1. Inleiding.

Hoofdassenanalyse is langzamerhand een al belegen data-analytische techniek en gemeengoed in de sociaal-wetenschappelijke literatuur. Er ligt aan hoofdassenanalyse evenwel geen theorie over de gegevens ten grondslag, in tegenstelling tot bijvoorbeeld faktoranalyse waar een - zij het zeer globale - theorie over de gegevens bestaat. Bij hoofdassenanalyse zoekt men slechts naar die lineaire kombinaties van de oorspronkelijke variabelen die, succesievelijk en onafhankelijk van elkaar, een zo groot mogelijk gedeelte van de in de gegevens aanwezige totale variantie voor hun rekening nemen.
Een andere, voor de sociale wetenschappen misschien iets gebruikelijkere formulering van het hoofdassenprobleem is de volgende.
Stel wij hebben een nxm matrix Z met gegevens, bijvoorbeeld de scores van n proefpersonen op m variabelen. We zoeken nu een mxp orthonormale matrix F, de zgn. "komponent ladingen" en een nxp koefficiëntenmatrix A, de zgn. "komponent scores" valt, zodanig dat $Z = AF'$ en dat $A'A$ een diagonale matrix is waarvan de elementen op de hoofddiagonaal van groot naar klein gerangschikt zijn. De eerste kolom van F zal dan korresponderen met die lineaire kombinatie van de kolommen van Z die het grootste gedeelte van de totale variantie vertegenwoordigt. Veelal is men niet geinteresseerd in alle hoofdassen (i.e. alle kolommen van F), maar alleen in die r hoofdassen, die het grootste deel van de variantie van Z voor hun rekening nemen. Men kan ook slechts de eerste drie of vier hoofdassen willen bekijken om een inzicht te krijgen van de belangrijkste strukturen die in de gegevens te vinden zijn. De rang van F (=r) is in dit soort gevallen dan veel kleiner dan de rang van Z. Een exakte faktorisering van Z in AF' is in een dergelijke situatie doorgaans niet meer mogelijk en men zal genoegen moeten nemen met een beste benaderende faktorisering, hopelijk zonder verlies van enige relevante informatie.

Van recentere datum is de poging om technieken te ontwikkelen die ingewikkelder gegevensbestanden aankunnen. Bij de hierboven geschetste hoofdassenanalyse beperkt men zich tot de analyse van gegevens die tweevoudig geklassificeerd zijn, bijvoorbeeld door middel van (proef)personen en door variabelen. Verschillende onderzoekers verzamelen evenwel gegevens die drie- of meervoudig geklassificeerd kunnen worden.

MDN, SWS/Vereniging voor Statistiek, 1978, 3(3), 30-53

TUCKALS 2: Een hoofdassenanalyse voor drieweggegevens.

Pieter Kroonenberg
Pedagogisch Instituut

Jan de Leeuw
Afdeling Datatheorie
Fakulteit der Sociale
Wetenschappen

Rijksuniversiteit Leiden.

Summary

In this paper we present a principal component analysis for three-mode data. The model used - christened the Tucker 2 model - is an asymmetric variant of the general Tucker three-mode model, i.e. the principal components of just two of the three modes are present in the model. The Tucker 2 model can be seen as a direct generalization of the INDSCAL-model. The alternating least squares principle is used to estimate the parameters of the model, and an algorithm is outlined for computing the estimates. Two examples from the Dutch political scene of 1968 are used to illustrate the functioning of the programme TUCKALS 2 written around the algorithm.

1. Inleiding.

Hoofdassenanalyse is langzamerhand een al belegen data-analytische techniek en gemeengoed in de sociaal-wetenschappelijke literatuur. Er ligt aan hoofdassenanalyse evenwel geen theorie over de gegevens ten grondslag, in tegenstelling tot bijvoorbeeld faktoranalyse waar een - zij het zeer globale - theorie over de gegevens bestaat. Bij hoofdassenanalyse zoekt men slechts naar die lineaire kombinaties van de oorspronkelijke variabelen die, succesievelijk en onafhankelijk van elkaar, een zo groot mogelijk gedeelte van de in de gegevens aanwezige totale variantie voor hun rekening nemen.
Een andere, voor de sociale wetenschappen misschien iets gebruikelijkere formulering van het hoofdassenprobleem is de volgende.
Stel wij hebben een nxm matrix Z met gegevens, bijvoorbeeld de scores van n proefpersonen op m variabelen. We zoeken nu een mxp orthonormale matrix F, de zgn. "komponent ladingen" en een nxp koefficiëntenmatrix A, de zgn. "komponent scores" valt, zodanig dat $Z = AF'$ en dat $A'A$ een diagonale matrix is waarvan de elementen op de hoofddiagonaal van groot naar klein gerangschikt zijn. De eerste kolom van F zal dan korresponderen met die lineaire kombinatie van de kolommen van Z die het grootste gedeelte van de totale variantie vertegenwoordigt. Veelal is men niet geinteresseerd in alle hoofdassen (i.e. alle kolommen van F), maar alleen in die r hoofdassen, die het grootste deel van de variantie van Z voor hun rekening nemen. Men kan ook slechts de eerste drie of vier hoofdassen willen bekijken om een inzicht te krijgen van de belangrijkste strukturen die in de gegevens te vinden zijn. De rang van F (=r) is in dit soort gevallen dan veel kleiner dan de rang van Z. Een exakte faktorisering van Z in AF' is in een dergelijke situatie doorgaans niet meer mogelijk en men zal genoegen moeten nemen met een beste benaderende faktorisering, hopelijk zonder verlies van enige relevante informatie.

Van recentere datum is de poging om technieken te ontwikkelen die ingewikkelder gegevensbestanden aankunnen. Bij de hierboven geschetste hoofdassenanalyse beperkt men zich tot de analyse van gegevens die tweevoudig geklassificeerd zijn, bijvoorbeeld door middel van (proef)personen en door variabelen. Verschillende onderzoekers verzamelen evenwel gegevens die drie- of meervoudig geklassificeerd kunnen worden.

