verspreiden

1.0 Inleiding

In dit inleidende hoofdstuk zullen we allereerst een definitie proberen te geven van multivariate analyse (MVA). Het spreekt vanzelf dat we hierbij rekening moeten houden met al eerder door anderen gegeven omschrijvingen. We doen dit door middel van een korte inhoudsanalyse van de voorwoorden en inleidende hoofdstukken van enkele van de meest gebruikte boeken over MVA. Over de definitie blijkt een behoorlijke mate van overeenstemming te bestaan, hoewel dat voornamelijk komt omdat de meest gebruikelijke definitie zo weinig zeggend is. Uit de door de diverse auteurs gegeven specifikaties, omschrijvingen, voorbehouden, en geloofsbelijdenissen komen een aantal problemen en tegenstellingen naar voren die interessant genoeg zijn om nader op in te gaan. We geven in dit hoofdstuk tevens een eerste voorbeeld van korrespondentie-analyse, toegepast om de inhoudsopgaven van de besproken MVA boeken.

1.1 Inhoudsanalyse MVA-boeken

1.1.1 Roy (1957)

Roy pretendeert niet een volledig overzicht van MVA te geven. Hij beperkt zich over het algemeen tot multinormaal verdeelde variabelen, en meer in het bizonder tot het begrenzen van betrouwbaarheids-intervallen voor bepaalde klassen parametrische funkties. Roy noemt als zijn belangrijkste omissies faktoranalyse, klassifikatie, en analyse van variantiekomponenten. Voor onze doeleinden is het vooral van belang dat hij in hoofdstuk 15 van zijn boek niet-parametrische generalisaties van MVA beschrijft, die vooruitlopen op de moderne analyse van meer-dimensionale kruistabellen. "Despite all the mathematical elegance and comparative simplicity of 'normal variate' analysis of variance and multivariate analysis, one cannot help feeling that the non-parametric approach (whether of this variety or of other varieties) is far more realistic and physically meaningful, and is likely, in the future, to supplant, to a large extent, the existing techniques of 'normal variate' analysis of variance and multivariate analysis, including those discussed in the first fourteen chapters of this monograph." (lc, pag viii). Hoewel Roy nergens een eksplisiete definitie geeft, lijkt het alsof hij MVA opvat als een stelsel technieken om een beperkt aantal typen hypothesen omtrent de relaties tussen gekorreleerde variabelen te toetsen.

1.1.2 Kendall (1957, 1975)

Kendall geeft wel een definitie. "We may thus define multivariate analysis as the branch of statistical analysis which is concerned with the relationships of sets of dependent variates." (1957, pag 6).

Maar hij geeft tevens een aantal voorbeelden van typische MVA problemen, en zegt dat een opsomming van dit soort problemen dikwijls beter werkt als een definitie. Hij maakt in zijn boek het fundamentele onderscheid tussen de analyse van afhankelijkheid ("dependence") en de analyse van samenhang ("interdependence"). In het eerste geval onderzoeken we hoe een bepaalde groep variabelen afhangt van de variabelen in een tweede groep. De eerste groep bevat de zogenaamde "afhankelijke", de tweede groep de zogenaamde "onafhankelijke" variabelen. Er is een zekere asymmetrie in de analyse, of, anders geformuleerd, er wordt een richting van kausale beinvloeding verondersteld. Daarentegen worden de groepen variabelen bij de analyse van samenhang symmetrisch behandeld, er is geen onderscheid tussen afhankelijke en onafhankelijke variabelen. In de gemoderniseerde versie van het boek, uit 1975, is de inleiding aanzienlijk uitgebreid. De definitie van MVA wordt op pagina 1 in ietwat andere bewoordingen herhaald. Als belangrijkste doeleinden van MVA noemt Kendall hier:

a: strukturele vereenvoudiging,
b: klassifikatie,
c: groepering van variabelen,
d: analyse van afhankelijkheid,
e: analyse van samenhang,
f: hypothese konstruktie en hypothese toetsing.

