvoeren van een supplementaire clusteranalyse in zekere zin redundant is. Toch kan clusteranalyse enig inzicht bijbrengen omdat in rekening kan gebracht worden dat (1) de punten een verschillend gewicht hebben, en (2) het in correspondentie-analyse geselecteerde hypervlak meer dan twee dimensies kan hebben.

Als de I categorieën van A opgedeeld zijn in k groepen, dan kan de totale inertie I_t opgesplitst worden in de inertie tussen de groepen (I_b) en de inertie binnen (I_w) de groepen (Bouroche & Saporta, 1983: 50-51):

$$I_t = I_b + I_w$$

waarbij I_t gelijk is aan χ^2_T/n en I_b aan χ^2_b/n , of de χ^2/n -waarde van de in k categorieën samengetrokken tabel, zodat nI_w de reductie in χ^2 weergeeft door de $(I \times J)$ -tabel samen te trekken tot een $(k \times J)$ -tabel.

Voor de concrete toepassing van clusteranalyse kan gesteund worden op verscheidene van de procedures die terzake werden ontwikkeld, en die ruwweg kunnen opgedeeld worden in twee groepen: hiërarchische agglomeratieve en iteratieve partitioneringsprocédés. De hiërarchische techniek komt neer op de toepassing van Ward's algoritme (Ward, 1967) op de matrix H_A , zij het dat de rijen differentieel gewogen worden (Israël, 1979). De iteratieve techniek kan uitgevoerd worden aan de hand van één van de vele algoritmes die het spoor van de 'pooled within-groups scatter-matrix' van H_A minimaliseren (ook met differentieel gewogen rijen) (Everitt, 1974: 26 e.v.).

3.2. Externe procedure: invoeren van supplementaire informatie

De externe invalshoek bedoelt een positieve cijferreeks met informatie over één van de variabelen voor te stellen als het gewogen zwaartepunt

van de puntenwolk van die variabele. Als ξ een vector is met positieve informatie over de categorieën van A, en

$$\xi_{\cdot} = \sum_{i=1}^I \xi_i$$

dan kunnen de elementen van $\xi_{\cdot}^{-1} \xi$ gebruikt worden als gewichten. De coördinaten van het gewogen zwaartepunt zijn dan:

$$\bar{z}' = \xi_{\cdot}^{-1} \xi' X$$

Dit representatie-model is vooral handig als de cijferreeks een frequentie-verdeling is, dit wil zeggen een kolom uit een kruistabel van A met een categorische variabele C. De gebruikte gewichten zijn dan de conditionele verdeling van de categorieën van A gegeven de gekozen categorie van C. De verdere logica ligt dan in de lijn van het 'principe barycentrique' dat hoger werd uiteengezet. In feite zijn de aldus geplote punten vergelijkbaar met de punten van B. Wel dient rekening gehouden met de wijze waarop A gerepresenteerd werd. In de symmetrische voorstelling wordt $z \Lambda^{-1/2}$ gerepresenteerd. Als de punten van A zo verschoven zijn dat B zwaartepunt is van A, dan wordt Z gerepresenteerd, enzovoort (Lebart, Morineau & Tabard, 1977: 56-57).

4. VERALGEMENINGEN VAN CORRESPONDENTIE-ANALYSE NAAR MEERDERE VARIABLEN

Zoals reeds eerder werd opgemerkt is correspondentie-analyse in voorgaand exposé benaderd vanuit de doelstelling een geometrisch beeld te krijgen van de structuur van een kruistabel. Een eerste voor de hand liggende manier om meerdere variabelen in de analyse te betrekken is dan de eerder besproken technieken toe te passen op tabellen waarin meerdere variabelen verwerkt zijn. Ten eerste kan dit gebeuren door de categorieën van een groep van variabelen die een meerdimensionele kruistabel constitueren te 'reduceren' tot een 'complexe' twee-dimensionele tabel. Zo kan bijvoorbeeld de tabel van A met 5, B met 6 en C met 3 categorieën

teruggebracht worden tot een twee-dimensionele tabel $A*(BC)$ met 5 rijen en 18 kolommen. Een tweede methode om dit te realiseren is over te stappen naar een gelijktijdige analyse van de marginale verdelingen van variabelen uit twee groepen. In die zin kan in het zopas beschreven voorbeeld de tabel $F = \begin{bmatrix} F_{AB} & F_{AC} \end{bmatrix}$ geanalyseerd worden met gewone correspondentie-analyse (Leclerc, 1975, 1976, 1980; Sikkel, 1981; Keller, 1980; Keller & Wansbeek, 1983).