Hieronder volgen enkele voorbeelden van zulke gegevens:

- Een klassiek voorbeeld van drievoudig geklassificeerde gegevens is te vinden bij het onderzoek van Osgood, Suci en Tannenbaum (1957), waarbij zij hun semantische differentiaal ontwikkelden.
  In hun geval werden de oordelen verkregen van een aantal individuen over de betekenis van bepaalde koncepten, met behulp van bipolaire schalen.
- Endler, Hunt en Rosenstein (1962) verzamelden de gegevens voor hun onderzoek door een "Stimulus-Response inventarisatie van anticipatieangst" af te nemen bij 169 studenten. De inventarisatie bestond hieruit dat de ondervraagde moest schatten hoe intens hij zou reageren in elf verschillende (penibele) situaties wat betreft elk van de veertien mogelijke antwoordkategoriën. De situaties waren bijvoorbeeld "Je gaat voor het eerst met een meisje uit". "Je gaat naar een sollicitatiegesprek voor een belangrijke baan", etc.; de antwoorden waren bijvoorbeeld: "hart slaat sneller", "zweten", "verheug me op de uitdaging", etc.
- Jones en Young (1972) verzamelden gegevens over de sociale struktuur van een kleine, gesloten en natuurlijk gevormde groep (staf en studenten van een psychologisch instituut).
  Staf, studenten en ander personeel van het instituut gaven hun oordeel over de gelijkenis tussen de leden van de wetenschappelijke staf en enkele doktoraal studenten, waarbij de gelijkenis van elk mogelijk paar werd aangegeven op een zevenpuntschaal. Dit soort gegevens worden typisch geanalyseerd met meerdimensionale schaalmodellen, of modellen voor individuele verschillen (e.g. INDSCAL, Carrol & Chang, 1970).
- Van de Geer (1974) geeft een voorbeeld van tijdreeks gegevens, waarbij relatief veel variabelen en weinig meetpunten in tijd voorkomen. In het zgn. ziekenhuisprojekt zijn gegevens beschikbaar over 188 ziekenhuizen t.a.v. 27 variabelen, gemeten op 11 opeenvolgende jaren.

In elk van de bovenstaande gevallen is er sprake van drieweg gegevens, waarbij een weg gedefinieerd is als een indexverzameling waardoor de gegevens geklassificeerd kunnen worden. In bovenstaande voorbeelden zijn de driewegen respektievelijk

- individuen, konsepten, schalen
- studenten, situaties, antwoordkategoriën

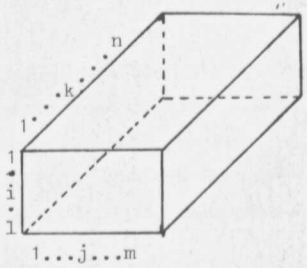
- beoordelaars, (stimulus)personen, (stimulus)personen
- ziekenhuizen, variabelen, tijdstippen.

Hogere orde klassificaties kunnen natuurlijk ook voorkomen. Men krijgt bijvoorbeeld vierweggegevens als men het Osgood e.a. experiment zou herhalen bij verschillende kulturen. In dit artikel zullen we ons echter alleen met drieweggegevens en de hoofdassenanalyse ervan bezighouden. Daarbij zullen we dus ook niet verder ingaan op andere analysetechnieken zoals de strukturele kovariantieanalyse van Jöreskog (1971), Werts, Jöreskog & Linn (1972);het individuele verschillen model van Carrol & Chang (1970) - INDSCAL, en de techniek van Campbell & Fiske (1959).

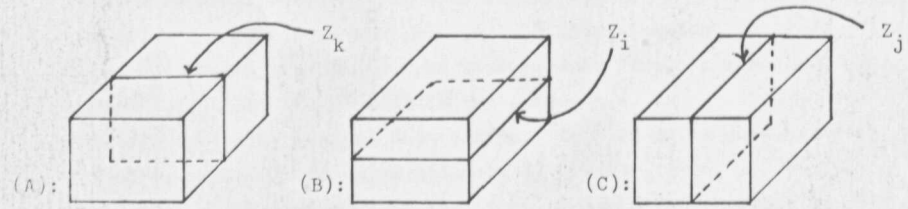
De basis voor de hoofdassenanalyse van drieweggegevens werd gelegd door Tucker (1963, 1964, 1966, 1972). Tucker echter koncentreerde zich voornamelijk op het faktoranalyse model, terwijl wij hier ons uitsluitend bezig zullen houden met de hoofdassenanalyse model en dan nog met een vereenvoudigde versie daarvan.

2. Notatie en terminologie

We zullen de klasse van alle reële n bij m matrices aanduiden met R^{nxm}. Verder duiden we de klasse van alle reële n bij m matrices, waarvan de kolommen orthonormaal zijn, aan met K^{nxm}, waarbij de afspraak geldt dat $n \geqslant m$.
Verder definiëren we een driewegmatrix Z als de kollektie elementen $\{z_{ijk} \mid i = 1,\ldots,1;\ j = 1,\ldots, m;\ k = 1,\ldots,n\}$. In de wiskundige literatuur noemt men Z tensor. Deze elementen kunnen in een blok geplaats worden met de index i langs de vertikale as, de index j langs de horizontale as en de index h langs de 'diepte' as.
Een drieweg matrix kan ook opgevat worden als een kollektie 'normale' (=tweeweg) matrices. Dit kan op drie manieren, waarvan de nu eerst genoemde manier in het vervolg verreweg het meeste voor zal komen.



Figuur 2-1 Driewegmatrix Z



Figuur 2-2
Drie verschillende manieren om de driewegmatrix Z te zien als een kollektie tweewegmatrices.

a. de verzameling voorvlakken: $Z = \{ Z_k \}_{k=1,\ldots,n}$ zie figuur 2-2 A

b. de verzameling bovenvlakken: $Z = \{ Z_i \}_{i=1,\ldots,l}$ zie figuur 2-2 B

c. de verzameling zijvlakken : $Z = \{ Z_j \}_{j=1,\ldots,m}$ zie figuur 2-2 C

Tenslotte duidt I_n de n bij n eenheidsmatrix aan.

3. Het TUCKER model

Het algemene drieweg hoofdanalyse model van Tucker is als volgt gedefinieerd:

Zij $Z = \{ z_{ijk} \}$ een driewegmatrix, dan kan deze Z gefaktoriseerd worden als:

$$z_{ijk} = \sum_{\alpha=1}^{s} \sum_{\beta=1}^{t} \sum_{\gamma=1}^{u} g_{i\alpha} h_{j\beta} e_{k\gamma} c_{\alpha\beta\gamma} \qquad \begin{matrix} i=1,\ldots,l;\ j=1,\ldots,m; \\ k=1,\ldots,n \end{matrix} \qquad (3.1.)$$

met s,t en u het aantal komponenten van resp. de eerste, tweede en derde weg.

De koefficienten $g_{i\alpha}, h_{j\beta}, e_{k\gamma}$ zijn elementen van de matrices G, H, E.

De koefficienten beschrijven de scores van de oorspronkelijke variabelen op de komponenten. De koefficienten $c_{\alpha\beta\gamma}$ zijn de elementen van de driewegmatrix C, de zgn. kernmatrix.
In de originele matrix Z representeerde ieder element van de matrix een specifieke kombinatie van kategorien van de oorspronkelijke variabelen. Op dezelfde manier stelt ieder element in de kernmatrix een unieke kombinatie van kategorien van de komponenten voor. Men kan zich voorstellen dat de kern matrix de basis relaties beschrijft die bestaan tussen de verschillende verzamelingen variabelen.