Als belangrijkste problemen bij de verdere ontwikkeling van MVA noemt hij de volgende vijf.

a: In veel situaties in de praktijk kunnen we niet de gebruikelijke statistische assumpties maken. Er is geen sprake van een steekproef uit een populatie, of er is wel een steekproef maar die is niet willekeurig. "It is a mistake to try and force the treatment of such data into a classical probabilistic mould, even though some subjective judgement in treatment and interpretation may be involved in the analysis of the results." (1975, pag 4).

b: Hoewel de komputer noodzakelijk is bij gebruik van MVA zijn er verschillende gevaren verbonden aan het onkritisch gebruik van ingeblikte komputerprogrammaas.

c: Zelfs als we een willekeurige steekproef hebben, dan is het vaak onmogelijk om multivariate normaliteit aan te nemen. "Most theoretical work in multivariate statistics is based on the assumption that the parent population is multinormal, and its robustness under departures from normality is very often difficult to determine with any exactitude." (1975, pag 4).

d: Bij MVA zijn grafische voorstellingen van groot belang. Het is echter uiterst moeilijk om bevredigende grafische representaties te bedenken voor gegevens in meer dan twee dimensies.

e: Omdat er geen natuurlijke definitie van een orde relatie tussen punten in een meerdimensionele ruimte is, is er tot nu toe geen bevredigende vorm van niet-parametrische MVA gevonden, die vergelijkbaar is met de niet-parametrische univariate statistiek.

We onthouden uit Kendall's diskussie dat er bij MVA sprake is van skores van een aantal individuen op een aantal samenhangende variabelen. Maar de variabelen zijn niet noodzakelijk stochastisch, niet noodzakelijk kontinu meetbaar, en de individuen zijn niet noodzakelijk onafhankelijke en identiek verdeelde replikaties. Opmerkelijk is tenslotte dat Kendall, evenals Roy, een laatste hoofdstuk over gekategoriseerde multivariate gegevens opneemt. In geen van de overige boeken, die we in dit hoofdstuk bespreken, vinden we een dergelijk hoofdstuk. Als we de twee edities vergelijken vinden we, zeker in het inleidende hoofdstuk, een verschuiving van multivariate statistische analyse en multinormale analyse naar een meer algemene omschrijving van MVA.

1.1.3 Anderson (1958)

"In this book we shall concern ourselves with statistical analysis of data that consist of sets of measurements on a number of individuals or objects. ... The mathematical model on which analysis is based is a multivariate normal distribution or a combination of multivariate normal distributions." (lc, pag 1). Dat is duidelijke taal. Hoe verdedigt Anderson zijn keuze voor multinormale statistische analyse ? We noemen hier zijn belangrijkste argumenten, verderop in dit hoofdstuk zullen we ze stuk voor stuk uitvoeriger bespreken.

a: De multinormaalverdeling blijkt in de praktijk in veel gevallen een goede beschrijving van de verdeling van multivariate observaties te zijn.

b: De multinormaalverdeling volgt uit de meerdimensionele versies van de centrale grenswaarde stelling, en levert dus een goed model op als de observaties opgevat kunnen worden als de som van een groot aantal onafhankelijke kleine effekten.

c: De verdelingstheorie gebaseerd op de aanname van multinormaal verdeelde observaties is mathematisch relatief eenvoudig, en mede daardoor in de loop van de laatste 75 jaar vrij volledig uitgewerkt.

Volgens Anderson is MVA voornamelijk generalisatie van de bekende univariate inferentiele statistiek naar multinormaal verdeelde

observaties. Als gevolg van deze opvatting nemen typisch multivariate technieken als principale komponenten analyse (voortaan PCA) en kanonische analyse (voortaan CA) een bescheiden plaats in zijn boek in.

