

Interpretatie van plots bij Correspondentie Analyse

Ben Pelzer, Jan Lammers¹

Samenvatting

In dit artikel gaan we in op de interpretatiemogelijkheden van vier verschillende plots die in zwang zijn bij correspondentie analyse (CA). In de literatuur worden deze interpretatiemogelijkheden weliswaar genoemd (Gifi 1990, Greenacre 1987 1989, Sikkel & Bethlehem 1981) maar een duidelijke richtlijn voor de vraag naar de specifieke mogelijkheden van elke plot wordt zelden gegeven (Israëls 1987). Bovendien is de meeste literatuur over CA geschreven voor de mathematisch geschoolde lezer. Deze tekst poogt inzicht te verschaffen in de belangrijkste interpretatiemogelijkheden zonder al te diep in te gaan op de mathematische achtergronden. Aan het slot bespreken we enkele software produkten en geven we aanstuurcommando's voor het opvragen van de vier typen plots.

¹Research Technische Afdeling Maatschappijwetenschappen
Postbus 9104, 6500 HE, Nijmegen. Tel. 024-3615568.

1. Inleiding

Een van de doelstellingen van CA is het grafisch weergeven van de samenhang tussen de rijen en kolommen van een tabel. Hoewel de techniek is ontstaan als analysemethode voor twee-dimensionele frequentietabellen (kruistabellen) is zij in de loop der jaren ook op andere typen data toegepast. In dit artikel zullen we ons echter beperken tot de toepassing op kruistabellen, waarvan de rijen en kolommen categorieën van nominale variabelen voorstellen en cel (i, j) de frequentie bevat van het aantal eenheden dat voorkomt in rijcategorie i en kolomcategorie j. Een voorbeeld van een dergelijke tabel is tabel 1, waarvan de gegevens ontleend zijn aan een onderzoek gedaan door Arts en anderen (Arts, 1990). De inkomensklassen Laag, Midden en Hoog staan respectievelijk voor inkomens tot fl.2500, inkomens van fl.2500 tot fl.5000 en inkomens hoger dan fl.5000.

Tabel 1: absolute frequenties van partijverwantschap tegen netto huishoudinkomen per maand				
	Laag	Midden	Hoog	marginale kolom
PvdA	130	97	14	241
CDA	76	110	13	199
VVD	32	52	15	99
D'66	24	26	6	56
marginale rij	262	285	48	595

Tabel 2: rijproporties van tabel 1			
Laag	Midden	Hoog	
.539	.402	.058	
.382	.553	.065	
.323	.525	.152	
.429	.464	.107	
.440	.479	.081	

CA kent aan de rij- en kolomcategorieën van de kruistabel coördinaten op een of meer dimensies toe. Als men, zoals meestal gebeurt, kiest voor twee dimensies behelst de grafische voorstelling of plot niet meer dan het weergeven van de ligging van een aantal punten in een plat vlak. Daarbij maken we onderscheid tussen rij- en kolompunten die achtereenvolgens de rij- en kolomcategorieën van de tabel voorstellen. De wijze waarop mathematisch voor elk punt de ligging - ofwel de coördinaten op de dimensies - bepaald kan worden, is voor de doorsnee gebruiker van de techniek misschien minder relevant. Wat hij/zij echter wel dient te weten, is dat er verschillende manieren zijn om tot coördinaten te komen. En wat belangrijker is: dat iedere manier leidt tot een andere plot die eigen, specifieke interpretatiemogelijkheden biedt.

Omdat deze tekst met name gaat over de diverse plots en bijhorende interpretatiemogelijkheden, hebben we ervoor gekozen geen uitvoerige mathematische uitwerking van CA te presenteren. We lichten alleen kort toe dat CA gezien kan worden als een techniek waarbij correlaties gemaximaliseerd worden. Vanuit deze blik op CA is het betrekkelijk eenvoudig om in te zien dat er verschillende manieren bestaan om tot schaalwaarden te komen voor de rij- en kolomcategorieën van een kruistabel.

2 Het bepalen van rij- en kolomscores met een maximale correlatie.

Een belangrijke gedachte achter CA is dat de volledige en vaak complexe samenhang die kan bestaan in een kruistabel van twee variabelen ontleed kan worden in een aantal los

van elkaar staande componenten die elk één bepaald aspect of dimensie van die complexe samenhang weergeven. We beschrijven nu kort hoe het ontrafelen van de totale samenhang in een aantal onafhankelijke delen bij CA wordt uitgewerkt.

Stel dat we, uitgaande van de frequenties in tabel 1, twee nieuwe metrische variabelen 'partij' en 'inkomen' willen construeren waarop we elk van de 595 personen in de tabel een score willen toekennen. Personen die dezelfde politieke partij kiezen willen we ook dezelfde score geven op de nieuwe variabele 'partij', zodat er slechts vier *verschillende* scores op 'partij' voorkomen, één voor elke politieke partij; ook zijn er slechts drie verschillende scores op 'inkomen' mogelijk, één voor elke inkomensklasse. Bij CA gaat het nu om de vraag: welke vier scores op 'partij' en welke drie scores op 'inkomen' moeten we kiezen opdat de Pearson correlatie van 'partij' en 'inkomen' berekend over de 595 personen een maximale waarde heeft? Daar de Pearson correlatie invariant is onder lineaire transformatie zijn de gezochte scores op 'partij' en 'inkomen' alleen te bepalen wanneer er vooraf restricties aan worden opgelegd. We kiezen als restricties dat de scores gestandaardiseerd moeten zijn: de gemiddelden van de 595 scores op 'partij' en 'inkomen' moeten gelijk zijn aan 0 en de standaardafwijkingen gelijk aan 1. Gegeven deze restricties kunnen we met behulp van CA de vier 'partij'-scores en de drie 'inkomen'-scores achterhalen evenals de waarde van de Pearson correlatie tussen de beide variabelen.

Zijn de gestandaardiseerde scores eenmaal gevonden dan worden via een lineaire transformatie van die scores de *coördinaten* bepaald van de partijen en inkomensklassen op de *eerste* dimensie van de plots die CA kan genereren. De gevonden standaardscores noemen we daarom wel de standaardscores op dimensie 1, terwijl de bijhorende correlatie ook wel de canonische correlatie van dimensie 1 wordt genoemd. (De term 'canonische correlatie' is afkomstig uit de techniek genaamd Canonische Correlatie Analyse, waarmee CA een sterke gelijkenis vertoont.) Aan de hand van de coördinaten van de partijen en inkomensklassen kan de onderzoeker komen tot een theoretische interpretatie van de eerste dimensie. De wijze waarop dat gaat lichten we verderop toe.

Er is nog een tweede paar gestandaardiseerde variabelen te construeren, zeg 'partij2' en 'inkomen2', zodanig dat 'partij2' en 'inkomen2' maximaal correleren met elkaar, onder de restrictie dat 'partij' en 'partij2' een correlatie van nul hebben, evenals 'inkomen' en 'inkomen2'. Aan de hand van dit tweede paar standaardscores worden via een lineaire transformatie van die scores de coördinaten berekend op de tweede dimensie van de plot. We spreken daarom wel over de standaardscores op dimensie 2 en de correlatie tussen 'partij2' en 'inkomen2' heet de canonische correlatie van dimensie 2. De coördinaten op de tweede dimensie maken een theoretische interpretatie van de tweede dimensie mogelijk.

Bij tabellen met meer dan drie rijen en kolommen zijn doorgaans meer dan twee paren, bestaande uit één gestandaardiseerde rij- en één gestandaardiseerde kolomvariabele, te construeren die onderling maximaal correleren, maar een correlatie van nul hebben met elke eerder geconstrueerde rij- c.q. kolomvariabele. Elk paar standaardscores kan lineair getransformeerd worden tot coördinaten van rijen en kolommen op een dimensie. Aan de hand van die coördinaten kan de onderzoeker elke dimensie theoretisch interpreteren.

Bij een kruistabel met I rijen en J kolommen is het maximale aantal D van onafhankelijke dimensies dat kan worden onderscheiden gelijk aan het minimum van I-1 en J-1. Elke dimensie draagt een deel bij aan de volledige samenhang in de tabel en die bijdrage is onafhankelijk van de bijdragen van alle andere dimensies. Vanwege de onafhankelijkheid van de dimensies is de volledige samenhang te beschouwen als de som van de bijdragen van alle afzonderlijke dimensies. Dit idee komt tot uitdrukking in de volgende formule die de door Pearson ontwikkelde associatiemaat $\phi^2 (= \chi^2/N)$ koppelt aan de canonische correlaties van de D te onderscheiden dimensies:

$$\phi^2 = \sum_{d=1}^D \lambda_d^2 \quad (1)$$

Hierin staat λ_d voor de canonische correlatie van dimensie d. Formule (1) laat zien dat de volledige mate van samenhang ϕ^2 , bij CA ook wel de totale inertie van de tabel genoemd, gelijk is aan de kwadratsom van de canonische correlaties van alle D te onderscheiden dimensies. De bijdrage van elke afzonderlijke dimensie d aan de totale inertie is gelijk aan λ_d^2 en deze bijdrage noemt men wel de inertie van dimensie d. Nu geldt in theorie voor elke kruistabel dat $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_D$ maar in de praktijk geldt voor de meeste tabellen dat $\lambda_1 > \lambda_2 > \lambda_3 > \dots > \lambda_D$. In woorden: de eerste dimensie levert van alle D te onderscheiden dimensies de grootste bijdrage aan de totale inertie; de tweede dimensie levert de op één na hoogste bijdrage van alle dimensies, de derde dimensie levert de op twee na hoogste bijdrage, en zo kunnen we verder gaan tot we komen bij de laatste dimensie D, die de kleinste bijdrage levert. Tot zover wat betreft de opsplitsing van de volledige samenhang in onafhankelijke dimensies. We keren nu nog even terug naar de standaardscores en coördinaten.

In alle verderop te bespreken CA-plots geldt voor elke dimensie dat de coördinaten van rijen en kolommen op die dimensie een lineaire transformatie zijn van de gevonden standaardscores op die dimensie. Het feit dat er verschillende CA-plots bestaan komt voort uit het feit dat er verschillende lineaire transformaties kunnen worden toegepast op de standaardscores om te komen tot coördinaten. Uit het oneindig grote aanbod van mogelijke lineaire transformaties kiest CA er enkele die het meest te bieden hebben qua interpretatiemogelijkheden van de plots. Bij elke CA-plot geven we aan hoe de betreffende transformatie van standaardscores in coördinaten eruit ziet. De aanduiding 'lineaire transformatie' is hier eigenlijk nog te ruim, omdat het bij elke plot gaat om een vermenigvuldiging van de kolom-standaardscores met een bepaalde constante en de vermenigvuldiging van de rij-standaardscores met een bepaalde constante.