4. Hoofdassenanalyse volgens Tucker.

De oplossing die Tucker voorgesteld heeft is in wezen erg simpel en recht toe recht aan. Kort samengevat komt deze op het volgende neer:

Voor elk van de wegen apart worden de hoofdassen berekend. Hiertoe vormt men eerst de gesommeerde kruisprodukten voor elk van de wegen:

weg 1 : $\sum_{k=1}^{n} Z_k Z_k'$, $Z_k \in R^{lxm}$; weg 2 : $\sum_{i=1}^{l} Z_i Z_i'$, $Z_i \in R^{mxn}$;

weg 3: $\sum_{j=1}^{m} Z_j Z_j'$, $Z_j \in R^{nxl}$

m.b.v. elk van deze gesommeerde kruisproduktenmatrices worden dan de hoofdassen berekend voor elk van de wegen. Deze procedure is identiek aan de procedure waarbij men eerst een nieuwe matrix, $\tilde{Z}$, konstrueert op de in figuur 4-1 aangegeven wijze en dan voor $\tilde{Z}\tilde{Z}'$ de hoofdassen berekent.

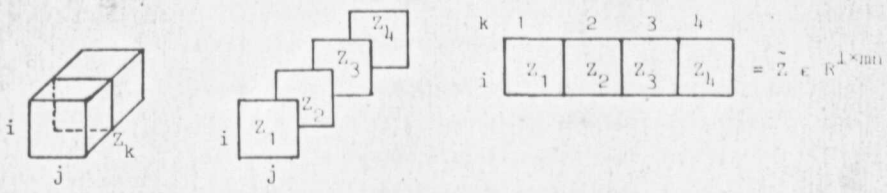
De hoofdassen worden berekend met behulp van de eigenwaarden - eigenvektoren dekompositie net zoals bij een gewone hoofdassenanalyse. (Voor verdere details zie met name: Tucker, 1966)

De door Tucker voorgestelde procedure kent twee nadelen en/of komplikaties:

a. Aangezien zeer vaak een van de wegen uit individuen bestaat en dit er nogal veel kunnen zijn in een enigszins goed opgezet onderzoek, kan het oplossen van het eigen probleem wel eens een dure zaak worden. Tucker zag dit probleem in en bedacht dan ook een iets gewijzigde procedure die echter een aantal nadelen heeft, nauw samenhangend met het volgende.

b. Zolang alle eigenvektoren gewenst zijn - wat in de praktijk nooit het geval is - valt er niet veel aan te merken op de gevolgde procedure. Wanneer men echter slechts een,twee,drie of vier eigenvectoren per weg wil hebben, dan komt men snel in moeilijkheden. Het weglaten van kleine eigenwaarden en bijbehorende faktoren geschiedt onafhankelijk van elkaar bij elk van de wegen.
De wegen zelf zijn evenwel niet onafhankelijk en er is geen garantie dat de gevonden oplossing de optimale is (Tucker, 1966, p.296).
Dit geldt in een nog sterkere mate voor de hierboven genoemde gewijzigde methode.
De gevonden hoofdassen zijn dan ook geen kleinste kwadraten schatters voor de model parameters.

Tucker merkt ten aanzien van deze problematiek op dat een kleinste kwadraten oplossing van het schattingsprobleem een gekompliceerde opeenvolging van benaderingen noodzakelijk maakt.



Figuur 4-1 Konstruktie van $\tilde{Z}$

Juist de kleinste kwadraten oplossing van het driewegprobleem is het doel van ons projekt. Op dit moment hebben we een (nog niet exporteerbaar) computer programma ontwikkeld voor een gedeeltelijke oplossing van het driewegprobleem.
Het gedeeltelijke zit hem in het feit dat slechts voor twee van de drie wegen de hoofdassen worden berekend. Afgezien van het feit dat de konstruktie van dit komputerprogramma een aardige vingeroefening is voor de algehele oplossing van het driewegmodel, heeft het vereenvoudigde model ook toepassingen als kontrole op het model bij een aantal meerdimensionele schaalmodellen, zoals PARAFAC, INDSCAL en IDIOSCAL. Een diskussie over de relatie tussen het algemene Tucker driewegmodel (m.n. de meerdimensionele schaal versie ervan) en de andere genoemde modellen is te vinden in Carroll en Wish (1974) en Takane, Young en De Leeuw (1977).

5. Het Tucker 2 model.

In plaats van het algemene Tucker model direkt aan te pakken hebben we eerst de vereenvoudiging doorgevoerd, dat E gelijk gekozen wordt aan de eenheidsmatrix, m.a.w. we hoeven slechts voor twee van driewegen de hoofdassen te berekenen.*
Zonder verlies aan algemeenheid zullen wij hiervoor altijd de eerste en de tweede weg kiezen.

*Onafhankelijk van elkaar hebben eerder Israelsson(1969), Jennrich (1972) en Carroll en Chang(1970) dit model voorgesteld.

Het vereenvoudigde - door ons Tucker 2 gedoopte - model is dus als volgt geformuleerd:

Zij $Z = \{z_{ijk}\}$ een drieweg matrix, dan kan deze Z gefaktoriseerd worden als:

$$z_{ijk} = \sum_{\alpha=1}^{s} \sum_{\beta=1}^{t} g_{i\alpha} h_{j\beta} c_{\alpha\beta k} \quad i=1,\ldots,l \; ; \; j=1,\ldots,m; \; k=1,\ldots,n \tag{5.1.}$$

In matrix notatie wordt dit:

$Z = \{Z_k\}_{k=1,\ldots,n}$, $Z_k \in R^{lxm}$ en $Z_k = GC_kH'$, waarbij

$g_{i\alpha}, h_{j\beta}$ en $c_{\alpha\beta k}$ de elementen zijn van respectievelijk G, H, C_k.

De C_k kunnen verzameld worden in de kernmatrix $C = \{C_k\}_{k=1,\ldots,n}$.

6. Kleinste kwadraten hoofdassenanalyse voor het Tucker 2 model.

Met behulp van een kleinste kwadraten procedures zoeken we dus de grootste s en t hoofdassen voor het Tucker 2 model, waarbij aangenomen mag worden dat s en t klein zijn vergeleken met de rijen-, respektievelijk kolommen rang van de Z_k 's. Meestal zullen s en t 2, 3 of 4 zijn. De nette formulering van het kleinste kwadraten probleem is als volgt:

Zij $Z = \{Z_k\}_{k=1,\ldots,n}$ met $Z_k \in R^{lxm}$ en laat s en t zodanige gehele getallen zijn dat $1 \leq s \leq l$ en $1 \leq t \leq m$.
De beste benadering voor het Tucker 2 model is dan de oplossing van de minimalisering van:

$$\sigma(G,H,C) = \sum_{k=1}^{n} \mathrm{Tr}(Z_k - GC_kH)'(Z_k - GC_kH') \tag{6.1.}$$

over alle $G \in K^{lxs}$, $H \in K^{mxt}$ en $C = \{C_k\}_{k=1,\ldots,n}$

met $C_k \in R^{sxt}$.

In Kroonenberg en De Leeuw (1977) hebben wij het volgende aangaande dit minimalisatieprobleem bewezen:

a. Er bestaat altijd een oplossing van het minimalisatieprobleem, m.a.w. er bestaat altijd een beste benadering voor $Z = \{Z_k\}_{k=1,\ldots,n}$ van de vorm $\hat{Z} = \{\hat{Z}_k\}_{k=1,\ldots,n}$ met $\hat{Z}_k = GC_kH'$.