1.1.4 Cooley en Lohnes (1962, 1971).

De oudere versie van dit boek is een uitvoerige handleiding hoe de gebruikelijke MVA technieken op de komputer geimplementeerd moeten worden. Het is daardoor nauwelijks meer interessant. De definitie van MVA is dezelfde als die van Kendall. In het boek uit 1971 lezen we in het voorwoord dat Cooley en Lohnes inmiddels hun identiteit gevonden hebben: ze zijn nu data analytici. Vandaar de verandering in de titel van het boek. In 1962 waren ze natuurlijk ook al data analytici, alleen ze wisten het nog niet. Ze leerden het vanzelfsprekend van Tukey, wiens befaamde artikel ook in 1962 verscheen. "His gift to us was our professional identity. ... Tukey made our interest in multivariate heuristics rather than multivariate inference sound respectable." (1971, pag v). De definitie van MVA is hetzelfde gebleven, Cooley en Lohnes nemen ook Kendall's onderscheid tussen de analyse van afhankelijkheid en de analyse van samenhang over. Voor de volledigheid vermelden we hier even dat dit onderscheid al eerder te vinden is in een aantal artikelen die Bartlett rond 1950 publiceerde, het is dus niet helemaal juist om het voortdurend aan Kendall toe te schrijven. Op het onderscheid tussen data analyse en inferentiele statistiek, en op het artikel van Tukey, komen we verderop in dit hoofdstuk nog uitvoerig terug. Met name lijkt het ons interessant om na te gaan of de heuristische multivariate technieken sinds Tukey inderdaad respektabel zijn geworden.

1.1.5 Morrison (1967, 1976)

De definitie is ondertussen bijna standaard aan het worden. "Multivariate statistical analysis is concerned with data collected on several dimensions of the same individual." (1967, pag vii). Niettemin beperkt Morrison zich meer dan Kendall, omdat hij eksplisiet aanneemt dat de individuen een willekeurige steekproef uit een populatie zijn, terwijl hij bovendien vrijwel overal uitgaat van multinormaal verdeelde gegevens. Misschien dat Morrison daarom, in navolging van Anderson, spreekt van multivariate statistische analyse. Het voornaamste verschil met Anderson is dat hij veel aandacht schenkt aan CA, PCA, en faktor analyse (voortaan FA), en dat met name PCA in de eerste plaats als een deskriptieve data-reduktie techniek beschreven wordt. Morrison's boek geeft ook een uitvoerige inleiding in de matrix algebra. De wijzigingen

in de nieuwe editie van 1971 zijn voor onze doeleinden nauwelijks van belang.

1.1.6 Van de Geer (1967,1971)

Op de flap van het nederlandse boek van Van de Geer lezen we: "Multivariate analyse, dat is de kunst om de samenhang tussen meerdere variabelen op eenvoudige wijze te beschrijven met behulp van wiskundige bewerkingen." Bij Cooley en Lohnes zagen we (vooral in de eerste versie) nadruk op de komputer programmaas, geen statistiek, en weinig algebra. Bij Van de Geer ontbreekt de statistiek eveneens, maar neemt de matrix algebra bijna de helft van het boek in beslag. MVA komt pas aan het eind. "In deze laatste hoofdstukken wordt de multivariate analyse uitsluitend behandeld als een data-reductie methodiek. M.a.w. het boek gaat nergens in op statistische vragen in de zin van de inferentiële statistiek." (1967, pag 12). De komputer wordt als belangrijke stimulans genoemd, en wel in die zin dat de tijd die men nu niet meer achter de tafelrekenmachine doorbrengt gebruikt kan worden voor het verkrijgen van inzicht in wat men eigenlijk aan het doen is. Het is duidelijk dat het inzicht dat Van de Geer bij probeert te brengen voornamelijk geometrisch van aard is, waar het maar mogelijk is worden figuren gebruikt. Om het even in de terminologie van kollegaas uit een geheel andere sektor te zeggen: hier is 'inzicht brengen' pas werkelijk 'in zicht brengen'. In de sterk uitgebreide engelse editie vinden we de uitgangspunten nog duidelijker weergegeven. "In my opinion, statistical theory is a substantially more advanced subject than is required to understand what multivariate techniques really do with data." (1971, p ix). Als belangrijkste voordelen van zijn benadering van MVA noemt Van de Geer weer het bevorderen van inzicht in plaats van het overschrijven van komputer output, het benadrukken van de overeenkomsten tussen de verschillende vormen van MVA, en de hieruit voortvloeiende 'kruisbestuiving' tussen de verschillende sociale wetenschappen. Belangrijke toevoegingen in het tweede boek zijn dan ook pad-analyse en strukturele vergelijkingen, die behandeld worden als twee nieuwe variaties op hetzelfde thema. Dat de diverse vormen van MVA omschreven kunnen worden als variaties van een algemeen eigenvektoren-eigenwaarden probleem werd overigens al rond 1950 door Bartlett en Tintner benadrukt. Zij bekeken echter vooral de rekenkundige overeenkomsten.