Tussen de standaardscores van de rijen en die van de kolommen bestaat een relatie die voor elke dimensie d geldt. De standaardscore r_{id} van rij i op dimensie d kan als volgt worden afgeleid uit de J standaardscores k_{jd} van de J kolommen van een kruistabel:

$$r_{id} = \lambda_d^{-1} \sum_{j=1}^J p_{ji} k_{jd} \quad (2)$$

Hierin staat λ_d voor de canonische correlatie van dimensie s. p_{ji} staat voor de rijproportie in kolom j gegeven rij i; voor tabel 1 staan deze proporties p_{ji} vermeld in tabel 2.

Formule (2) staat bekend onder de naam 'overgangs-formule' of 'transitie-formule' omdat ze aangeeft hoe de standaardscores van de kolommen overgaan in die van de rijen. Er bestaat ook een 'omgekeerde' overgangsformule die laat zien hoe de standaardscores van de rijen overgaan in die van de kolommen. Naast overgangsformules voor de standaardscores, bestaan er overgangsformules die laten zien hoe de *coördinaten* van kolommen overgaan in coördinaten van rijen en omgekeerd. Omdat het voor de interpretatie van een plot relevant kan zijn te weten hoe de coördinaten van rijen en kolommen aan elkaar gerelateerd zijn, geven we bij elke van de vier te bespreken plots de overgangsformule voor de coördinaten van rijen naar kolommen of omgekeerd.

Alvorens bij elke afzonderlijke plot stil te staan gaan we nu eerst in op enkele belangrijke interpretatiemogelijkheden die voor alle plots in meer of mindere mate gelden.

3 De drie belangrijkste interpretatiemogelijkheden

Ca is een techniek bedoeld om de samenhang tussen nominale variabelen bloot te leggen. Het lijkt daarom wellicht vreemd dat bij CA standaardscores voor rijen en kolommen worden gezocht met een zo hoog mogelijke Pearson correlatie. Het blijkt echter dat de daaruit via lineaire transformatie afgeleide coördinaten de rij- en kolompunten zodanig in een plat vlak plaatsen dat interpretaties mogelijk worden die juist uitstekend aansluiten bij ons begrip van samenhang voor het nominale meetniveau. De eerste 'nominale' interpretatiemogelijkheid die zich aanbiedt is die van de chikwadraat-afstand.

De chikwadraat-afstand is een afstandsmaat voor twee rijen (of kolommen) van een kruistabel. Bij drie van de vier te bespreken plots is deze afstandsmaat van toepassing. Laten we ons voorlopig even tot de rijen van de kruistabel beperken. Als we de chikwadraat-afstand tussen twee rijen i en k aangeven met het symbool δ_{ik} dan is het kwadraat van deze afstand als volgt gedefinieerd:

$$\delta_{ik}^2 = \sum_{j=1}^J \frac{(p_{ji} - p_{jk})^2}{p_j} \quad (3)$$

Hierin staat J voor het aantal kolommen in de tabel. Verder staat p_{ji} voor de proportie eenheden binnen rij i in cel (i, j) . Deze proportie is gelijk aan de frequentie in cel (i, j) gedeeld door de marginale frequentie in rij i . Laten we een dergelijke proportie binnen een rij een rijproportie noemen. Ook p_{jk} is een rijproportie, te weten de proportie eenheden die zich binnen rij k in cel (k, j) bevindt. Voorts is p_j een marginale proportie: de marginale frequentie in kolomcategorie j gedeeld door N , het totaal aantal eenheden in de tabel. Voor tabel 1 zijn de rijproporties en marginale proporties gegeven in tabel 2.

Een rij met rijproporties optellend tot 1 heet bij CA een 'rijprofiel'. In tabel 2 zien we dat iedere politieke partij gekenmerkt wordt door een bepaald rijprofiel. Ook voor alle respondenten tezamen kunnen we een rijprofiel berekenen, het marginale rijprofiel. Dit staat in de onderste regel van tabel 2. De chikwadraat-afstand is zoals uit (3) blijkt een maat voor het verschil tussen twee rijprofielen. Elke term in de som is groter naarmate de overeen-

komstige proporties (te weten die in dezelfde kolom) meer van elkaar afwijken. In tabel 2 zien we dat de rijen PvdA en CDA meer op elkaar lijken dan de rijen PvdA en VVD. De chikwadraat-afstand tussen PvdA en CDA (.32) is dan ook kleiner dan die tussen PvdA en VVD (.50). In het algemeen geldt: hoe sterker twee rijen verschillen wat betreft hun profiel des te groter is de (chikwadraat) afstand tussen de twee rijpunten in de plot.

Uit de noemer p_j in (3) blijkt dat de verschillen tussen twee rijproporties in een bepaalde kolomcategorie 'gewogen' worden door ze te relateren aan de marginale proportie in die kolomcategorie. Verschillen in kolommen met een kleine marginale proportie dragen meer bij aan de chikwadraat-afstand dan even grote verschillen in kolommen met een grote marginale proportie. De uitspraak van zoëven kunnen we nu aanscherpen tot: de chikwadraat-afstand tussen twee rijen neemt toe, wanneer de gewogen verschillen tussen overeenkomstige rijproporties toenemen. De weging van kolombijdragen aan de afstand tussen twee rijen is intuïtief plausibel. Zouden we CA op andere data dan frequenties toepassen, dan kan dat alleen wanneer ook dan de weging zinvol is.

Op soortgelijke wijze als in (3) kunnen we de afstand bepalen tussen bijvoorbeeld rij PvdA en de marginale rij. Als we een dergelijke afstand tussen rij i en de marginale rij aangeven met symbool δ_i dan geldt:

$$\delta_i^2 = \sum_{j=1}^I \frac{(p_{ji} - p_j)^2}{p_j} \quad (4)$$

Men kan aantonen dat $\sum_{i=1}^I f_i \delta_i^2 = \chi^2$, waarin I staat voor het aantal rijen van de tabel, f_i voor de marginale frequentie van rij i en χ^2 voor de Pearson chikwadraatwaarde van de tabel. Het is vanwege deze relatie dat men spreekt over 'chikwadraat-afstanden'. Het symbool δ_i noemt men kortweg de 'chikwadraat-afstand' van rij i . In sommige plots vinden we de chikwadraat-afstand van een rij terug als de afstand van het rijpunt tot de oorsprong. De oorsprong (het punt dat op alle dimensies een schaalwaarde van nul heeft) geeft dan de ligging weer van de marginale rij. Hoe groter de chikwadraat-afstand van een rij is, des te sterker wijkt de rij af van het marginale rijprofiel en des te verder verwijderd van de oorsprong vinden we in de plot het betreffende rijpunt. Op basis van tabel 2 kunnen we nu al stellen dat rij D'66 het dichtst bij de oorsprong van de plot zal liggen, omdat het profiel van deze rij het meest op het marginale rijprofiel lijkt.

Naast rijprofielen kunnen we ook kolomprofielen berekenen en een marginaal kolomprofiel. Op basis van de kolomprofielen zijn chikwadraat-afstanden te berekenen voor de kolommen met soortgelijke formules als (3) en (4). Wat de plots betreft zien we dat in één plot de chikwadraat-afstand alleen van toepassing is op de rijpunten, in een andere plot alleen op de kolompunten, en in een derde plot op beide puntensets.

De tweede interpretatiemogelijkheid van de plots van CA is de mate van over/ondervertegenwoordiging. Met oververtegenwoordiging in een cel bedoelen we dat de geobserveerde celfrequentie groter is dan de - op basis van statistische onafhankelijkheid van rij- en kolomvariabele - verwachte frequentie. Laten we de mate van over/ondervertegenwoordiging in cel (i, j) v_{ij} noemen en deze definiëren als

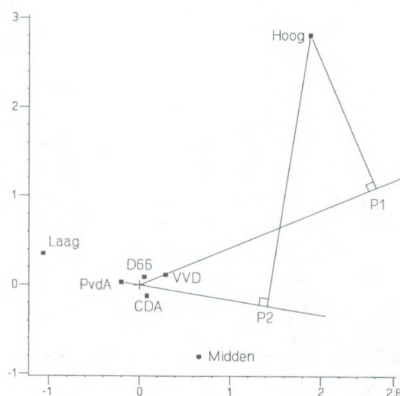
$$v_{ij} = (f_{ij}/e_{ij}) - 1 = (p_{ji}/p_j) - 1 \quad (5)$$

Hierin staat f_{ij} voor de geobserveerde en e_{ij} voor de verwachte frequentie in cel (i, j) . Als f_{ij} groter is dan e_{ij} , ofwel $p_{j|i}$ groter dan p_j , dan is v_{ij} positief en is er sprake van oververtegenwoordiging; is f_{ij} kleiner dan e_{ij} , ofwel $p_{j|i}$ kleiner dan p_j , dan is v_{ij} negatief en is er ondervertegenwoordiging. Kijken we in tabel 2 naar cel $(1, 3)$ dan zien we dat binnen de PvdA aanhang sprake is van een ondervertegenwoordiging van inkomenscategorie Hoog; v_{13} is gelijk aan $(.058/.081)-1=-.28$. Binnen de VVD aanhangers is categorie Hoog oververtegenwoordigd; v_{33} is gelijk aan $(.152/.081)-1=+.88$.

De waarde van v_{ij} is af te leiden uit de coördinaten van rijpunt i en kolompunt j in sommige plots die CA levert. In figuur 1 presenteren we een dergelijke plot, gemaakt aan de hand van de data in tabel 1. De punten van partijen, inkomenscategorieën en de oorsprong (symbool +) zijn rechtstreeks overgenomen uit een met CA gemaakte plot. De punten P1 en P2 en de lijnen zijn naderhand toegevoegd. Om de waarde van v_{33} voor de combinatie VVD en Hoog te achterhalen, trekken we een lijn door de oorsprong en het punt VVD. Daarop projecteren we loodrecht punt Hoog hetgeen projectiepunt P1 oplevert. Als we nu de afstand van VVD tot de oorsprong (uitgedrukt in de schaaleenheid van de assen) vermenigvuldigen met de afstand van P1 tot de oorsprong, dan krijgen we de waarde voor v_{33} , ofwel .88. Op dezelfde manier kunnen we de waarde van v_{13} afleiden: projectie van Hoog op de lijn oorsprong-PvdA levert P2. Vermenigvuldigen van de afstand PvdA-oorsprong met de afstand P2-oorsprong levert .28; deze uitkomst moet nog met -1 worden vermenigvuldigd, omdat P2 aan de andere zijde van de oorsprong ligt dan punt PvdA. P1 ligt aan dezelfde zijde als punt VVD zodat vermenigvuldigen met -1 niet nodig is. In de geschetste werkwijze trekken we eerst een lijn door de oorsprong en het rijpunt en projecteren daarop het kolompunt. Andersom te werk gaan kan eveneens en levert dezelfde waarde voor v_{ij} .