De grote lijnen van het bewijs zijn als volgt:

i. voor vaste G en H wordt (6.1) geminimaliseerd door voor C_k $C_k = G'Z_kH$ te kiezen en de minimalisatie over C is niet afhankelijk van de minimalisatie over G en H.

ii. dus kan (6.1) herschreven worden als:

$$\sigma(G,H) = \sum_{k=1}^{n} \mathrm{Tr}(Z_k - GG'Z_kHH')'(Z_k - GG'Z_kHH') \qquad (6.2.)$$

en verder als:

$$\sigma(G,H) = \sum_{k=1}^{n} \mathrm{Tr}\, Z_k'Z_k - p(G,H), \text{ met}$$

$$p(G,H) = \sum_{k=1}^{n} \mathrm{Tr}G'Z_kHH'Z_kG \qquad (6.3.)$$

iii. dus is het minimalisatieprobleem van σ equivalent met het maximalisatieprobleem van p.

iv. p is gedefinieerd op de kompakte deelverzameling $S = \{(G,H) \mid G \in K^{lxs}$ en $H \in K^{mxt}$ uit de $R^{lsmt}\}$

v. p is een kontinue begrensde funktie op S en dus heeft p een supremum in S en neemt dit aan.

b. Zij $P(G) = \sum_{k=1}^{n} Z_k' GG'Z_k$ en $Q(H) = \sum_{k=1}^{n} Z_kHH'Z_k'$. (6.4.)

p neemt zijn maximum aan voor $(\hat{G},\hat{H})$ dan en slechts dan als $\hat{G}$ ($\hat{H}$) een orthonormale rotatie is van de eigenvektorenmatrix behorende bij de s (t) grootste eigenwaarden van $Q(\hat{H})$ ($P(\hat{G})$).

c. Onder bepaalde voorwaarden kan er een exakte oplossing gevonden worden, maar omdat in dat geval de oplossing van volledige rang is, is deze in vrijwel alle praktische gevallen niet erg interessant.

7. De alternerend kleinste kwadraten benadering.

Het is duidelijk dat we een zodanig algoritme voor de maximalisatie van p zouden willen konstrueren dat het konvergeert naar het globale maximum van p. Ongelukkig genoeg is p een kruisprodukt term van een vierdegraads multivariaat popynoom. Voor dit soort niet-lineaire problemen is het in het algemeen niet mogelijk om algoritmen te konstrueren, waarvan te bewijzen valt dat ze naar een globaal maximum konvergeren.
Ook hier is dit het geval. Wanneer er dan ook van konvergentie sprake is dan bedoelen we dat het algoritme naar een van de stationaire punten van p konvergeert, dat geen minimum is.
De methode die wij gebruikt hebben om de beste benadering te vinden van het Tucker 2 model, maakt gebruik van een zogenaamde alternerende kleinste kwadraten (ALS - Alternating Least Squares) methode. De meest karakteristieke eigenschap van ALS methoden is dat optimaliseringsproblemen waarbij meer dan één verzameling parameters bij betrokken is, opgelost worden door iedere verzameling op zijn beurt te schatten met konstant houden van de andere verzameling(en).
Nadat elke verzameling eenmaal geschat is wordt de procedure steeds weer herhaald tot dat een stationair punt van de te maximaliseren funktie bereikt is. Deze techniek is natuurlijk al veel langer bekend, n.l. bij het schatten van parameters in regressievergelijkingen, waarin de storingstermen met zichzelf gekorreleerd zijn (cf.Cochrane en Orcutt, 1949). Verdere details en toepassingen in multivariate analyse van de ALS-benadering kunnen, bijvoorbeeld, gevonden worden in Young, De Leeuw en Takane (1977) en een voorbeeld werd in een vorig nummer van MDN gegeven door De Leeuw en Van Rijckevorsel (1977).
Dat er in het huidige probleem inderdaad sprake van een kleinste kwadraten probleem volgt uit de formuleringen (6.2) en (6.3) van het minimalisatie probleem.
Het is duidelijk dat de verzamelingen parameters hier G en H zijn en dat de minimalisatie van σ, en dus de maximalisatie van p, over G met H vast het ene kleinste kwadratenprobleem is en dat de minimalisatie van σ, en dus de maximalisatie van p, over H met G vast het andere is.

Hoe het maximalisatieprobleem van p opgelost moet worden is nu ook duidelijk. Kies eerst een willekeurige (of een slimme) H_0, maximaliseer p over G met deze vaste H_0, maximaliseer vervolgens met de gevonden G vast over H, etc. Uit het bewijs van punt 6.4. volgt dat de oplossingen van de maximalisaties niets anders zijn dan de eigenvektoren matrices P en Q. Per stap zoeken we natuurlijk die eigenvektoren die behoren bij de grootste s en t eigenwaarden van Q en P. Omdat we maar een paar eigenvektoren nodig hebben, willen we graag gebruik maken van een techniek die efficient is in het vinden van enkele van een mogelijk groot aantal eigenvektoren.
In onze situatie hebben we gebruik gemaakt van de simultane iteratiemethode van Bauer-Rutishauser (Rutishauser, 1969).
De maximalisatie van p bestaat dus uit een in principe oneindig iteratief proces, waarin bij iedere stap twee eigenwaarden-eigenvektoren problemen moeten worden opgelost. Elk van deze eigenwaarden - eigenvektoren problemen wordt aangepakt met een in principe oneindige iteratieve methode. Deze geneste iteratieprocedures dreigen, als we ze zonder meer zouden toepassen, zeer lange rekentijden te vergen. Om dit probleem te omzeilen voeren we slechts één enkele stap uit van de berekening van de eigenwaarden-eigenvektoren problemen, in plaats van de gehele iteratie. Een soortgelijke aanpak is gebruikt door De Leeuw c.s. bij hun toepassingen van de ALS techniek. Hun ervaringen hiermee waren dat het uitvoeren van de komplete iteratie om de eigenvektoren te vinden uitsluitend de algehele efficientie verlaagde, maar geen effekt had op het uiteindelijke konvergentiepunt (cf. Young, De Leeuw en Takane, 1977).

8. Het TUCKALS 2 algoritme

Veronderstel weer dat $Z = \{Z_k\}_{k=1,\ldots,n}$ een gegeven datamatrix is, en laten $G \in K^{l \times s}$ en $H \in K^{m \times t}$ iteratiematrices zijn. Als we G en H zoals ze zijn na i iteratiestappen van het algoritme aanduiden met G_i en H_i, dan wordt een iteratie stap van het TUCKALS 2 algoritme gedefinieerd als:

G substap

$$Q_i = \sum_{k=1}^{n} Z_k H_i H_i' Z_k' \qquad (8.1)$$

$$G_{i+1} = Q_i G_i (G_i' Q_i^2 G_i)^{-\frac{1}{2}} \qquad (8.2)$$

$V W^{-\frac{1}{2}} V'$

H substap

$$P_i = \sum_{k=1}^{n} Z_k' G_{i+1} G_{i+1}' Z_k \qquad (8.3)$$

$$H_{i+1} = P_i H_i (H_i' P_i^2 H_i)^{-\frac{1}{2}} \qquad (8.4)$$

De regels (8.2) en (8.4) zijn in feite verkorte schrijfwijzen voor één enkele stap van de Bauer-Rutishauser methode.(zie Kroonenberg en De Leeuw, 1977, p. 4-3). De inverse wortel in (8.2) en (8.4) wordt berekend via een Jacobi eigenroutine. Zeer kleine eigenwaarden en singulariteiten vormen een potentieel gevaar voor het algoritme. Er zijn dan ook in het programma een aantal kontroles ingebouwd om ze te signaleren en maatregelen om ze in bepaalde gevallen op te vangen.