Volgens Van de Geer is MVA dus de lineaire analyse van data-matrixen, met als belangrijkste doeleinden data-reduktie en, bovenal, het maken van geometrische afbeeldingen. MVA is toegepaste lineaire algebra of, en dat is natuurlijk hetzelfde, toegepaste lineaire

meetkunde. Deze nadruk op plaatjes (ook plaatjes in de vorm van pijldiagrammen) vinden we later, en in mindere mate, by Cooley en Lohnes (1971) en bij Kendall (1975). Voor mensen die vertrouwd zijn met de psychometrische traditie, en met name met multipele FA en de data theorie van Coombs, is deze nadruk op de geometrische interpretatie van gegevens niet vreemds. In de binnenkort te verschijnen nieuwe editie van het engelse boek wordt er zo mogelijk nog meer aandacht geschonken aan de meetkundige aspekten van MVA.

1.1.7 Dempster(1969)

Het boek van Dempster heeft een prima inleidend hoofdstuk. We vinden daar de volgende omschrijving. "The purpose of this book is to describe certain methods of analysis of statistical data arising from multivariate samples. A basic aim of such data analysis is to reduce large arrays of numbers to provide meaningful and reasonably complete summaries of whatever information resides in sample aggregates. Another aim is to draw inferences from sample aggregates to larger population aggregates from which the samples are drawn; that is, to understand how certain information about a sample provides uncertain information about a population." (1c, pag 3). We zien dat data analyse en statistiek van het begin af onderscheiden worden. Ze worden in de loop van het boek ook gescheiden behandeld, waarbij de data analyse het eerst aan bod komt. "While most books on statistical theory start out with sampling theory and attempt to make methods of data analysis follow, the attitude in this book is that the methods of data analysis are carried out largely because of the intrinsic appeal of the sample quantities computed. Such, at least, were the historical origins of the methods described here. Moreover, when viewed as producing descriptive or summary statistics, the methods have value even when assumptions like randomness of samples and normality of populations are quite unwarranted." (1c, pag 3). De gescheiden behandeling geeft echter toch min of meer de indruk dat de twee benaderingen weinig of niets gemeen hebben. In de inleiding geeft Dempster als relatie dat de statistiek gebruikt kan worden om te laten zien dat bepaalde data analytische technieken bepaalde aantrekkelijke eigenschappen hebben, of juist dat andere technieken die eigenschappen niet hebben en daarom verbeterd kunnen worden. In het boek zelf vinden we van dit soort interaktie echter weinig voorbeelden.

Voordat hij aan MVA toekomt geeft Dempster, evenals Van de Geer, een uitvoerige inleiding in de theorie van matrices en lineaire ruimten. Zijn inleiding gebruikt zo min mogelijk koordinaten, en

is daardoor in hoge mate meetkundig. Niettemin wordt ook de rekenkundige vertaling van lineaire operaties uitvoerig besproken. We noemen nog even het lijstje van zaken die Dempster als zijn belangrijkste omissies beschouwd.

a: Kategorische data.

b: Gespecialiseerde, discipline-specifieke technieken als FA en lineaire strukturele vergelijkingen.

c: Cluster analyse en verwante technieken.