De richting waarin een rij- en kolompunt liggen vanuit de oorsprong geeft blijkbaar aan of er in een bepaalde cel een over- dan wel ondervertegenwoordiging is. In figuur 1 is de hoek VVD-oorsprong-Hoog scherp, waardoor aan lijnstuk P1-oorsprong een positief teken wordt toegekend evenals aan v_{33} . De hoek PvdA-oorsprong-Hoog is daarentegen stomp, waardoor aan lijnstuk P2-oorsprong een negatief teken wordt toegekend evenals aan v_{13} . We kunnen ook zeggen: liggen een rijpunt i en kolompunt j in dezelfde richting vanuit de oorsprong gezien (scherpe hoek) dan is er in de cel sprake van een oververtegenwoordiging. Liggen ze in tegengestelde richting (stompe hoek) dan is er ondervertegenwoordiging. Naarmate bij gegeven richtingen (hoek) de afstanden van rij- en kolompunt tot de oorsprong toenemen, des te sterker is de mate van over/ondervertegenwoordiging.

Figuur 1: over/ondervertegenwoordiging als produkt van twee afstanden



De derde interpretatiemogelijkheid is van toepassing bij elk van de vier te bespreken plots en betreft de twee assen of dimensies van de plot. Om de assen te kunnen interpreteren projecteren we de rij- en kolompunten loodrecht op de assen. Doen we dat bij de plot in figuur 3, die we verderop uitvoerig zullen bespreken, op de horizontale as (de eerste dimensie) dan levert dit figuur 2.

In figuur 1 is de gekwadeerde afstand van een rijpunt i tot de oorsprong precies gelijk is aan δ_i^2 . In figuur 2 is de gekwadeerde afstand van een rijpunt i tot de oorsprong nog slechts een deel van δ_i^2 . Het is dat deel, dat door de eerste dimensie wordt 'gereproduceerd'. Laten we dit deel aanduiden met het symbool $\delta_{i(1)}^2$.

We stelden eerder dat $\chi^2 = \sum_{i=1}^I f_i \delta_i^2$. We introduceren nu de grootheid $\chi_{(1)}^2$ en definiëren deze als: $\chi_{(1)}^2 = \sum_{i=1}^I f_i \delta_{i(1)}^2$. In woorden is $\chi_{(1)}^2$ het door de eerste dimensie gereproduceerde deel van de totale geobserveerde Pearson chikwadratwaarde χ^2 van de tabel. Zo zien we dat de plot gemaakt op basis van maximaal gecorreleerde en vervolgens linear getransformeerde standaardcores meer te bieden heeft dan alleen de interpretatie van chikwadrat-afstanden: de plot maakt het ons mogelijk de bijdrage van elke dimensie te bepalen aan de totale mate van samenhang, gemeten via Pearson's χ^2 grootheid.

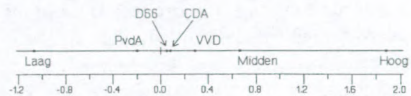
Er bestaat een eenvoudige relatie tussen de in paragraaf 2 besproken inertie λ_1^2 van dimensie 1 en $\chi_{(1)}^2$, te weten $\chi_{(1)}^2 / N = \lambda_1^2$. Het zoeken naar maximaal correlerende scores voor rijen en kolommen blijken we nu te kunnen 'vertalen' in: het zoeken naar een dimensie die zo veel mogelijk van de totale χ^2 waarde van de tabel reproduceert of verklaart.

Door rijpunt i in figuur 3 te projecteren op de verticale as (de tweede dimensie) krijgen we $\delta_{i(2)}^2$ zijnde het deel van δ_i^2 dat gereproduceerd wordt door de tweede dimensie. De grootheid $\chi_{(2)}^2 = \sum_{i=1}^I f_i \delta_{i(2)}^2$ geeft het deel van χ^2 dat gereproduceerd wordt door de tweede dimensie. Deze dimensie wordt zo gekozen dat zij loodrecht op de eerste staat en dat $\chi_{(2)}^2$ zo groot mogelijk is. Voor dimensie 2 geldt dat $\chi_{(2)}^2 / N = \lambda_2^2$. De tweede dimensie is meestal 'minder belangrijk' dan de eerste omdat $\chi_{(2)}^2 \leq \chi_{(1)}^2$.

Wetende dat de dimensies de componenten aangeven waaruit de totale samenhang in de kruistabel is opgebouwd, rijst de vraag hoe we de afzonderlijke dimensies kunnen interpreteren. Om deze vraag te beantwoorden leggen we eerst uit, aan de hand van figuur 2, hoe dimensie 1 de mate van over/ondervertegenwoordiging v_{ij} reproduceert. We zullen deze gereproduceerde grootheid aanduiden met het symbool $v_{ij(1)}$. Aan de hand van $v_{ij(1)}$ komen we tot een interpretatie van de dimensie.

Voor cel (1,3) van de kruistabel, ofwel partij PvdA en inkomen Hoog, leiden we $v_{13(1)}$ af uit figuur 2 op de wijze waarop we dat eerder bij figuur 1 deden. De lijn door het punt PvdA en de oorsprong is hier de horizontale as zelf; het projectiepunt van Hoog op die lijn valt samen met punt Hoog, zodat de afstand projectiepunt-oorsprong overeen komt met de afstand Hoog-oorsprong. De waarde van $v_{13(1)}$ vinden we nu door de afstand PvdA-oorsprong te vermenigvuldigen met de afstand Hoog-oorsprong. De uitkomst vermenigvuldigen we met -1 omdat PvdA en Hoog in tegengestelde richting liggen ten opzichte van de oorsprong. Er is dus sprake van een gereproduceerde *ondervertegenwoordiging*. D66, CDA en VVD liggen aan dezelfde kant van de oorsprong als Hoog; bij deze partijen en Hoog wordt dan ook een *oververtegenwoordiging* gereproduceerd. Voorts zien

Figuur 2: punten geprojecteerd op de horizontale as.



we in figuur 2 dat de afstand D66-oorsprong kleiner is dan afstand CDA-oorsprong en dat deze laatste weer kleiner is dan afstand VVD-oorsprong. Daarom geldt dat $v_{43(1)} < v_{23(1)} < v_{33(1)}$. Zo kunnen we uit figuur 2 afleiden dat dimensie 1 slechts twee typen partijen onderscheidt, te weten

- 1 partijen die positief scoren op dimensie 1, ofwel rechts van de oorsprong liggen in figuur 2; zij zijn oververtegenwoordigd in inkomensklasse Hoog, zwakker oververtegenwoordigd in inkomensklasse Midden, en ondervertegenwoordigd in inkomensklasse Laag; verder geldt: hoe groter de oververtegenwoordiging bij Midden en Hoog, des te groter de ondervertegenwoordiging bij Laag;
- 2 partijen die negatief scoren op dimensie 1, ofwel links van de oorsprong liggen in figuur 2; zij zijn oververtegenwoordigd in inkomensklasse Laag, ondervertegenwoordigd in inkomensklasse Midden, en sterker ondervertegenwoordigd in inkomensklasse Hoog; hoe groter de oververtegenwoordiging bij Laag, des te groter de ondervertegenwoordiging bij Midden en Hoog.

Binnen deze twee typen bestaat enkel variatie in de grootten van genoemde over- c.q. ondervertegenwoordigingen. Met deze partijen-indeling hebben we meteen een interpretatie gegeven van dimensie 1. De dimensie schetst een bepaald beeld van de samenhang tussen inkomen en partijkeuze. In het algemeen is het beeld weergegeven door één dimensie een vereenvoudigde weergave van de complexe samenhang tussen rijen en kolommen van de kruistabel.

We benadrukken nog eens dat we in figuur 2 de relatieve plaatsing van de rijen vergelijken met de relatieve plaatsing van de kolommen. Het gaat dus niet om de afstanden van rijpunten tot kolompunten. De strategie luidt: bewegen we ons bij de partijen van PvdA via D66 en CDA naar VVD dan verschuiven de oververtegenwoordigingen bij de inkomens van Laag naar Midden en Hoog. Hoewel de afstand van VVD tot Laag ongeveer even groot is als die tot Hoog, concluderen we toch dat bij de VVD hoge inkomens oververtegenwoordigd zijn en lage inkomens ondervertegenwoordigd.

Omdat de interpretatie van dimensies bij alle te behandelen CA-plots mogelijk is gaan we er niet verder op in bij de bespreking van iedere afzonderlijke plot. Bij elke plot komt de interpretatie van dimensie 1 overeen met die welke in deze paragraaf is vermeld. De interpretatie van de tweede dimensie verloopt op analoge wijze.

Hiermee hebben we de drie belangrijkste interpretatiemogelijkheden van CA-plots besproken. We zullen nu de specifieke kenmerken van vier typen plots afzonderlijk bekijken. We duiden ze voor het gemak aan als plot A, B, C en D.

3 Plot A: asymmetrisch, rijpunten binnen kolompunten

Figuur 3 geeft plot A behorend bij de tabel 1. De plot komt overeen met die in figuur 1. Formules (6a) en (6b) geven aan hoe bij deze plot de coördinaten van rijen en kolommen op een willekeurige dimensie d zijn gerelateerd aan de standaardcores op dimensie d ; overgangsformule (6c) geeft weer hoe voor dimensie d de rijcoördinaten zijn af te leiden uit de kolomcoördinaten.

$$R_{id} = \lambda_d r_{id} \quad (6a)$$

$$K_{jd} = k_{jd} \quad (6b)$$

$$R_{id} = \sum_{j=1}^J p_{j|i} K_{jd} \quad (6c)$$

R_{id} en K_{jd} staan voor de coördinaten van respectievelijk rij i en kolom j op dimensie d ; r_{id} en k_{jd} staan voor de standaardscores van rij i respectievelijk kolom j op dimensie d ; λ_d staat voor de canonische correlatie van dimensie d .