In Kroonenberg en De Leeuw(1977) hebben wij ten aanzien van het TUCKALS 2 algoritme bewezen dat:

a. alle limietpunten van de rij $\{(G_i, H_i)\}$ $i=0,1,2,\ldots$ stationaire punten zijn van p;

b. bij iedere stap en zelfs iedere substap van het algoritme de waarde van p toeneemt en dat uitsluitend in een limietpunt $p(G_i, H_i) = p(G_{i+1}, H_{i+1})$;

c. de rij $\{(G_i, H_i)\}$ konvergente deelrijen heeft, hetgeen impliceert dat de rij of konvergeert of een kontinuum van limietpunten heeft.

d. $\| (G_i, H_i) - (G_{i+1}, H_{i+1}) \| \to 0$ indien tijdens het iteratieproces in de stappen (8.2) en (8.4) geen singulariteiten zijn opgetreden.

9. Voorbeelden

Voor twee voorbeelden ter illustratie van het algoritme gebruiken we enige verouderde politieke gegevens uit 1968. Er wordt slechts een vrij globale analyse gegeven en niet alle mogelijkheden van

het programma zullen worden uitgebuit. Waarschijnlijk zullen we op een later tijdstip pogen een exemplarische analyse van een recentere data set te geven.

Kenmerken, partijen en psychologen (De Leeuw data).

In 1968 legde De Leeuw (1973) een vragenlijst voor aan elf leden van de staf en de doktoraal studenten M&T van het Psychologisch Instituut van de Leidse Universiteit om na te gaan welke kenmerken geassocieerd zouden zijn met welke politieke partijen. Alle mogelijke kombinaties van twaalf partijen en zeventien kenmerken werden voorgelegd aan de proefpersonen, die moesten aangeven of een kenmerk al dan niet bij een partij behoorde. De gegevens bestonden dus uit elf 12x17 matrices met enen en nullen. Deze gegevens werden dubbel gecentreerd en vervolgens geanalyseerd met het TUCKALS 2 programma. De hoofdassen van de partijen-ruimte en die van de kenmerkenruimte zijn aangegeven in tabel 9.1 en zijn afgebeeld in de figuren 9.2 en 9.3.

TABEL 9.1. Hoofdassen van de kenmerken- en van de partijenruimte

(De Leeuw data)

partij	komponenten($x10^{-2}$)			kenmerk	komponenten($x10^{-2}$)		
	1	2	3		1	2	3
D'66	50	-17	19	homogeen	40	9	- 1
PPR	40	10	2	duidelijk	34	2	-18
PvdA	35	- 5	-33	konsistent	32	-13	-27
PSP	15	37	-30	negatief	30	4	42
CPN	- 2	54	3	dogmatisch	23	-46	-17
ARP	- 4	-33	-36	links	9	19	-29
VVD	- 5	4	19	konservatief	8	-41	41
KVP	- 7	-48	26	up-to-date	-02	41	1
BP	-10	8	69	progressief	-06	23	-18
CHU	-28	-38	-21	opportunistisch	-08	16	59
GPV	-41	14	- 7	belangrijk	-08	27	08
SGP	-42	14	-11	sympathiek	-16	17	-14
				konstruktief	-22	1	-08
				intelligent	-22	11	-01
				verantwoordelijk	-25	-32	-11
				tolerant	-35	-28	4
				demokratisch	-36	-10	-11
gewicht	30	15	6		32	13	6

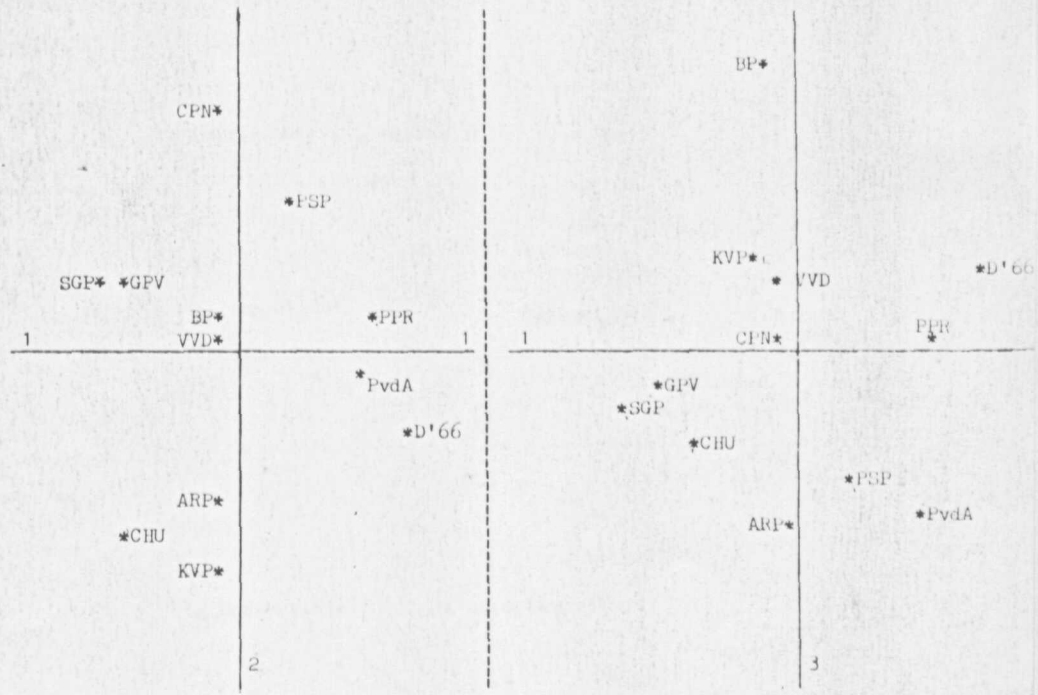


Fig. 9.2 Politieke Partijen ruimte (De Leeuw data)