Als gevolg van deze omissies is de data analyse van Dempster beperkt tot het bespreken van diverse rekenkundige aspekten van het lineaire model. PCA en CA worden genoemd, er is een voorbeeld van eksploratief gebruik van CA, maar veel aandacht krijgen deze typisch multivariate technieken niet.

1.1.8 Tatsuoka (1971)

In dit veelgebruikte boek vinden we op de eerste bladzijde twee verschillende definities van MVA. "Multivariate statistical analysis, or multivariate analysis for short, is that branch of statistics which is devoted to the study of multivariate (or multidimensional) distributions and samples from those distributions." (1c, pag 1) Dit noemt Tatsuoka de definitie van de statistikus, in de praktijk kunnen we er weinig mee beginnen omdat hij nogal tautologisch klinkt. We kunnen daar aan toevoegen dat de definitie ook nogal imperialistisch klinkt, en dat multivariate analyse definieren als een afkorting van multivariate statistische analyse een nogal doorzichtige truuk is. De tweede definitie is die van de data analytikus. "In applied contexts, particularly in educational and psychological research, multivariate analysis is concerned with a group (or several groups) of individuals, each of whom possesses values or scores on two or more variables such as tests or other measures. We are interested in studying the interrelations among these variables, in looking for possible group differences in terms of these variables, and in drawing inferences relevant to these variables concerning the populations from which the sample groups were chosen." (1c, pag 1).

Het boek bevat veel matrix algebra, en aardig wat statistiek. Er is aanzienlijk minder geometrie als bij Van de Geer of Dempster, de aanpak van het lineaire model is die van Anderson of Morrison. De bekende univariate technieken worden zo goed en zo kwaad als het gaat gegeneraliseerd. "Pointing out the analogy between a given multivariate technique and the corresponding univariate method is one of the principal didactic strategies used throughout this book." (1c, pag 3). Deze benadering heeft het nadeel dat we het

implisiet als een soort van lastig beschouwen dat gegevens multivariaat zijn, en dat typisch multivariate technieken moeilijk vanuit dit gezichtspunt behandeld kunnen worden. PCA krijgt dan ook weinig aandacht, CA krijgt meer aandacht, maar de benadering van CA is statistisch als bij Anderson en gebaseerd op het multinormale model.

1.1.9 Harris (1975)

Er zit een zeker patroon in de besprekingen tot nu toe. In de later verschenen boeken gaat de data analytische kant van MVA over het algemeen meer plaats innemen, en worden de beperkingen van de inferentiele multinormale analyse wat duidelijker ingezien. Het boek van Harris is in sommige opzichten een reaktie op die trend. Het heeft een interessant eerste hoofdstuk, waarin de inferentiele statistiek verdedigd wordt. Harris durft het zelfs op te nemen voor de nulhypothese met bijbehorende signifikantietest, terwijl dit illustere duo in sommige data analytische kringen als morsdood en al jaren begraven geldt. We bespreken zijn argumenten hier niet in detail, daar hebben we verderop nog genoeg gelegenheid voor.

In grote lijnen gaat de redenering van Harris als volgt. Statistische methoden zijn een vorm van kwaliteitskontrole op wetenschappelijke produktie. Ze zijn nodig omdat omtelbare keren is gebleken dat alleen maar afgaan op de opinie van de onderzoeker over de generaliseerbaarheid van zijn resultaten tot niets leidt. We hebben dus formele methoden nodig, we kunnen kiezen tussen deskriptieve en inferentiele. De inferentiele methoden zijn bedoeld om te voorkomen dat de onderzoekers konklusies uit hun analyses trekken die niet generaliseerbaar zijn, dat wil zeggen die alleen maar volgen uit eigenschappen van de bepaalde steekproef die ze onderzocht hebben, of uit eigenschappen van de techniek die ze gebruikt hebben. "As will become obvious in the remaining sections of this chapter, the present Primer attempts in part to plug a 'loophole' in the current social control exercised over researchers' tendency to read too much from their data. (It also atteempts to add a collection of rather powerful techniques to the descriptive tools available to behavioral reasearchers. Van de Geer (1971) has in fact written a textbook on multivariate statistics which deliberately omits any mention of their inferential applications." (lc, pag 4-5). En welke gereedschappen gebruikt Harris voor dit loodgieterswerk ? In de eerste plaats kiest hij het bekende uitgangspunt dat MVA-technieken de gebruikelijke univariate technieken trachten te generaliseren. Ze doen dit vaak door groepen van variabelen te