Uit (6a) blijkt dat de rij-standaardscores r_{id} worden vermenigvuldigd met λ_d om te komen tot de rijcoördinaten R_{id} in de plot. Omdat λ_d meestal dicht bij 0 ligt dan bij 1 (een waarde van 0.4 is bijvoorbeeld al tamelijk hoog) liggen de rijcoördinaten R_{id} meestal dicht bij nul dan de rij-standaardscores r_{id} . De kolomcoördinaten K_{jd} daarentegen ondergaan een dergelijk 'krimping' niet: zij zijn gelijk aan de kolom-standaardscores k_{jd} , zoals te zien in (6b). Het gevolg van deze ongelijke behandeling van rijen en kolommen is in de plot duidelijk zichtbaar: de rijpunten liggen, zoals te zien in figuur 3, tamelijk dicht rond de oorsprong en de kolompunten bevinden zich meer in de periferie van de plot.

Overgangsformule (6c) toont aan dat, voor elke dimensie, rijcoördinaat R_{id} het gewogen gemiddelde is van de J kolomcoördinaten K_{jd} , waarbij de gewichten bestaan uit de rijproporties $p_{j|i}$ van rij i . Men zegt wel dat elk rijpunt in het zwaartepunt van de gewogen kolompunten ligt. We zouden ons bij de plot in figuur 3 de voorstelling kunnen maken dat elk perifeer kolompunt j een bepaalde aantrekkingskracht ter grootte van $p_{j|i}$ uitoefent op het meer centraal gelegen rijpunt i . Een bepaald kolompunt j waarvoor $p_{j|i}$ relatief groot is heeft een sterkere aantrekkingskracht op rijpunt i dan andere kolompunten met kleinere $p_{j|i}$ waarden. In figuur 3 is bijvoorbeeld te zien dat de PvdA dicht bij Laag ligt dan bij Midden en Hoog als gevolg van het feit dat, zoals we in tabel 2 zien in rij PvdA, categorie Laag harder aan de PvdA trekt dan de twee andere inkomensklassen doen. Ook zien we figuur 3 dat Hoog relatief ver verwijderd ligt van *alle* vier de partijen: Hoog heeft blijkbaar op geen enkele partij een sterke aantrekkingskracht, hetgeen is af te lezen aan de kleine proporties in kolom Hoog van tabel 2. Het kan ook voorkomen dat slechts één kolompunt j aan rijpunt i trekt en dat de andere kolompunten geen enkele aantrekkingskracht uitoefenen. In dat extreme geval is $p_{j|i} = 1$ en zijn alle overige rijproporties van rij i gelijk zijn aan 0 en ligt rijpunt i op exact dezelfde plaats als kolompunt j .

Via formule (6c) willen we een belangrijk kenmerk van de plot toelichten. De kolom-standaardscores k_{jd} op dimensie d hebben een gemiddelde van nul, berekend over alle eenheden in de kruistabel, ofwel $(\sum_{j=1}^J f_{.j} K_{jd}) / N = \sum_{j=1}^J p_j K_{jd} = 0$. Omdat volgens (6b) de coördinaten van de J kolompunten gelijk zijn aan de standaardscores geldt eveneens dat $\sum_{j=1}^J p_j K_{jd} = 0$. Uit (6c) nu blijkt dat, wanneer we alle kolomcoördinaten kennen, we eenvoudig de coördinaat van een rij kunnen uitrekenen door het betreffende rijprofiel in (6c) te substitueren. We kunnen zelfs met (6c) de coördinaat berekenen van een fictief rijprofiel dat helemaal niet bij een van de rijen in de kruistabel hoort. Stel bijvoorbeeld eens dat we met (6c) de coördinaat R_{marg} zouden willen berekenen van het rijprofiel dat hoort bij de marginale rij, zoals onder tabel 2 weergegeven. Op basis van (6c) zou deze coördinaat gelijk zijn aan $\sum_{j=1}^J p_j K_{jd}$. Zou juist echter toonden we aan dat $\sum_{j=1}^J p_j K_{jd} = 0$, zodat we kunnen concluderen dat de waarde nul op dimensie d de plaats aangeeft waar de marginale rij op die dimensie thuishoort. Omdat dit voor elke dimensie geldt, geeft de oorsprong van de plot aan waar het rijpunt behorend bij de marginale rij zich bevindt. Bij het interpreteren van de chikwadraatafstand van een rij i tot de marginale rij kunnen we

daarvan gebruik maken. Op soortgelijke wijze is met behulp van de 'omgekeerde' overgangsformule aantoonbaar dat het nulpunt tevens de plaats is waar het kolompunt ligt dat hoort bij de marginale kolom. Dat is voor de interpretatie van plot A niet relevant maar wel voor plot B en C.

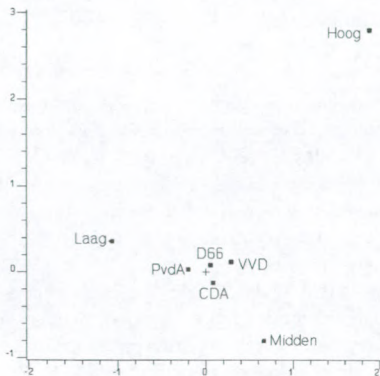
Het verschil in de ligging van rij- en kolompunten weerspiegelt het asymmetrische karakter van de plot: de perifere kolompunten vormen het 'interpretatie-kader' voor de centraal gelegen rijpunten. De plot sluit aan bij situaties waarin men zich afvraagt hoe groot de gelijkenis is tussen diverse categorieën onderzoekseenheden (te weten de rijcategorieën) wat betreft hun frequentieverdeling over een aantal kenmerken (te weten de kolomcategorieën). De interpretatiemogelijkheden van plot A zijn:

- a De spreiding van de rijpunten. Grote spreiding betekent dat de rijpunten niet op een kluitje dicht rond de oorsprong zijn gelegen, zoals in figuur 3, maar meer in de richting van de kolompunten in de periferie. We kunnen grote spreiding interpreteren als een sterke samenhang tussen rij- en kolomvariabele. Hoe sterker de samenhang, des te meer de rijpunten de perifere posities van de kolompunten naderen. De samenhang is maximaal wanneer in de kruistabel in iedere rij slechts één cel gevuld is en de overige celfrequenties nul zijn. Iedere rijcategorie correspondeert in dat geval met precies één kolomcategorie. In de plot vinden we dan de rijpunten op exact dezelfde perifere plaatsen als de kolompunten waar ze mee corresponderen. In het andere uiterste geval waarin er geen enkele samenhang is en waarin ieder rijprofiel gelijk is aan het marginale profiel, hebben de rijpunten geen enkele spreiding. Ze liggen dan alle in de oorsprong, hetgeen in dat geval ook de kolompunten doen.

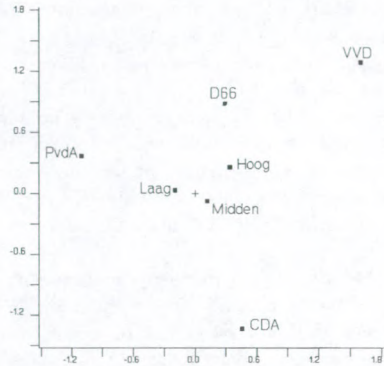
Uiteraard is er in de praktijk bijna nooit sprake van een dergelijke extreme situatie. Meestal corresponderen rijcategorieën met sommige kolomcategorieën wat meer (in de betekenis van: sterke oververtegenwoordiging of positieve v_{ij}) en met andere wat minder (negatieve v_{ij}). Vaak lijkt de plot, qua globale ligging van rij- en kolompunten, op die in figuur 3. De kolompunten liggen in de periferie en geven de extreme posities weer waar de rijpunten zich zouden bevinden wanneer de samenhang maximaal zou zijn. De rijpunten bevinden zich meer centraal in de buurt van de oorsprong en ze liggen, vanuit de oorsprong gezien, in de richting van die kolompunten waarmee de oververtegenwoordiging het sterkst is. De diverse rijprofielen moeten sterk van elkaar afwijken, wil er een redelijke spreiding van rijpunten in de plot te zien zijn.

- b Afstand van een rijpunt tot de oorsprong. Deze is gedefinieerd als de chikwadraat-afstand van de betreffende rij tot de marginale rij. Een grote afstand staat voor een grote gewogen afwijking van het marginale rijprofiel. De afstand is minimaal, te weten nul, wanneer het rijpunt in de oorsprong is gelegen. In dat geval komt het rijprofiel overeen met het marginale rijprofiel. In figuur 3 zien we dat, zoals eerder afgeleid uit formule (4), D'66 het dichtst bij de oorsprong ligt. Het rijprofiel van VVD wijkt het meest af van het marginale rijprofiel. Van alle partijen ligt dit punt dan ook het verst van de oorsprong verwijderd.
- c Afstand tussen twee rijpunten. Deze is gedefinieerd als chikwadraat-afstand tussen twee rijprofielen. Hoe groter de afstand, hoe meer twee rijprofielen van elkaar afwijken, in gewogen zin. Liggen de rijpunten op dezelfde plaats dan zijn de beide rijprofielen aan

Figuur 3: plot A van tabel 1



Figuur 4: plot B van tabel 1



elkaar gelijk. In figuur 3 zien we dat, qua gewogen rijprofiel, de VVD en de PvdA het meest van elkaar verschillen, terwijl PvdA en D'66 meer op elkaar lijken.

d De mate van over/ondervertegenwoordiging. Voor een bepaalde rij-kolomcombinatie is v_{ij} uit de plot af te leiden op de wijze zoals eerder is beschreven. Zoals gezegd, is daarbij de richting van een rij- en een kolompunt gezien vanuit de oorsprong van belang. In figuur 3 zien we dat PvdA en Laag in dezelfde richting liggen. Dat geldt ook voor CDA en Midden en voor VVD en Hoog. In deze drie combinaties is dan ook sprake van een oververtegenwoordiging. In de combinatie VVD-Hoog is de oververtegenwoordiging het sterkst. Een ondervertegenwoordiging vinden we bij combinaties die in tegenstelde richting van de oorsprong liggen, te weten CDA-Laag, VVD-Laag en PvdA-Hoog.

e Richting van twee rijpunten vanuit de oorsprong gezien. Als twee rijpunten in dezelfde richting liggen, betekent dit dat beide rijcategorieën hetzelfde patroon van over- en ondervertegenwoordiging bezitten. Hiermee bedoelen we dat voor elke kolom geldt: als in de éne rij een oververtegenwoordiging voorkomt, dan is dat ook in de andere rij het geval; komt in de éne rij een ondervertegenwoordiging voor, dan geldt dat ook voor de andere rij. Kijken we alleen naar het positief/negatief zijn van de v_{ij} 's in alle cellen in twee rijen (en niet naar de grootte) dan vertonen deze rijen een maximale overeenkomst. Liggen daarentegen twee rijpunten precies in tegengestelde richting van de oorsprong, dan geldt voor elke kolom: een oververtegenwoordiging in de éne rij gaat gepaard met een ondervertegenwoordiging in de andere rij. Enkel lettend op het teken van de v_{ij} 's kunnen we nu stellen dat beide rijen maximaal van elkaar verschillen.