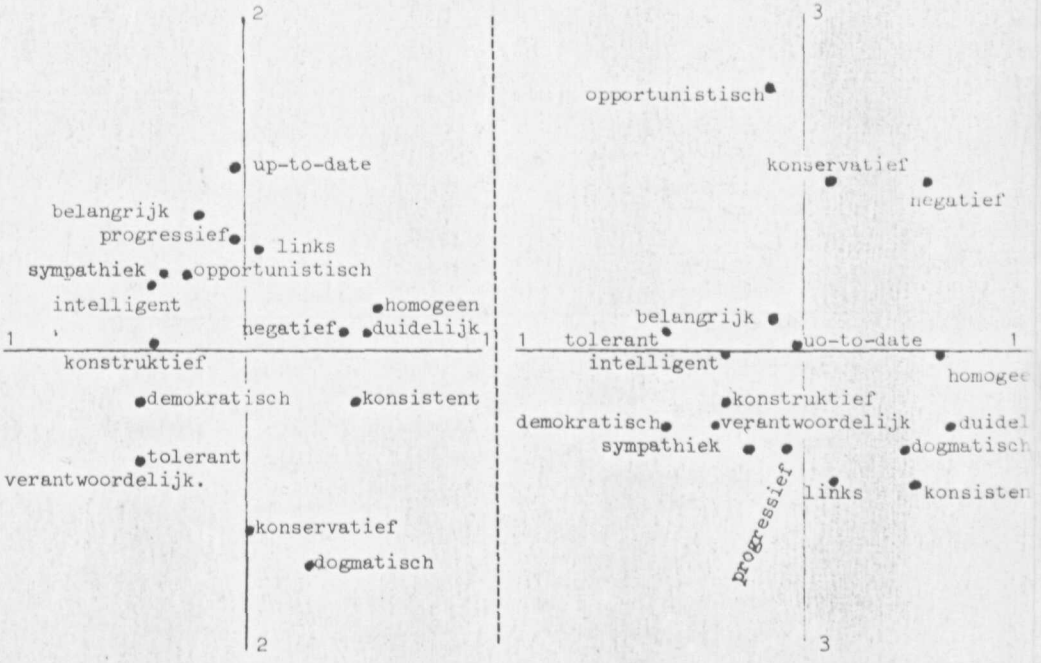


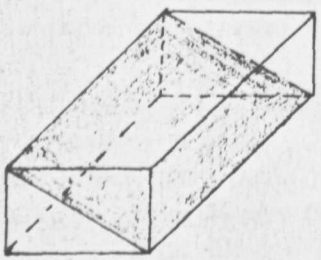
Fig. 9.3. Kenmerken ruimte (De Leeuw data)

Als we in het achterhoofd houden dat het komponentgewicht van de derde as aanzienlijk kleiner is dan die van de andere twee assen, dan kunnen we in figuur 9.2. drie groepen partijen onderscheiden:

-- christen-demokraten KVP, ARP, CHU
-- gematigd links PvdA, PPR, D'66
-- echt links CPN, PSP

Alhoewel de derde as geen hoog gewicht heeft, lijkt het alsof hij dient om het verschil aan te geven tussen de Boerenpartij en de rest. Figuur 9.3 vertoont veel minder struktuur dan figuur 9.2, al is er een zekere neiging voor gelijke kenmerken om samen te gaan. Dit is nauwelijks verrassend als men bedenkt dat de proefpersonen de opdracht kregen om kenmerken aan partijen toe te kennen en niet om kenmerken te schalen. De betekenis van de kenmerkenruimte komt pas goed naar voren in relatie met de partijenruimte. In het TUCKALS 2 programma is een routine opgenomen die de hoofdassen van de eerste en tweede weg op elkaar past. Het resultaat van deze procedure voor dit voorbeeld staat in figuur 9.4. Gezien de kombinaties van kenmerken en partijen is het duidelijk dat de psychologen een duidelijke voorkeur hadden voor de gematigd linkse partijen. Het is ook aardig om te zien dat de CPN en de SGP/GPV gezien werden als even homogeen, helder en konsistent, terwijl ze op vele politieke strijdpunten mijlen van elkaar af staan.

Zoals eerder opgemerkt geeft elk van de voorvlakken van de kernmatrix aan hoe voor iedere psycholoog de komponenten van de kenmerken- en de politieke partijenruimte op elkaar laden. Aangezien de voorvlakken voor de De Leeuw data niet veel verschillen te zien geven - m.a.w. alle elf beoordelende psychologen hadden min of meer gelijke visies op de Nederlandse partijen - laten we hier alleen het gemiddelde voorvlak van de kernmatrix zien (tabel 9.5A).