vervangen door een lineaire kombinatie van de variabelen in die groepen. De koefficienten van de lineaire kombinaties worden gekozen door een bepaald optimaliteitskriterium te maksimaliseren. Data analytici zijn voornamelijk geinteresseerd in die koefficienten, ze noemen ze "ladingen" (van de geobserveerde variabelen op de gekonstrueerde variabelen), statistici zijn voornamelijk geinteresseerd in de maksimale waarde van het optimaliteitskriterium. Maar het is duidelijk dat deze twee zaken elkaar aanvullen, je kunt het ene zelfs niet berekenen zonder het andere te berekenen. En dus legt Harris de nadruk op optimaliteitskriteria met een relatief eenvoudige statistische interpretatie, die bovendien nog tot optimale koefficienten leiden. Hij prefereert daardoor dus bijvoorbeeld de "union-intersection" benadering van Roy boven de meer gebruikelijke "likelihood ratio" methode; die laatste methode levert meestal geen stel koefficienten op om de data analytisch georienteerde gebruiker tevreden te stellen. Matrix algebra krijgt weinig aandacht, omdat het alleen maar een techniek is om de optimalisaties efficient uit te voeren, en vanzelfsprekend is er weinig meetkunde in het boek te vinden. Meer nog dan bij Tatsuoka is de multivariaatheid van de gegevens iets lastigs, ze moeten dan ook zo snel mogelijk univariaat gemaakt worden.

1.1.10 Dagnelie (1975)

Het is leuk om de gebruikelijke definitie eens in het frans herhaald te zien. "Au sens large, l'analyse à plusieurs variables ou analyse multidimensionelle ou analyse multivariée ou analyse 'multivariate' peut être considerée comme formée de l'ensemble des méthodes statistiques qui ont pour objet l'étude des relations existant entre plusieurs variables dépendantes ou interdépendantes." (lc, pag 11). Dat klinkt toch gelijk een stuk mooier. Dagnelie komt overigens tot deze omschrijving door de definities van Anderson, Cooley en Lohnes, Kendall, Morrison, en Press naast elkaar te leggen. Dat is dan ook min of meer het teleurstellende van zijn boek, het is erg anglo-amerikaans georienteerd, brengt weinig nieuws, en laat geen typisch franse benadering zien. De gebruikelijke technieken worden behandeld, waarbij als belangrijkste onderscheid gebruikt wordt de gebruikelijke indeling van variabelen in subgroepen van ieder één of meerdere variabelen. Weinig algebra, weinig meetkunde, aardig wat kookboek-statistiek, en een aantal leuke voorbeelden uit de ekologie. "L'utilisateur doit évidemment avoir une idée suffisamment précise des principes généraux et des conditions d'application de ces methodes, mais il do-

pouvoir consacrer essentiellement son attention à l'interprétation des résultats obtenus." (lc, pag 18).