In de praktijk liggen twee rijpunten meestal niet in exact dezelfde richting maar 'min of meer'. Vaak zien we dat beide rijcategorieën dan niet precies hetzelfde patroon van over- en ondervertegenwoordiging bezitten, maar sterk op elkaar lijkende patronen. De tekens van de twee v_{ij} 's zijn dan in de meeste kolommen gelijk, maar niet in alle. In figuur 3 zien we dat VVD en D'66 bijna in dezelfde richting liggen. Uit tabel 2 leiden we af dat in rij VVD de tekens van v_{ij} voor Laag, Midden en Hoog respectievelijk -, +

en + zijn. Voor D'66 is het patroon: - - +. De twee patronen verschillen alleen in de categorie Midden. Kijken we daarentegen naar PvdA en VVD, dan zien we dat die nagenoeg in tegengestelde richting liggen. Dit betekent dat de patronen van PvdA en VVD, zoals afleidbaar uit tabel 2, nagenoeg tegengesteld moeten zijn. Dat is inderdaad het geval: het patroon van PvdA is + - - en dat van VVD - + +.

f Richting van twee kolompunten vanuit de oorsprong gezien. Wat onder punt e is gezegd voor twee rijpunten geldt ook voor twee kolompunten. In figuur 3 liggen Midden en Laag bijna in tegengestelde richting. Dit betekent dat de twee patronen van over- en ondervertegenwoordiging bijna tegengesteld zijn. Uit tabel 2 leiden we af dat het patroon van kolom Midden gelijk is aan '- + + -' (van boven naar beneden). Voor kolom Laag is dat '+ - - -'. Op de laatste rij D'66 na zijn beide patronen tegengesteld. Het patroon in kolom Hoog, te weten '- - + +', vertoont op twee plaatsen een overeenkomst met dat van Midden. In de plot zien we dan ook dat Midden en Hoog meer in dezelfde richting liggen dan Midden en Laag.

g De afstanden van één rijpunt tot alle kolompunten. We hebben aan de hand van overgangsformule (6c) geconstateerd dat de afstanden van één rijpunt i tot alle kolompunten gerelateerd zijn aan de rijproporties in het rijprofiel van rij i: hoe groter de rijproportie des te kleiner meestal de afstand is tot een kolompunt. Kijken we in figuur 3 naar de PvdA dan zien we dat van de drie kolompunten Laag daar het dichtste bij ligt, Midden wat verder weg ligt en Hoog het verst verwijderd is. Deze volgorde in nabijheid komt overeen met de volgorde van groot naar klein van de rijproporties in tabel 2 rij PvdA. Ook voor de andere partijen geldt dat de twee volgorden van inkomklassen, de volgorde qua nabijheid in de plot en de volgorde qua grootte van rijproporties in tabel 2, met elkaar overeenkomen.

De gelijkheid van de beide ordeningen van kolommen gaat helaas niet altijd op: als elke kolom in de kruistabel evenveel onderzoekseenheden bevat en bovendien de totale samenhang in de tabel in niet meer dan twee dimensies is op te splitsen, alleen dan geldt voor elk tweetal kolommen dat de kolom met de grotere rijproportie in rij i tevens de kolom is met de kleinste afstand in de plot tot rijpunt i. Er is echter zelden aan beide voorwaarden voldaan. Bij het toepassen van deze interpretatiemogelijkheid moeten we daarom wat voorzichtig zijn: inspectie van de kruistabel (c.q. de rijprofielen) is aan te bevelen.

h De afstanden van één bepaald kolompunt tot alle rijpunten. Ook deze afstanden zijn te relateren aan rijproporties, en wel aan de rijproporties in de betreffende kolom. In het algemeen geldt: hoe kleiner de afstand des te groter de proportie. Ordenen we bijvoorbeeld de rijpunten in figuur 3 van meest tot minst nabij punt Laag, dan geeft dit de volgorde: PvdA, D'66, CDA en VVD. Ordenen we vervolgens in tabel 2 binnen kolom Laag de partijen van grootste proportie naar kleinste, dan krijgen we precies dezelfde volgorde. Voor inkomenscategorie Midden zijn de beide volgorden eveneens aan elkaar gelijk, te weten CDA, VVD, D'66 en PvdA. Voor Hoog tenslotte zijn beide volgorden gelijk aan VVD, D'66, CDA en PvdA.

Deze koppeling tussen afstanden tot rijpunten en rijproporties binnen een kolom, gaat net als die van punt g niet altijd op, maar er treden minder vaak schendingen op dan bij die van punt g. Greenacre (1989) maakt daarom wel expliciet melding van deze

interpretatiemogelijkheid en niet van die bij punt g. Het lijkt ons verstandig om ook bij het hanteren van deze interpretatiemogelijkheid voor alle zekerheid de kruistabel te inspecteren ter verificatie.

Samenvattend kunnen we stellen dat plot A veel informatie verschaft. Een niet zelden optredend nadeel van de plot is de geringe spreiding van de rijpunten, hetgeen de leesbaarheid bemoeilijkt. Wanneer de samenhang in de tabel niet erg groot is, zal de waarde van de canonische correlatie van dimensie 1 tamelijk dicht bij nul liggen. Als deze correlatie kleiner is dan laten we zeggen 0.25 betekent dit op grond van (6a) en (6b) dat de coördinaten van de rijen op de eerste dimensie ruwweg minstens $1/0.25$ ofwel 4 maal zo dicht bij nul liggen dan de coördinaten van de kolommen doen. Plot A lijkt ons in dat geval niet meer goed leesbaar. De plots C en D bieden dan, zoals zal blijken, een goed alternatief.

4 Plot B: asymmetrisch, kolompunten binnen rijpunten

Plot B verschilt van plot A in die zin, dat de rijen en kolommen van rol gewisseld hebben. Bij veel onderzoeksproblemen zullen we zowel de gelijkenis van rijen qua rijprofiel, als de gelijkenis van kolommen qua kolomprofiel willen beschrijven. In dat geval zijn we behalve in plot A, ook in plot B geïnteresseerd.

De formules voor de relatie tussen coördinaten en standaardscores alsmede de overgang formule voor de berekening van kolomcoördinaten uit rijcoördinaten zien er bij plot B als volgt uit:

$$R_{id} = r_{id} \quad (7a) \qquad K_{jd} = \lambda_d k_{jd} \quad (7b) \qquad K_{jd} = \sum_{i=1}^I p_{ij} R_{id} \quad (7c)$$

Hierin staat I voor het aantal rijen in de tabel; p_{ij} staat voor de proportie eenheden binnen kolom j in cel (i,j) , ofwel voor de frequentie in cel (i,j) gedeeld door de marginale frequentie van kolom j . De drie formules behoeven geen verdere toelichting: de interpretatie is identiek aan die van formules (6a), (6b) en (6c) bij plot A, zij het dat rijen en kolommen verwisseld zijn. Voor tabel 1 presenteren we plot B in figuur 4.

Over de interpretatiemogelijkheden van de plot kunnen we eveneens kort zijn. Alles wat bij plot A is gezegd over de rijen geldt hier voor de kolommen en omgekeerd. Kijken we naar de chikwadraat-afstanden tussen de kolompunten, die zich nu in het centrum van de plot bevinden, dan zien we dat de inkomenscategorieën Laag en Midden het dichtst bij elkaar liggen: het kolomprofiel van Midden lijkt, in gewogen zin, meer op dat van Laag dan op dat van Hoog. Ook zien we dat Midden en Laag vanuit de oorsprong gezien nageenog in tegenstelde richting liggen: qua patroon van over- en ondervertegenwoordiging lijkt Midden meer op Hoog dan op Laag. Tenslotte vermelden we nog dat van de kolompunten het punt Hoog het verst van de oorsprong verwijderd is: het kolomprofiel van Hoog wijkt gewogen het sterkst af van het marginale kolomprofiel.

5 Plot C: symmetrisch, rijpunten uit plot A, kolompunten uit plot B

De formules (8a) en (8b) laten zien hoe de coördinaten van rijen en kolommen gerelateerd zijn aan hun respectievelijke standaardscores; overgang formule (8c) geeft aan hoe de rijcoördinaten berekend kunnen worden uit de kolomcoördinaten en (8d) beschrijft het omgekeerde.

$$R_{id} = \lambda_d r_{id} \quad (8a) \quad K_{jd} = \lambda_d k_{jd} \quad (8b)$$

$$R_{id} = \lambda_d^{-1} \sum_{j=1}^I p_{j|i} K_{jd} \quad (8c) \quad K_{jd} = \lambda_d^{-1} \sum_{i=1}^I p_{i|j} R_{id} \quad (8d)$$

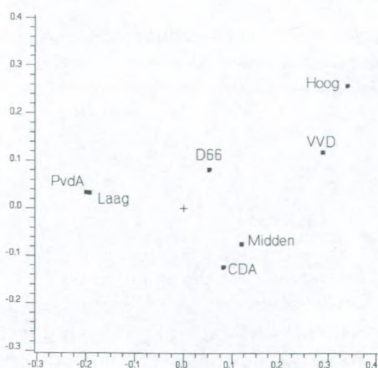
Formule (8a) is gelijk aan (6a) en (8b) gelijk aan (7b): wanneer we de rijpunten uit plot A en de kolompunten uit plot B afbeelden in één assenstelsel krijgen we plot C, weergegeven in figuur 5. Omdat in deze plot de rij- en kolomcoördinaten R_{id} en K_{jd} berekend worden door een gelijke 'krimping' van λ_d toe te passen op de rij- en kolom-standaardscores r_{id} en k_{jd} , is het probleem van de ongelijke spreiding van rij- en kolompunten verdwenen.