Een tweede type veralgemening, meer in de strikte zin van het woord, gaat uit van de eigenschap dat de eerste component kan worden opgevat als categorie-kwantificatie van de variabelen die maakt dat de pearson correlatie-coëfficiënt tussen de twee variabelen maximaal is. Correspondentie-analyse kan derhalve beschouwd worden als een canonische correlatie-analyse op de indicatormatrices van de twee variabelen (Lebart, Morineau & Tabard, 1977: 106 e.v.; Lancaster, 1957; Williams, 1952; Kendall & Stuart, 1967). Aan de hand van Carroll's veralgemening van canonische correlatie (Carroll, 1968) naar meer sets van variabelen kan ook correspondentie-analyse veralgemeend worden (Lebart, Morineau & Tabard, 1977: 130 e.v.). Vermits Kettenring (1971) heeft aangetoond dat Carroll's veralgemeende canonische correlatie neerkomt op het maximaliseren van de eerste eigenwaarde van de correlatie-matrix van de verkregen lineaire combinaties, komt de toepassing op indicatormatrices neer op een optimale scalering van de categorieën in functie van maximale homogeniteit, met andere woorden op HOMALS (Gifi, 1981, 1982). Gewone correspondentie-analyse kan ook gezien worden als een optimale scalering van één variabele zodat de discriminatie tussen de categorieën maximaal is (Fisher, 1940). De Leeuw (1973) heeft aangetoond dat dit in HOMALS veralgemeend wordt.

Een andere invalshoek die tot hetzelfde resultaat leidt is het toepassen van gewone correspondentie-analyse op de superindicatormatrix. In het geval van twee variabelen geeft dit hetzelfde resultaat, waardoor de eenheden in de gecombineerde plots kunnen bijgetekend worden als zwaartepunt van de categorieën waartoe ze behoren. Dit principe kan ook toegepast worden op een superindicatormatrix van meer dan twee variabelen. Deze veralgemeningen komen neer op volgende analyse. Definieer

$$B = G'G \quad B = \begin{bmatrix} D_1 & F_{12} & F_{13} & \dots & F_{1m} \\ F_{21} & D_2 & F_{23} & \dots & F_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ F_{m1} & F_{m2} & F_{m3} & \dots & D_m \end{bmatrix}$$

$$D = \text{diag}(B)$$

Vervolgens wordt volgend eigenprobleem opgelost:

$$\frac{1}{m} D^{-1/2} B D^{-1/2} - \frac{1}{mn} D^{1/2} 1_J 1_J' D^{1/2} = U \Lambda U'; U'U = I$$

$$X = \sqrt{n/m} D^{-1/2} U \Lambda^{1/2}$$

Er zijn in het algemeen $J-m$ niet-triviale eigenwaarden (J is het totaal aantal categorieën in de m variabelen. Er is een relatie met de bivariate χ^2 -statistiek aantoonbaar:

$$m^2 \sum_s \lambda_s^2 = \sum_{k \neq 1} \sum \frac{\chi_{k1}^2}{n} + J - m$$

Het betreft hier dus een methode die de bivariate marginalen van de m variabelen onderzoekt.