1.1.11 Green en Carroll (1976)

Dit is een ander soort boek. Op het eerste gezicht hoort het zelfs helemaal niet in onze opsomming thuis, omdat het alleen maar de wiskundige gereedschapskist van de onderzoeker wil aanvullen. Maar bij nadere bestudering wil het boek aanzienlijk meer, en zet het de tendens voort die we tot nu toe het meest uitgesproken aantroffen bij Van de Geer. MVA is toegepaste lineaire algebra en lineaire meetkunde. Er is een aardig inleidend hoofdstuk, waarin de opvattingen van de auteurs nader toegelicht worden. We geven wat representatieve citaten. "Completion of this book should provide both a technical base for tackling most applications-oriented multivariate texts and, more importantly, a geometric perspective for aiding one's intuitive grasp of multivariate methods." (lc, pag xii). "In function, as well as in structure, multivariate techniques form a unified set of procedures that can be organized around a relatively few prototypical problems." (lc, pag 1). "The heart of any multivariate analysis consists of the data matrix, or in some cases, matrices. The data matrix is a rectangular array of numerical entities whose informational content is to be summarized and portrayed in some way." (lc, pag 3). "To a large extent, the study of multivariate techniques is the study of linear transformations." (lc, pag 14). De boodschap is duidelijk. Wie deze "mathematical tools" beheerst kan MVA technieken in diverse vermommingen herkennen, en kan bovendien in de situatie waar hij mee te maken heeft eventueel zelf zijn eigen vorm van MVA konstrueren. De statistiek wordt verder in dit boek in het geheel niet meer genoemd. In het voorwoord lijkt het alsof Green en Carroll hun boek mede bedoelen als een inleiding op bijvoorbeeld Tatsuoka, Harris, of Morrison, maar het zou natuurlijk ook best zo kunnen zijn dat veel mensen na bestudering van dit boek besloten dat ze dat soort vervolgonderwijs niet meer nodig hadden. Er is weinig fantasie voor nodig om dit boek te zien als een boek over MVA, en niet als alleen maar een boek over mathematische benodigdheden.

1.1.12 Caillez en Pages (1976)

Er is in Frankrijk een sterke data analytische stroming, voornamelijk door het werk en de invloed van J.P. Benzécri. Op verschillende franse universiteiten zijn er zelfs statistische afdelingen onder leiding van data analytici die van de inferentiele statistiek niets moeten hebben. Het boek van Caillez en Pages gaat over de franse vorm van data analyse, die in een aantal opzichten van de

anglo-amerikaanse vorm verschilt. De franse lineaire algebra, bijvoorbeeld, is aanzienlijk abstrakter en moderner, en maakt veel minder gebruik van koordinaten en matriksen. Dit geheel in de traditie van moderne franse wiskundigen als Bourbaki en Dieudonné. Het boek van Caillez en Pages bevat een zeer uitvoerige inleiding in de lineaire algebra, en bovendien nog uitvoerige hoofdstukken over regressie, PCA, CA, en over Benzécri's versie van metrische meerdimensionele schaalmethoden. Bovendien is er een hoofdstuk over korrespondentie analyse, een eveneens door Benzécri ontwikkelde vorm van niet-lineaire PCA die we in het verloop van deze kursus nog vele malen tegen zullen komen. Er is een nuttig inleidend hoofdstuk over verzamelingen, relaties, afbeeldingen, en een nuttig uitleidend hoofdstuk over klassifikatie en kluster methoden, maar voor de rest is het lineaire algebra wat de klok slaat. "Un choix a été fait dans la facon de présenter les techniques d'analyse des données: nous utilisons constamment les notions de projecteur, d'application M-symétrique, de codage, et surtout le 'schéma de dualité' qui est un instrument de langage efficace permettant de présenter d'un seul jet et sous toutes leurs facettes les techniques relevant de l'algèbre linéaire." (lc, pag vii). Het boek heeft een voorwoord van Morlat, wat uitvoerig ingaat op de verschillen tussen data analyse en klassieke statistiek. Maar daar komen we later op terug.

1.1.13 Giri (1977)

Giri neemt de invariantie van toetsproblemen over groepen transformaties als uitgangspunt, zijn boek kan dan ook beschouwd worden als toepassing van aspekten van de beslissingstheorie op multinormaal verdeelde gegevens. De nadruk ligt op toetsen, niet op schatten. Om invariantie te bestuderen is vanzelfsprekend wat matriksalgebra nodig, en wat informatie over groepen transformaties. PCA, FA, CA worden zeer summier besproken.