Figuur 9.6

Een gediagonaliseerde kernmatrix

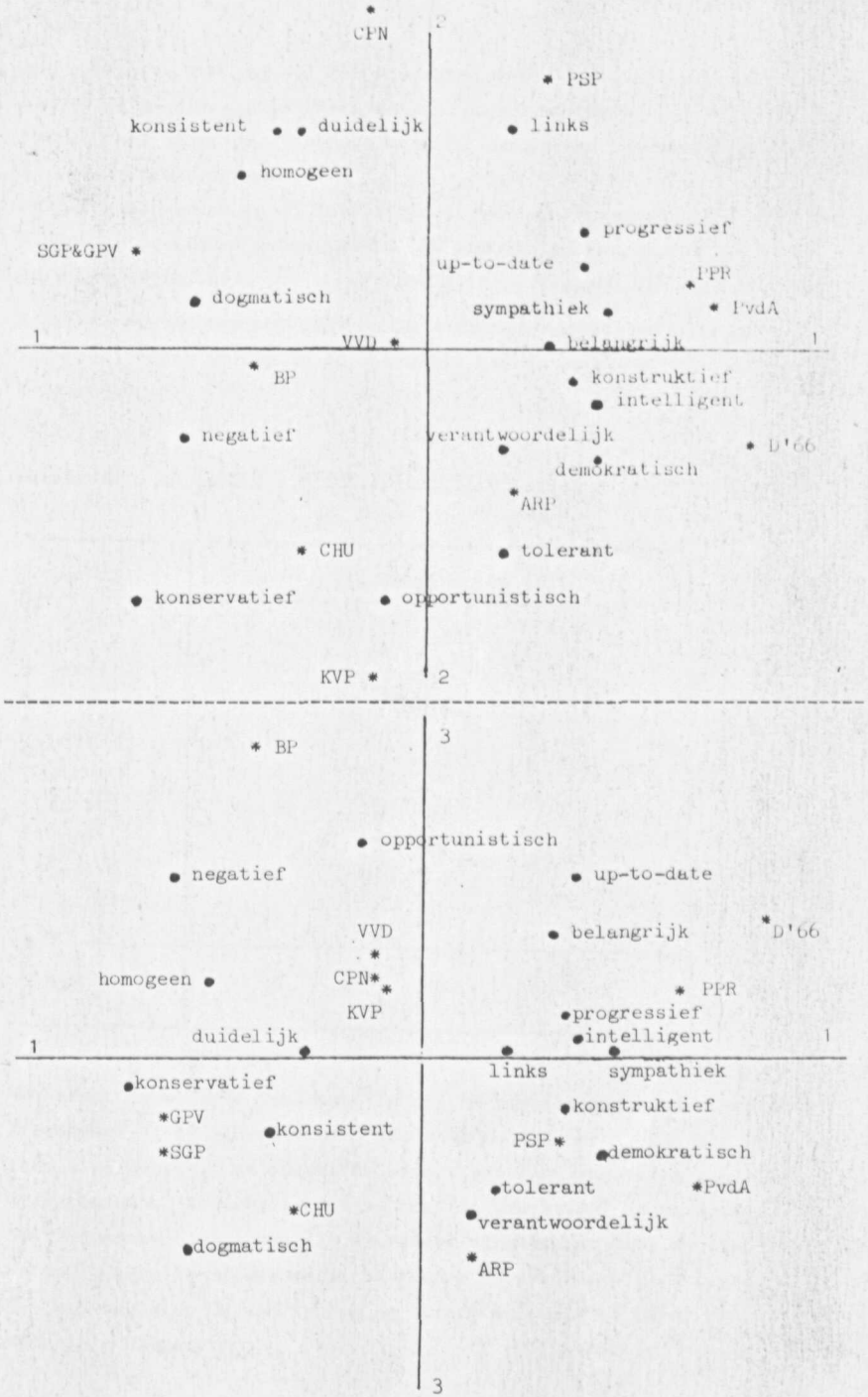


Fig. 9.4. Gekombineerde kenmerken en politieke partijen ruimte (De Leeuw data)

TABEL 9.5. Gemiddeld voorvlak van de kernmatrix van de De Leeuw data

($\times 10^{-1}$)

KENMERKEN

voor rotatie

PARTIJEN	1	2	3
1	53	44	-26
2	35	15	-26
3	13	7	28

(A)

na rotatie

1	2	3
-75	- 3	2
2	45	1
- 2	3	27

(B)

Een erg overzichtelijk beeld van de onderlinge relatie krijgt men hier niet van. Aanzienlijke verheldering krijgen we door de kern matrix zo te transformeren dat hij optimaal (in de kleinste kwadraten zin) 'diagonaal' wordt (zie fig. 9.6). Met de gevonden transformaties moeten natuurlijk ook de komponentenmatrices bijgesteld worden. Deze procedure veranderde (in dit geval) de komponentenruimten niet wezenlijk.

Tabel 9.5.B laat een zeer eenvoudige interpretatie toe van de relaties tussen de komponenten van beide wegen. De i-de komponent (i=1, 2, 3) van de partijen laadt vrijwel uitsluitend op de i-de komponent van de kenmerken en omgekeerd, m.a.w. de partijen die hoog laden op de i- de komponent van de partijenruimte behoren bij de kenmerken die hoog laden op de i-de komponent van de kenmerkenruimte. Het minteken van het (1,1)-element duidt er op dat de eerste komponentassen gespiegeld zijn, zoals ook blijkt uit vergelijking van de figuren 9.2 en 9.3 met 9.4.

Gelijkenis van politieke partijen (V.d.Kamp data)

Ook in 1968 gebruikte V.d.Kamp de methode van successieve intervallen voor stimulusparen om gelijkenisoordelen te verkrijgen van de belangrijkste politieke partijen.

Een groep van 100 psychologie studenten werd gebruikt als beoordelaars. De gegevens zijn dus symetrische 9x9 matrices. Dit zijn dus typisch het soort gegevens die men zou analyseren met meerdimensionele schaalprogramma's.
De gegevens werden eerst dubbel gecentreerd en vervolgens met het TUCKALS 2 programma geanalyseerd. De resultaten van de analyse staan getabuleerd in tabel 9.7 en afgebeeld in de figuren 9.8 en 9.9.

Zoals te verwachten met symmetrische invoergegevens waren G en H identiek, zodat we ons in de diskussie kunnen beperken tot een van de twee wegen.

TABEL 9.7. Gelijkenis tussen politieke partijen (V.d.Kamp data)

partij	komponenten ($x10^{-2}$)		
KVP	15	33	- 3
PvdA	30	-16	27
VVD	2	-16	-54
ARP	19	50	2
CHU	6	46	-12
CPN	-20	-21	49
PSP	7	-22	65
BP	-86	0	-13
D'66	27	-54	-40
Gewicht	39	15	12

Het meest opvallende resultaat is dat de psychologie studenten de Boerenpartij, zó anders vonden dan enig andere partij dat dit effekt alle verschillen tussen de andere partijen onderling geheel overschaduwde. Een gedeeltelijke analyse van hetzelfde materiaal door De Leeuw (1973) had dit al eerder aangetoond. Voor de relaties tussen de andere partijen kunnen we het beste naar figuur 9.9 kijken, waar de waarden op de tweede en derde as tegen elkaar zijn uitgezet.

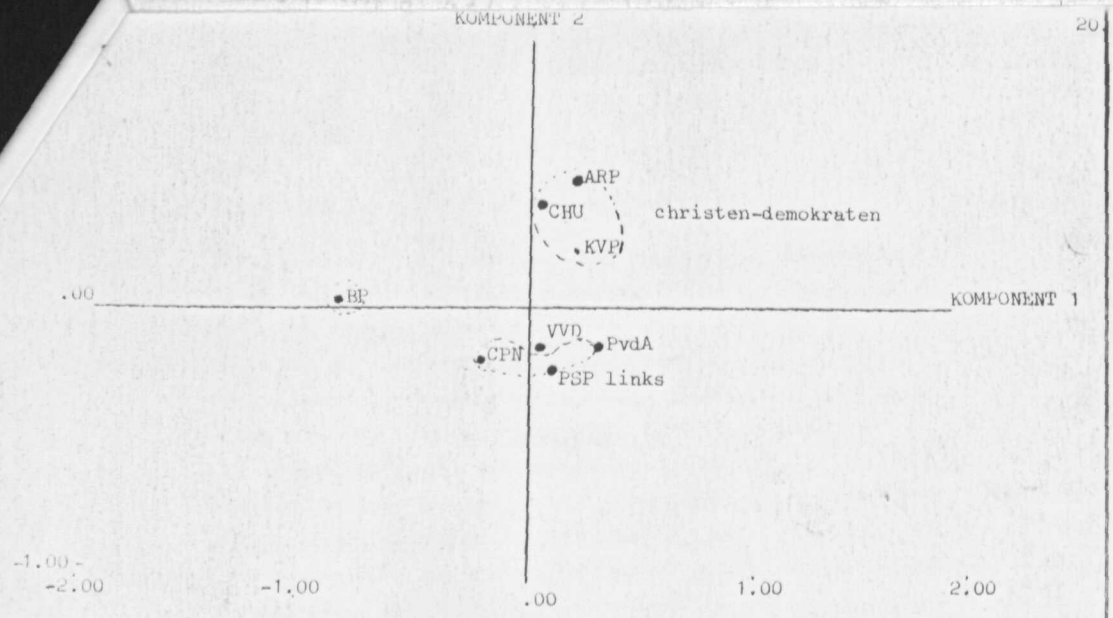


Fig. 9.8 Gelijkenissen van Nederlandse politieke partijen
(Komponent 1 tegen 2)