Uit (8c) blijkt dat ook hier de coördinaten K_{jd} van alle kolompunten een bijdrage leveren aan de coördinaat R_{id} van rijpunt i op dimensie d , en dat aan de bijdragen van kolommen waarvoor $p_{j|i}$ relatief hoog is meer gewicht wordt toegekend. Bij plot A konden we aan de hand van de overgang formule laten zien dat er een verband bestond tussen de afstand van kolompunt j tot rijpunt i enerzijds en de aantrekkingskracht $p_{j|i}$ van dat kolompunt j op rijpunt i anderzijds. Bij plot C gaat dit verband niet meer op. Wel is aan te tonen dat een rijpunt en een kolompunt samenvallen wanneer zij elkaar maximaal aantrekken in die zin dat zowel $p_{j|i} = 1$ als $p_{i|j} = 1$. Het komt echter ook voor dat een rij- en een kolompunt bijna samenvallen terwijl er helemaal niet van een extreem grote aantrekkingskracht sprake is. Zo is in figuur 5 te zien dat PvdA en inkomenscategorie Laag bijna op dezelfde plaats liggen, terwijl zowel $p_{PvdA|Laag}$ als $p_{Laag|PvdA}$ bij lange na niet de maximale waarde van 1 bereiken. Kortom, er bestaat bij deze plot geen eenvoudige relatie zoals bij plot A tussen de afstanden van een rijpunt tot alle kolompunten en de grootte van rijproporties.

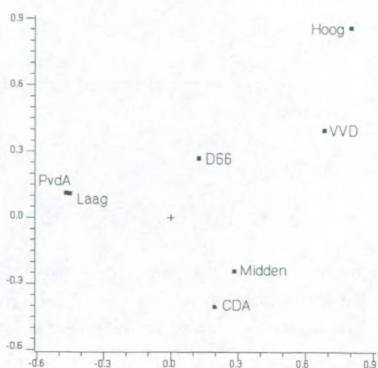
Plot C verschaft zowel informatie over de gelijkheid van rijen op basis van hun rijprofiel als over de gelijkheid van kolommen op basis van hun kolomprofiel. Daarom sluit de plot aan bij een symmetrische vraagstelling waarbij men in beide zaken geïnteresseerd is. De plot is echter ook bruikbaar bij een asymmetrisch probleem, waarbij men bijvoorbeeld enkel de gelijkheid van de rijcategorieën wilt onderzoeken. De interpretatiemogelijkheden komen deels overeen met die van plot A en B, omdat de puntenwolken van rij- en kolompunten eenvoudig een uitvergroting zijn van die in A en B. We noemen:

- a De afstand van een rij- c.q. kolompunt tot de oorsprong. Deze kunnen we interpreteren als de chikwadraat-afstand tot het marginale rij- c.q. kolomprofiel.
- b De afstand tussen twee rij- c.q. kolompunten. Ook deze afstand is te interpreteren als chikwadraat-afstand tussen de twee rij- c.q. kolomprofielen.
- c De over- en ondervertegenwoordiging. In tegenstelling tot plot A en B is het in plot C niet mogelijk de waarde van v_{ij} exact uit de plot af te leiden en datzelfde geldt voor het

Figuur 5: plot C van tabel 1



Figuur 6: plot D van tabel 1



teken van v_{ij} . Globaal geldt dat 'in dezelfde richting' duidt op oververtegenwoordiging, en 'in tegengestelde richting' op ondervertegenwoordiging.

d De richting van twee rij- c.q. kolompunten vanuit de oorsprong. Dezelfde richting staat voor hetzelfde patroon van over/ondervertegenwoordiging. We moeten hier wel bedenken (zie c) dat het teken van v_{ij} zoals afleidbaar uit de plot niet per se gelijk hoeft te zijn aan het werkelijke teken.

e Het belang van dimensie 1 in verhouding tot het belang van dimensie 2. De hoogte van de gereproduceerde chikwadratwaarden van beide dimensies is gerelateerd aan de spreiding van de puntenwolken van rij- en kolompunten: is $\chi^2_{(1)}$ veel groter dan $\chi^2_{(2)}$ dan is in het algemeen de spreiding van de rij- en kolompunten in de richting van dimensie 1 veel groter dan in de richting van dimensie 2. De spreiding in beide richtingen, horizontaal en verticaal, weerspiegelt beter dan bij de andere plots het relatieve belang van de twee dimensies. Voor alle duidelijkheid zij vermeld dat de spreiding in de plot als een ruwe indicatie voor het relatieve belang van beide dimensies moet worden opgevat. Voor een exacte indicatie van het relatieve belang dienen we de bijdragen $\chi^2_{(1)}$ en $\chi^2_{(2)}$ aan de totale χ^2 waarde van de tabel met elkaar te vergelijken, dan wel de inerties λ_1^2 en λ_2^2 . Deze behoren tot de standaard output van de vier te bespreken software pakketten.

Een eerste voordeel van plot C boven de eerder besproken varianten heeft te maken met de spreiding van de punten in de plot. Omdat de punten die bij A en B in de periferie liggen buiten beschouwing blijven, wordt de totale ruimte van de plot benut voor de punten die bij A en B in het centrum liggen. De spreiding van deze punten is daarom groter dan bij plot A en B en samenklontering rond de oorsprong komt niet voor. Tevens zegt de spreiding in horizontale en verticale richting ons vaak iets over het relatieve belang van de twee dimensies. Een tweede voordeel is het feit dat we ons bij een symmetrische vraagstelling tot slechts één enkele plot kunnen beperken.

Voor deze winstpunten moeten we echter een prijs betalen. Allereerst verliezen we de interpretatiemogelijkheden g en h van plot A: de afstanden van één rijpunt tot alle kolompunten en de afstanden van één kolompunt tot alle rijpunten zeggen niets meer over de hoogte van de rijproporties binnen één rij respectievelijk binnen één kolom. Een tweede nadeel is dat de mate van over/ondervertegenwoordiging niet exact is af te leiden uit de plot.

Vanwege de fraaie punten-spreiding die deze plot laat zien moeten we enigszins op onze hoede zijn bij de interpretatie. De optimale spreiding suggereert altijd een behoorlijke mate van samenhang tussen rijen en kolommen, ook wanneer die in werkelijkheid zeer gering is. Alvorens de plot te gaan interpreteren doen we er daarom goed aan de tabel te inspecteren om er zeker van te zijn dat de samenhang groot genoeg is om een nauwgezette analyse ervan te rechtvaardigen. Hoe groot 'voldoende groot' is laten we hier even in het midden: de ene onderzoeker zal vooral willen letten op de relevantie van verschillen tussen rij- of kolomprofielen in de tabel, en de ander zal vooral kijken naar het al of significant zijn van de algehele samenhang bijvoorbeeld getoetst met Pearson's chikwadraat. Wij willen hier enkel ervoor waarschuwen dat duidelijk 'zichtbare' samenhangen in plot C kunnen berusten op zeer minieme of niet relevant geachte profiel-verschillen in de tabel, terwijl plot C ons daaromtrent geheel in het ongewisse laat. Bij plot A en B daarentegen waarschuwt de geringe spreiding van de centraal gelegen punten ons voor het feit dat de profielverschillen klein zijn.

Tenslotte willen wij expliciet beklemtonen dat bij plot C niet de afstand van één rijpunt tot één kolompunt geïnterpreteerd mag worden. Bij deze plot zouden we in de verleiding kunnen komen dat te doen, omdat door de goede spreiding een rij- en kolompunt veel dicht bij elkaar kunnen liggen dan bij plot A en B. Het interpreteren van de afstand van een rijpunt tot een kolompunt is echter alleen toegestaan bij plot A (zie interpretatiemogelijkheid g en h) en B, wanneer we bij die twee plots de rij-tot-kolom afstand vergelijken met de afstanden van het rijpunt tot alle andere kolompunten of met de afstanden van het kolompunt tot alle andere rijpunten. Bij plot C bestaan deze interpretatiemogelijkheden echter niet. We lichten een en ander nog eens toe met een voorbeeld. Uit de nabijheid van Hoog en VVD in figuur 5 mogen we geenszins concluderen dat mensen met een hoog inkomen voor het merendeel VVD stemmen, en ook niet dat VVD-stemmers voor het merendeel een hoog inkomen hebben. Beide conclusies zouden ook niet passen bij onze data: kijken we in tabel 1 naar kolom Hoog dan zien we dat er onder de 48 mensen met een hoog inkomen nagenoeg evenveel PvdA-stemmers voorkomen als VVD-stemmers; kijken we in tabel 2 naar rij VVD dan zien we dat de meerderheid van de VVD-stemmers niet te vinden zijn in inkomensklasse Hoog maar in Midden.

Valt er dan helemaal niets af te leiden uit de nabijheid van een rij- en kolompunt bij plot C? Wanneer die nabijheid gepaard met een grote afstand tot de oorsprong van de plot is er sprake van oververtegenwoordiging. In figuur 5 liggen VVD en Hoog tamelijk ver van de oorsprong: onder VVD-stemmers is het percentage hoge inkomens (15.2%) groot in verhouding tot het percentage hoge inkomens onder alle 495 mensen (8.1%), te weten 1.88 maal zo groot. Dat feit levert voor cel VVD-Hoog verreweg de grootste van alle oververtegenwoordigingen v_{ij} in de tabel, te weten $15.2/8.1 - 1 = 0.88$. Wat we wel mogen concluderen is 'onder VVD-stemmers komen procentueel veel meer hoge inkomens voor dan onder stemmers op de andere partijen'.

Al met al is het eenvoudigste om te onthouden dat het bij plot C enkel gaat om de mate van over-/ondervertegenwoordiging wanneer we iets willen zeggen over de relatie tussen één bepaalde rij en één bepaalde kolom. Het interpreteren van een rij-tot-kolom afstand kunnen we beter helemaal vergeten en reserveren voor plots A en B.

6 Plot D: symmetrisch

Formules (9a) en (9b) laten zien hoe bij plot D de coördinaten van rijen en kolommen gerelateerd zijn aan hun respectievelijke standaardscores. De overgangsformules geven we niet: ze zijn gelijk aan die bij plot C.

$$R_{id} = \sqrt{\lambda_d} \ r_{id} \quad (9a) \quad K_{jd} = \sqrt{\lambda_d} \ k_{jd} \quad (9b)$$

Net als bij plot C worden ook bij plot D de rij- en kolomcoördinaten R_{id} en K_{jd} berekend door een gelijke 'krimping', in dit geval van $\sqrt{\lambda_d}$, toe te passen op de rij- en kolom-standaardscores r_{id} en k_{jd} . De ongelijke spreiding van rij- en kolompunten, kenmerkend voor plot A en B, treffen we daarom bij plot D niet aan, zoals we die ook bij C niet aantreffen.

Men kan plot D net als C gebruiken bij symmetrische en asymmetrische vraagstellingen. Figuur 6 geeft plot D voor tabel 1. Vergelijken we figuur 5 en 6, dan zien we dat plot D veel lijkt op C. In het algemeen kunnen we het verschil zo voorstellen dat bij plot D de tweede dimensie enigszins uitgerekt is, om precies te zijn met een factor $\sqrt{(\lambda_1/\lambda_2)}$, waarin λ_1 en λ_2 de canonische correlaties van de eerste en tweede dimensie zijn.