Een derde mogelijke veralgemening vertrekt van de observatie dat correspondentie-analyse in feite een SVD is van de gestandaardiseerde residuen van het onafhankelijkheidsmodel. Als meerdere variabelen in de analyse worden opgenomen dient met een tweetal opmerkingen rekening gehouden. Ten eerste zijn er vele mogelijke (hiërarchisch log-lineaire) modellen die kunnen afgezet worden tegenover de geobserveerde frequenties. Ten tweede brengen deze residuen vrijwel steeds verscheidene effecten tot uitdrukking. Zo bijvoorbeeld betreft een onderzoek van de residuen van het model AC,BC niet alleen de analyse van het effect AB, maar ook van het interactie-effect ABC. In de mate dat dit interactie-effect verwaarloosbaar is stelt dit geen probleem. Wanneer dit effect nochtans niet verwaarloosbaar is, is het interessant dit effect te elimineren. Dit kan gebeuren door de residuen van bijvoorbeeld model AC,BC niet te

berekenen ten opzichte van de geobserveerde frequenties (dus de verwachtingen onder het gesatureerd model), maar ten opzichte van de verwachte frequenties onder het model zonder derde-orde interactie-effecten, namelijk AC,BC,AC. Voor verdere theorie rond loglineaire modellen, wordt verwezen naar Bishop, Fienberg & Holland (1974) en Haberman (1977).

De situatie met meerdere variabelen heeft ook als probleem dat de residuen geen matrix meer vormen. In het geval van drie variabelen vormen deze blok, en in het geval van meer variabelen hoger-dimensionele analogen hiervan. Derhalve dienen analogen van de SVD ontwikkeld te worden die de structuur van deze blokken uit elkaar haalt. Voor drie variabelen kan verwezen worden naar Tucker (1963, 1964, 1966) die een veralgemening van de SVD heeft uitgewerkt. Deze heeft echter niet meer de geneste kleinste kwadraten eigenschappen die een SVD zo aantrekkelijk maakt (Eckart & Young, 1936). Een methode om voor dit probleem een oplossing te vinden wordt gegeven door Kroonenberg, namelijk TUCKALS3 (Kroonenberg & De Leeuw, 1980; Kroonenberg, 1983). Lastovicka (1981) heeft een veralgemening van de SVD voor 4 ingangen uitgewerkt. Kapteyn, Neudecker & Wansbeek (1981) geven een veralgemening voor gegevens met n ingangen, evenals een ALS-algoritme voor een kleinste kwadraten-benadering voor gefixeerde dimensionaliteiten. Kroonenberg (1983) heeft deze invalshoek gebruikt voor de veralgemening van correspondentie-analyse naar meerdere variabelen.

Bibliografie

BENZECRI, J.-P., L'analyse des données. I. La Taxonomie; II. L'analyse des correspondances. Parijs, Dunod, 1973.

BISHOP, Y., FIENBERG, S., HOLLAND, P., Discrete Multivariate Analysis. Cambridge, M.I.T. press, 1977.

BOUROCHE, J.M., SAPORTA, G., L'analyse des données. Paris, P.U.F. (que sais-je ? 1854), 1983.

CAILLEZ, F., PAGES, J.P., Introduction à l'analyse des données. Paris, S.M.A.S.H., 1976.

CARROLL, J.D., A generalization of canonical correlation analysis to three or more sets of variables. Proceedings of 76th Convention of American Psychological Association, 1968.

CARROLL, J.D., Individual differences and multidimensional scaling. In: SHEPARD, R.N., ROMNEY, A.K., NERLOVE, S.B. (eds.), Multidimensional scaling. Theory and applications in the behavioral sciences. Vol.I. New York, Seminar Press, 1972, 105-155.

DOBBELAERE, K., BILLIET, J., Godsdienst in Vlaanderen. Leuven, Davidsfonds, 1975.

ECKART, C., YOUNG, G., The approximation of one matrix by another of lower rank, in: Psychometrika, 1(1936)3, 211-218.

ESCOUFIER, Y., L'Analyse des tableaux de contingence simple et multiples, in: Metron, (1982)1-2, 53-77.

EVERITT, B., Cluster analysis. New York, Halsted Press, 1981.

FISHER, R.A., The precision of discriminant functions. In: Annals of eugenetics, 10(1940), 422-429.

GEER, J.P. VAN DE, Introduction to multivariate analysis for the social sciences. San Francisco, Freeman, 1971.

GIFI, A., Niet-lineaire multivariate analyse. Leiden, 1980.

GIFI, A., Non-linear multivariate analysis. Leiden, 1981.

GUTTMAN, L., The principal components of scale analysis. In: STOUFFER, S. (ed.), Measurement and prediction. New York, John Wiley (1966) 1950, 312-361.