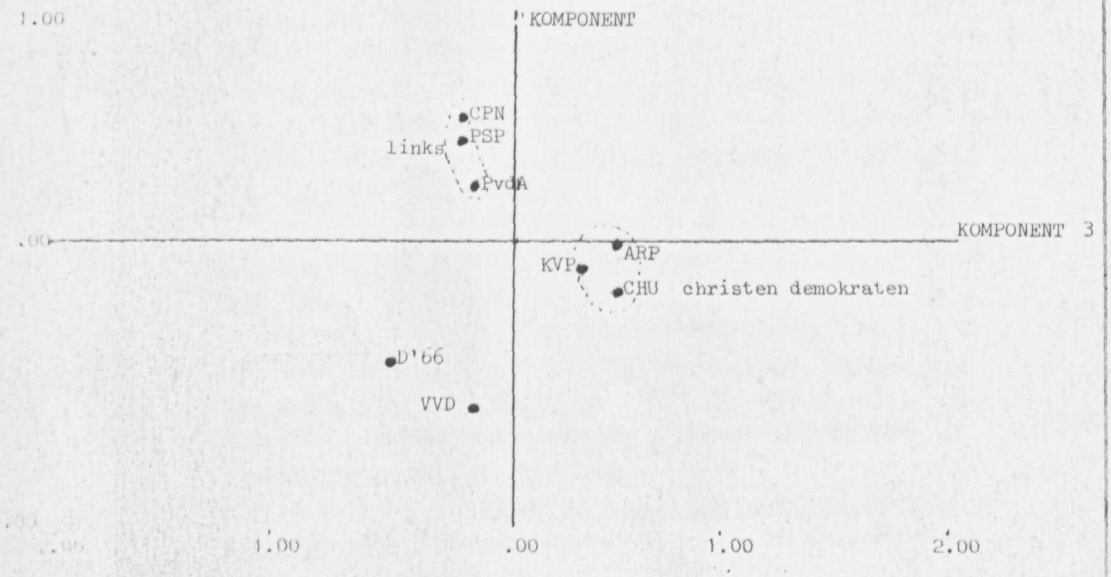


Fig. 9.9 Gelijkenissen van Nederlandse politieke partijen
(Komponent 2 tegen 3)

Als men de assen zou willen benoemen dan zou men de tweede as de religieuse en de derde de links-rechts as kunnen noemen. Vele andere studies hebben dergelijke assen gevonden voor het Nederlandse politieke bestel. Opvallend is de positie van D'66. Waarschijnlijk zagen de studenten D'66 in het begin van zijn bestaan meer als een rechtse (of misschien wel liberale) partij in plaats van een gematigd linkse. De positie van D'66 hier is trouwens nogal verschillend van die bij het vorige voorbeeld. De redenen hiervoor zijn niet duidelijk, al kunnen er natuurlijk wel allerlei aardige verklaringen verzonnen worden.
Gezien het min of meer gelijke gewicht van de tweede en derde as zou men kunnen veronderstellen dat zij beide een gelijke rol hebben gespeeld bij de bepaling van de gelijkenisoordelen.

10.Uitbreidingen van het Tucker 2 model en het TUCKALS 2 algoritme

In het kader van het TUCKALS projekt staan nog een aantal uitbreidingen en verfijning van het hier beschrevene op het programma.

1. een ALS programma voor het algemene Tucker model: TUCKALS 3;
2. uitbreiding van TUCKALS 2 (en TUCKALS 3) naar andere schaalnivo's volgens de principes geformuleerd in Young, De Leeuw en Takane (1977);
3. met name in TUCKALS 2: faciliteiten voor de controle op het model bij zulke multidimensionele schaalmodellen als INDSCAL, IDIOSCAL en PARAFAC.
4. inbouwen van 'missing data' procedures.

Literatuurverwijzingen

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Carroll, J.D. & Chang, J.J. Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika, 1970, 35,283-319.

Carroll, J.D. & Wish, M. Models and methods for three-way multidimensional scaling. In D.H. Krantz, et. al., Contemporary Developments in Mathematical Psychology (Vol.2). San Francisco: Freeman, 1974.

Cochrane,D. & Orcutt, G.H. Applications of least-squares regressions to relationships containing auto-correlated error terms. Journal of the American Statistical Association, 1949, 44, 32-61.

Endler, N.S.,Hunt, J. McV.& Rosenstein, A.J. An S-R inventory of anxiousness. Psychological Monographs, 1962, 76 (17, Whole No. 536).

Geer, J.P. v.d., Toepassing van drieweg-analyse voor de analyse van multiple tijdsreeksen. In: Interim rapport onderzoek groei en ontwikkeling van ziekenhuisorganisaties in Nederland. Leiden: Sociologisch Instituut, Rijksuniversiteit Leiden, 1974.

Israelsson, A. Three-way (or second order) component analysis. In H. Wold en E. Lyttkens, Non linear iterative partial least-squares (NIPALS) estimation procedures, Bulletin of the International Statistical Institute, 1969, 43, 29-51.

Jennrich, R. A generalisation of the multidimensional scaling model of Carroll and Chang (Working papers in phonetics No. 22). Los Angeles: University of California, 1972.

Jones, L.E., & Young, F.W. Structure of a social environment: Longitudinal individual differences scaling of an intact group. Journal of Personality and Social Psychology, 1972, 24, 108-221.

Jöreskog, K.G.; Simultaneous factor analysis in several populations. Psychometrika, 1971, 36, 409-426.

Kroonenberg, P. & de Leeuw, J. TUCKALS 2. A principal component analysis of three mode data (Research Bulletin RB 001-77). Leiden: Datatheorie, University of Leiden, 1977.

Leeuw, J. de Canonical analysis of categorical data. Leiden: Rijksuniversiteit Leiden, 1973.

Leeuw, J. de & van Rijckevorsel, J. HOMALS. Methoden en data nieuwsbrief, 1977, 2, 30-43.

Levin, J. Three-mode factor analysis. Psychological Bulletin, 1965, 64, 442-452.

Osgood, C.E., Suci, G.J. & Tannenbaum, T.H. The measurement of meaning. Urbana: University of Illinois Press, 1957.

Rutishauser, H. Computational aspect of F.L. Bauer's simultaneous iteration method. Numerische Mathematik, 1969, 13, 4-13.

Takane, Y., Young, F.W. & de Leeuw, J. Nonmetric individual differences multidimensional scaling: an alternating least-squares method with optimal scaling features. Psychometrika, 1977, 42, 7-67.

Tucker, L.R. Implications of factor analysis of three way matrices for measurement of change. In C.W. Harris, Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Tucker, L.R. The extension of factor analysis to three dimensional matrices. In H. Gullikson en N. Frederikson, Contributions to Mathematical Psychology. New York: Holt, Rinehart and Winston, 1964.

Tucker, L.R. Some mathematical notes on three-mode factor analysis. Psychometrika, 1966, 31, 279-312.

Tucker, L.R. Relations between multidimensional scaling and three-mode factor analysis. Psychometrika, 1972, 37, 3-27.

Werts, C.E., Jöreskog, K.G., & Linn, R.L. A multitrait-multimethod model for studying growth. Educational and Psychological Measurement, 1972, 32, 655-678.

Young, F.W., de Leeuw, J. & Takane, Y. Quantifying qualitative data, in press.