Deze uitrekking brengt ons op een belangrijk verschil van deze plot met de drie andere: de afstanden van rijpunten (c.q. kolompunten) tot elkaar en tot de oorsprong zijn geen chikwadraat-afstanden meer. De interpretatiemogelijkheid van de chikwadraat-afstanden verliezen we dan ook. Verder wordt door de uitrekking van dimensie 2 het belang van deze dimensie wat aangedikt: het belang van dimensie 1 in verhouding tot dat van dimensie 2 kunnen we niet goed uit de puntenspreiding in de plot afleiden, in tegenstelling tot plot C. Wat overblijft zijn de interpretaties gericht op over-/ondervertegenwoordiging. We noemen nog eens kort:

- a De mate van over-/ondervertegenwoordiging v_{ij} is op de eerder beschreven wijze uit plot D af te leiden. Door het verschijnsel van samenklontering is deze interpretatiemogelijkheid bij plots A en B niet altijd goed mogelijk. Bij plot D doet dit probleem zich niet voor, waardoor deze plot bij uitstek geschikt is om over-/ondervertegenwoordigingen weer te geven.
- b Richting van twee rijpunten (c.q. kolompunten) vanuit de oorsprong. Twee rijpunten (c.q. kolompunten) die in dezelfde richting liggen, zijn gelijk wat betreft het patroon van over- en ondervertegenwoordiging over de kolomcategorieën (c.q. rijcategorieën). Liggen ze in tegengestelde richting dan zijn ze in dit opzicht elkaars tegenpool.

Plot D heeft van alle plots de minste interpretatiemogelijkheden. De richting vanuit de oorsprong, ofwel het kijken naar over- en ondervertegenwoordigingen, dat is waar het bij

deze plot om te doen is. Wanneer we de plot op dit aspect vergelijken met plots A en B dan is hij te prefereren vanwege de goede spreiding van zowel rij- als kolompunten. Willen we echter chikwadraat-afstanden interpreteren, dan kunnen we beter een van de andere plots gebruiken.

De plot kent dezelfde 'nadelen' van optimale spreiding die we bij plot C al tegenkwamen. Bij totaal niet relevante/significante verschillen tussen de diverse rij- c.q. kolomprofielen loopt de argeloze gebruiker het risico de toch duidelijk zichtbare over/ondervertegenwoordigingen in de plot als belangrijk te gaan bestempelen. Voorzichtigheid is hier weer op zijn plaats. Verder geldt ook voor deze plot dat geen rij-tot-kolom afstanden geïnterpreteerd mogen worden, zoals bij plot A en B, maar enkel over/ondervertegenwoordigingen.

7 Conclusies voor de keuze tussen plots.

Alvorens we besluiten met CA een plot te maken doen we er goed aan na te gaan of de mate van samenhang in de tabel wel zo groot is dat het zinvol is die samenhang nauwkeurig te analyseren en te beschrijven. Behalve de geringe spreiding van de centraal gelegen punten in plot A en B geven de CA-plots geen verdere indicatie van de mate van samenhang in een tabel. Het spreekt voor zich dat het raadzaam is alvorens plots te gaan interpreteren eerst de tabel met 'huis-, tuin- en keukengereedschap' onder de loupe te nemen, zoals de rij- en kolomprofielen, de marginale profielen en het al of niet significant zijn van de totale samenhang in de tabel. Pas wanneer de geconstateerde profiel-verschillen relevant zijn en/of de totale samenhang in de tabel significant is lijkt ons een genuanceerde beschrijving van die samenhang aan de hand van een CA-plot gerechtvaardigd.

Als het gaat om de keuze van de juiste plot speelt om te beginnen het type vraagstelling dat we hebben een rol. Bij een asymmetrische vraagstelling is het aan te raden te beginnen met plot A of B. Deze zijn qua interpretatiemogelijkheden immers het meest compleet: interpretaties van zowel chikwadraat-afstanden als over/ondervertegenwoordigingen zijn mogelijk. Bovendien zijn rij-tot-kolom afstanden gerelateerd aan rijproporties c.q. kolomproporties. De neiging tot samenklonteren van punten rond de oorsprong vormt in de praktijk echter vaak een probleem, vooral wanneer het aantal punten toeneemt en de samenhang in de tabel niet al te sterk is. Weergave via grafische software leidt dan soms toch tot een acceptabele plot. Wanneer de waarde van λ_1 , ofwel de canonische correlatie van dimensie 1, echter kleiner is dan 0.25 lijken ons plot C en D meestal meer geschikt dan A of B. De keuze voor C of D hangt daarbij af van de vraag wat we zo goed mogelijk in de plot willen weergeven. Richten we ons op het wel of niet 'bij elkaar horen' van specifieke rijen en kolommen in de zin van over/ondervertegenwoordigingen, dan is plot D de beste keuze. Leggen we meer de nadruk op de gelijkenis tussen rijen of kolommen onderling, op basis van de rij- of kolomprofielen, dan komt plot C het meest in aanmerking.

Is de vraagstelling symmetrisch, dan zouden we enerzijds kunnen kiezen voor weergave van zowel plot A als B. Anderzijds zouden we voor één van de plots C of D kunnen kiezen op basis van zojuist genoemde overwegingen.

Tenslotte willen nog eens herhalen wat we op verschillende plaatsen al opmerkten: het is niet verstandig conclusies louter en alleen te baseren op wat een CA-plot laat zien. Zelfs wanneer niet meer dan twee dimensies nodig zijn om de volledige samenhang in de tabel weer te geven, kunnen de plots hier en daar een onzuiver beeld van de werkelijkheid voorspiegelen. Zo kunnen bij plot A en B de interpretaties van rij-tot-kolom afstanden tot verkeerde conclusies leiden. Bij plot C kan het interpreteren van over/ondervertegenwoordigingen conclusies opleveren die strijdig zijn met de data in de kruistabel. Plot D heeft dergelijke vertekeningen niet maar heeft dan ook de minste interpretatiemogelijkheden.

In de praktijk zien we echter dat CA juist wordt toegepast in situaties waarbij de samenhang tussen rij- en kolomvariabele zo complex is dat deze niet in twee dimensies is te vatten. Zo hebben bijvoorbeeld tabellen met vier of meer rijen en kolommen meestal drie of meer dimensies nodig. Doordat we ons in de plots beperken tot de eerste twee dimensies zien we slechts een beperkte weergave van de volledige samenhang. Het gevolg is dat alle interpretatiemogelijkheden die we eerder bij de plots apart noemden tot conclusies kunnen leiden die niet overeenkomen met de data in de kruistabel. Zo kan wat in drie dimensies een oververtegenwoordiging is, in twee dimensies uitzien als een ondervertegenwoordiging. Ook kan de chikwadraatafstand van een rij in twee dimensies een minieme fractie zijn van dezelfde afstand in drie dimensies, waardoor we ten onrechte zouden concluderen dat de betreffende rij veel lijkt op de marginale rij. Vanwege deze vertekeningen ten gevolge van de beperking tot twee dimensies adviseren we gebruikers niet blind te varen op de plot, maar specifieke interpretaties altijd te verifiëren aan de hand van de kruistabel zelf.

In de paragraaf die volgt bespreken we een aantal kengetallen die binnen CA bestaan om de kwaliteit van de weergave in de diverse dimensies te kunnen evalueren.

8 Kengetallen betreffende de weergave-kwaliteit.

Zoals gezegd gaat de weergave in twee dimensies doorgaans gepaard met een zeker verlies. De interpretatiemogelijkheid van de dimensies, heeft onder dit verlies niet te lijden. Ook wanneer de eerste twee dimensies tezamen bijvoorbeeld slechts 40% van de totale geobserveerde chikwadraatwaarde reproduceren, geven zij nog steeds de twee belangrijkste richtingen aan waarlangs de totale samenhang in de tabel zich voltrekt.

De beperking tot twee dimensies kan echter wel afbreuk doen aan de andere interpretatiemogelijkheden. Zo kunnen de in de plot weergegeven chikwadraat-afstanden kleiner zijn dan de geobserveerde afstanden, die direct uit de kruistabel zijn te berekenen. En de uit de plot afleidbare maten v_{ij} van over/ondervertegenwoordiging kunnen groter of kleiner zijn dan de geobserveerde, en zelfs tekenwisselingen kunnen voorkomen. Alvorens gedetailleerde conclusies te trekken uit de plot doen we er daarom goed aan eerst de kwaliteit van de weergave te bestuderen. Zo zouden we aan de hand van de rijprofielen eenvoudig kunnen nagaan of oververtegenwoordigingen in de tabel gepaard gaan met plaatsingen in dezelfde richting (t.o.v. de oorsprong) in de plot; en andersom of ondervertegenwoordigingen in de tabel als plaatsingen in tegengestelde richting te vinden zijn in de plot. De vraag naar de correcte weergave van chikwadraatafstanden is echter niet zo eenvoudig te beantwoorden op basis van de geobserveerde rij- of kolomprofielen.

Naast de oorspronkelijke tabel staan ons bij CA echter nog enkele grootheden ter beschikking op basis waarvan we de kwaliteit van de weergave kunnen beoordelen.