GUTTMAN, L., The principal components of scalable attitudes. In: LAZARSFELD, B.F. (ed.), Mathematical thinking in the social sciences. Glencoe, Free Press, 1954.

HABERMAN, S.Y., Analysis of qualitative Data. Vol. I, II. New York, Academic Press, 1978.

HEISER, W.J., Unfolding analysis of proximity data. Leiden, dissertatie, 1981.

HEISER, W.J., MEULMAN, J., Analyzing rectangular tables by joint and constrained multidimensional scaling. In: Journal of Econometrics, 22(1983), 139-167.

ISRAELS, A.Z., Clusteranalyse met behulp van de χ^2 -afstand. Interne CBS-nota, nr. 12806-79-M1, 1979.

KAPTEYN, A., NEUDECKER, H., WANSBEEK, T., n-mode component analysis, Interne CBS-nota, nr. 10875-82-M1, 1982.

KELLER, W.J., Het 'error-in-variables'-model (EVM) en andere multivariate technieken, Interne CBS-nota, nr. 12008-80-M1, 1980.

KELLER, W.J., WANSBEEK, T., Multivariate Methods for quantitative and qualitative data. In Journal of econometrics. 22(1983), 91-111.

KENDALL, M.G., STUART, A., The advanced theory of statistics. Vol. II, London, Griffin, 1961.

KETTENRING, J.R., Canonical Analysis of several sets of variables. In: Biometrika, 58(1971), 433-460.

KROONENBERG, P.M., Three-mode principal component analysis. Leiden, DSWO-press, 1983.

KROONENBERG, P.M., LEEUW DE, J., Principal component analysis of three-mode data by means of alternating least squares algorithms. In: Psychometrika, 45(1980)1, 69-97.

LASTOVICKA, J.L., The extension of component analysis to four-mode matrices. In: Psychometrika, 46(1981)1, 47-57.

LANCASTER, H.O., Some properties of the bivariate normal distribution considered in the form of a contingency table. In: Biometrika, 44(1957), 289-292.

LEBARD, L., MORINEAU, A., FENELON, J.-P., Traitement des données statistiques. Paris, Dunod, 1979.

LEBARD, L., MORINEAU, A., TABARD, N., Techniques de la description statistique. Paris, Dunod, 1977.

LECLERC, A., L'analyse des correspondances sur juxtaposition de tableaux de contingence, in: Revue de statistique appliquée, 23(1975)3, 5-16.

LECLERC, A., Une étude de la relation entre une variable qualitative et un groupe de variables qualitatives. In: International statistical revue. 44(1976)2, 241-248.

LECLERC, A., Quelques propriétés optimales en analyse des données en termes de corrélation entre variables. In: Mathématique et sciences humaines, 18(1980), 51-67.

LEEuw, J. DE, Canonical analysis of categorical data. Leiden, dissertatie, 1973.

SCHRIEVER, B.F., Scaling of order dependent categorical variables with correspondance analysis, in: International statistical Review, 51(1983), 225-238.

SIKKEL, P., An application of correspondance-analysis to leisure time activities. Den Haag, staatsuitgeverij, Statistical studies, 75, 1980.

SIKKEL, P., The relationship between canonical correlations and correspondance analysis. Interne CBS-nota, nr. 8696-81-M1, 1981.

TUCKER, L.R., Implications of factor analysis of three-way-matrices for measurement of change. In HARRIS, C.W., Problems in measuring change. Madison, Univ. of Wisconsin Press, 1963, 122-137.

TUCKER, L.R., The extension of factor analysis to three-dimensional matrices. In FREDRIKSEN, H., GULLIKSEN, H., Contributions to mathematical psychology. New York, Holt, Rinehart & Winston, 1964, 110-127.

TUCKER, L.R., Some mathematical notes on three-mode factor-analysis. In: Psychometrika, 31(1966), 279-311.

WARD, J.H., Hierarchical grouping to optimize an objective function. In: Journal of the American statistical association, 68(1963), 236-244.

WILLIAMS, E.J., The use of scores for the analysis of association in contingency tables. In: Biometrika, 39(1962), 274-289.

Ontvangen: 21-8-1984