De meest globale grootheid is gebaseerd op de in twee dimensies gereproduceerde chikwadraat-waarde. We merkten eerder op dat de geobserveerde chikwadraat-waarde χ_{obs}^2 gelijk is aan $\sum_{i=1}^I f_i \delta_{i(\text{obs})}^2$, waarin $\delta_{i(\text{obs})}$ staat voor de geobserveerde, met formule (4) te berekenen, chikwadraat-afstand van rij i tot het marginale rijprofiel. We definiëren nu de in twee dimensies gereproduceerde chikwadraat-waarde χ_{rep}^2 als zijnde gelijk aan $\sum_{i=1}^I f_i \delta_{i(\text{rep})}^2$, waarin $\delta_{i(\text{rep})}$ staat voor de, in bijvoorbeeld plot A, weergegeven (chikwadraat) afstand van rijpunt i tot de oorsprong. De verhouding $\chi_{\text{rep}}^2 / \chi_{\text{obs}}^2$ is een globale indicatie voor de kwaliteit van de weergave in twee dimensies. Omdat de waarden χ_{obs}^2 / N en χ_{rep}^2 / N bij CA wel 'inerties' worden genoemd, respectievelijk de geobserveerde en gereproduceerde inertie, heet genoemde verhouding ook wel 'proportie weergegeven inertie'. In paragraaf 2 vermeldde we dat $\chi_{\text{obs}}^2 / N = \sum_{d=1}^D \lambda_d^2$. Evenzo is $\chi_{\text{rep}}^2 / N = \sum_{d=1}^D \lambda_d^2$ en daarom kunnen we $\chi_{\text{rep}}^2 / \chi_{\text{obs}}^2$ ook weergeven als $\sum_{d=1}^D \lambda_d^2 / \sum_{d=1}^D \lambda_d^2$, waaruit volgt dat het hier inderdaad een proportie betreft, begrensd door de waarden 0 en 1. Hoe meer de grootheid de waarde 1 nadert, des te hoger is de kwaliteit van de weergave in twee dimensies. Een 'hoge' waarde van bijvoorbeeld .90 zou kunnen optreden wanneer de geobserveerde chikwadraat-afstanden van alle rijen behoorlijk goed in de plot zijn weergegeven, en er een monotone relatie bestaat tussen geobserveerde en weergegeven afstanden. Deze waarde van .90 zou echter ook kunnen optreden wanneer de geobserveerde chikwadraat-afstanden van alle rijen op één na zo goed als perfect zijn weergegeven, terwijl voor één rij de geobserveerde en weergegeven afstand sterk van elkaar afwijken, waardoor van een monotone relatie tussen geobserveerde en weergegeven afstanden geen sprake is. Het is daarom raadzaam om ook bij hoge waarden van $\chi_{\text{rep}}^2 / \chi_{\text{obs}}^2$ de nu volgende grootheden te inspecteren.

De tweede grootheid is meer gedetailleerd. Zij is gebaseerd op de in twee dimensies gereproduceerde chikwadraat-afstand $\delta_{i(\text{rep})}$ die voor iedere rij en kolom is te bepalen. De verhouding $\delta_{i(\text{rep})}^2 / \delta_{i(\text{obs})}^2$, waarvoor geldt $0 < \delta_{i(\text{rep})}^2 / \delta_{i(\text{obs})}^2 \leq 1$, geeft aan welke proportie van de gekwadrateerde geobserveerde chikwadraat-afstand van rijpunt i wordt weergegeven in de plot. Omdat men $f \delta_{i(\text{obs})}^2$ en $f \delta_{i(\text{rep})}^2$ ook wel de geobserveerde resp. gereproduceerde inertie van rij i noemt, heet de verhouding $\delta_{i(\text{rep})}^2 / \delta_{i(\text{obs})}^2$ ook wel 'proportie gereproduceerde inertie van rij i '. Soms spreekt men kortweg over de 'kwaliteit' van de weergave van rijpunt i in twee dimensies. Bij een 'lage' kwaliteit van bijvoorbeeld .60 zullen we voorzichtig moeten zijn met conclusies ten aanzien van rij i . Blindvaren op de plot om te komen tot uitspraken als 'de chikwadraat-afstand tussen rij i en k is groter dan die tussen rij i en l ' en 'in cel (i, j) is sprake van een oververtegenwoordiging' is dan zeker niet mogelijk. We zullen in dat geval altijd terug moeten naar de geobserveerde tabel om ons te vergewissen van de juistheid van zulke beweringen. Bij rijen en kolommen met een 'hoge' kwaliteit van bijvoorbeeld .90 zal de overeenkomst tussen plot en tabel meestal groot zijn, maar blijft verificatie aanbevelenswaard.

De vraag naar de mate waarin chikwadraat-afstanden in de plot gerealiseerd worden, kunnen we ook stellen voor chikwadraat-afstanden tussen rijpunten of kolompunten onderling. Anders gezegd: we kunnen voor ieder tweetal rijen (of kolommen) i en j de verhouding $\delta_{ij(\text{rep})}^2 / \delta_{ij(\text{obs})}^2$ berekenen. Hierin staat $\delta_{ij(\text{obs})}$ voor de geobserveerde chikwadraat-

afstand van rij i tot rij j , te berekenen volgens formule (3), en $\delta_{ij(rep)}$ voor de in de plot weergegeven afstand. Er geldt weer $0 < \delta_{ij(rep)}^2 / \delta_{ij(obs)}^2 \leq 1$. De grootheid $\delta_{ij(rep)}^2 / \delta_{ij(obs)}^2$ stelt ons in staat de weergavekwaliteit van iedere 'onderlinge' rijafstand (of kolomafstand) te evalueren. Zo is het mogelijk dat de afstand van twee rijpunten tot de oorsprong redelijk goed in de plot is weergegeven, terwijl dat voor de onderlinge afstand in veel mindere mate het geval is. Jammer genoeg leveren de vier te bespreken software producten deze gedetailleerde informatie niet.

Dan rest nog de vraag naar de kwaliteit van de weergave van alle over/ondervertegenwoordigingen in de plot. Het antwoord op deze vraag kunnen we krijgen door iedere geobserveerde en met formule (5) te berekenen waarde $v_{ij(obs)}$ te vergelijken met de waarde $v_{ij(rep)}$, die op de wijze besproken onder figuur 1 uit de plot is af te leiden. Grote (absolute) verschillen en tegengesteldheid van teken duiden dan op slecht weergegeven over/ondervertegenwoordigingen in de plot. Helaas treffen we de waarden $v_{ij(obs)}$, $v_{ij(rep)}$, de absolute verschillen tussen beide (eventueel voorzien van een minteken wanneer $v_{ij(obs)}$ en $v_{ij(rep)}$ tegengesteld van teken zijn), of een andere afgeleide maat eveneens niet aan in de output van de vier software producten.

De vraag naar de kwaliteit van de weergave van over/ondervertegenwoordigingen is extra relevant bij plot C. Daarin worden over/ondervertegenwoordigingen immers ook bij een perfecte 'fit' in twee dimensies slechts globaal weergegeven. De reductie tot twee dimensies kan dan nog eens extra tot vertekeningen leiden. De $v_{ij(rep)}$ waarden bij plot C hebben als zodanig weinig betekenis. Wel is het wenselijk dat ze qua teken overeenkomen met de $v_{ij(obs)}$ waarden en dat de rangordening van alle $v_{ij(rep)}$ waarden geen al te grote afwijkingen vertoont met die van de $v_{ij(obs)}$ waarden.

9 Software

We geven hier voor drie software producten aan welke CA-plots er mee gemaakt kunnen worden, de commando's om ze te genereren, en de benamingen waaronder men in de output de diverse maten voor de weergavekwaliteit aantreft. We beperken ons tot de pakketten, SPSS, SAS en BMDP.

De procedure ANACOR in de module Categories van SPSS levert de plots A, B en D. We krijgen plot A met de subcommando's /NORMALIZATION RPRINCIPAL, plot B met /NORMALIZATION CPRINCIPAL en plot D met /NORMALIZATION CANONICAL. Bij alle versies dient men tevens het subcommando /PLOT JOINT te gebruiken. Plot D is de standaard-plot. Plot C kan niet rechtstreeks opgevraagd worden, maar is wel via een omweg te produceren, en wel door de rijcoördinaten van plot A en de kolomcoördinaten van plot B uit de output te selecteren en deze tezamen in een datafile op te slaan. Daarmee kan dan in tweede instantie met procedure PLOT alsnog een scatterplot worden gegenereerd. De maat $\chi_{rep}^2 / \chi_{obs}^2$ vinden we als 'cumulative proportion' in een tabel naast de 'singular value' en 'inertia' voor iedere dimensie. De kwaliteit van elk punt vinden we in de kolom 'total' van een tabel met titel 'contribution of dimensions to the inertia of each row (c.q. column) point'.

PROC CORRESP in SAS maakt alle vier de plots. We sturen plot A, B en C het eenvoudigst aan met de optie PROFILE= binnen het PROC CORRESP statement. Plot A roepen we aan via PROFILE=ROW, plot B met PROFILE=COLUMN en plot C met hetzij PROFILE=BOTH hetzij door de optie achterwege te laten, want plot C is de standaard. Plot D sturen we op een andere wijze aan. In plaats van PROFILE= gebruiken we nu de opties ROW=DAD1/2 en COL=DBD1/2. De maat $\chi_{rep}^2 / \chi_{obs}^2$ vinden we, vermenigvuldigd met 100, in kolom 'percents' van de tabel getiteld 'inertia and chi-square decomposition'. De percentages van de eerste twee dimensies moeten wel nog opgeteld worden. De kwaliteit per punt vinden we door optelling van de getallen in de kolommen 'dim1' en 'dim2' van de tabel met titel 'squared cosines for the row (c.q. column) points'.

Voor het maken van de plots moeten we nog wat extra werk doen. Allereerst moeten we via de optie OUT= in het PROC CORRESP statement de schaalwaarden van rijen en kolommen op de twee dimensies in een bestand zetten. Daarmee kunnen we in tweede instantie met PROC PLOT of PROC GPLOT de gewenste plot genereren.

Het programma CA in BMDP kent alleen de plots A, B en C. De PLOT paragraaf geeft standaard deze drie plots. Met het keyword NO ROWS onderdrukken we plot A, met NO COLS plot B en met NO BOTH plot C. De maat $\chi_{rep}^2 / \chi_{obs}^2$ staat, vermenigvuldigd met 100, in kolom 'cum %' in een tabel met 'eigenvalues' en '% of inertia' per 'axis'. De kwaliteit per punt staat in kolom 'qlt' van een tabel getiteld 'rows' (c.q. 'columns').

Literatuur

- * Arts, K., Hollander, E., Renckstorf, K., Verschuren, P. (1990). *Grootschalig Onderzoek naar Media-uitrusting, Media-exposure en Mediagebruik in Nederland 1989*. Instituut voor Toegepaste Sociale Wetenschappen, Nijmegen.
- * Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley.
- * Greenacre, M. (1981). Practical Correspondence Analysis. In: Barneff, V. (red.), *Interpreting Multivariate Data*. Wiley.
- * Greenacre, M., Hastie, T. (1987). The Geometric Interpretation of Correspondence Analysis. *Journal of the American Statistical Association*, 82, 437-447.
- * Greenacre, M. (1989). The Carroll-Green-Schaffer scaling in correspondence analysis: a theoretical and empirical appraisal. *Journal of Marketing Research*, 26, 358-365.
- * Israëls, A. (1987). *Eigenvalue techniques for qualitative data*. DSWO Press, Leiden.
- * Sikkel, D., Bethlehem, J. (1981). *Interpretatie van correspondentieanalyse-resultaten of wat stellen die plaatjes nu eigenlijk voor?*. Centraal Bureau voor de Statistiek, Voorburg.

ontvangen: 05-10-1995
geaccepteerd: 14-01-1998