1.1.14 Gnanadesikan (1977)

De gebruikelijke definitie van MVA wordt gehanteerd, maar de aanpak is ongebruikelijk. Gnanadesikan is in de eerste plaats data analytikus. "Much of the theoretical work in multivariate analysis has dealt with formal inference procedures, and with the associated statistical distribution theory, developed as extensions of and by analogy with quite specific univariate methods, such as tests of hypotheses concerning location and/or dispersion parameters. The resulting methods have often turned out to be of very limited value for multivariate data analysis." (lc, pag 2). De hier gepresenteerde technieken geven ook een visuele representatie van de data, maar in een wat

andere zin van het woord als we tot nu toe gewend zijn. Het boek gebruikt grafische technieken en waarschijnlijkheidsplots, er is minder lineaire algebra als bij Caillez en Pages en er is minder matriks algebra als bij Van de Geer. Zoals bij een data analytikus uit de school van Tukey te verwachten is, wordt er de nodige (grafische) aandacht besteed aan "goodness-of-fit" (van de multinormaal verdeling) en aan "outliers". Als belangrijkste doeleinden van MVA noemt Gnanadesikan de volgende.

a: Dimensionaliteits-reduktie.
b: Studie van afhankelijkheid.
c: Meerdimensionele klassifikatie.
d: Onderzoek van statistische modellen.
e: Samenvatting en verduidelijking.

Als belangrijkste problemen bij de ontwikkeling van MVA, in vergelijking met univariate analyse, geeft hij een lijstje dat veel lijkt op dat van Kendall (zie boven).

a: Het blijkt moeilijk te zijn om vast te stellen wat de onderzoeker precies wil weten.
b: Zodra we besloten hebben ons onderzoek multivariaat aan te pakken staan we voor het probleem het aantal variabelen te bepalen.
c: Zelfs met de tegenwoordige generaties komputers zijn er een aantal multivariate technieken die alleen maar toegepast kunnen worden voor een relatief klein aantal variabelen en/of individuen.
d: Het is moeilijk om in veel dimensies plaatjes en grafieken te maken.
e: Er is geen natuurlijke ordening gedefinieerd in de meerdimensionele ruimte.

1.1.15 Kshirsagar (1978)

Dit is een dik boek, en ook een goed boek, maar toch kunnen we kort zijn. Sinds Anderson hebben de multinormalen niet stilgezeten. Er is heel wat afgerekend, verschillende ekspansies hebben er één of meerdere termen bijgekregen. Nieuwe statistieken zijn bedacht, er zijn een aantal nieuwe technieken ontwikkeld om sneller en eleganter gekompliceerde verdelingen af te leiden. Het boek van Kshirsagar is tot nu toe de meest komplete samenvatting van multinormale analyse. Er zijn geen voorbeelden, er is geen FA, terwijl PCA en CA multinormaal behandeld worden. "The theory of multivariate analysis developed so far almost invariably assumes that the joint distribution of the random variables is a multivariate normal distribution." (lc, pag 1). Als men MVA definieert als multinormale analyse is dit natuurlijk een juiste konstatering.

Als belangrijkste omissie noemt Kshirsagar FA, tijdreeksen, en kategorische data. Hij beschouwt regressieanalyse als de belangrijkste statistische techniek, een typisch multinormale konstatering. Niettemin noemt hij data reduktie eveneens. "The aim of the statistician undertaking multivariate analysis is to reduce the number of variables by employing suitable linear transformations and to choose a very limited number of the resulting linear combinations in some optimum manner, disregarding the remaining linear combinations in the hope that they do not contain much significant information. The statistician thus reduces the dimensionality of the problem." (lc, pag 2).

1.1.16 Thorndike (1978)

Thorndike beschouwt zich, evenals Cooley en Lohnes, als data analytikus. En in zijn data analytische praktijk ondervindt hij veel hinder van het probleem dat zijn klienten dikwijls een aardig begrip hebben van variantie analyse, maar slecht thuis zijn in technieken gebaseerd op korrelatiekoefficienten. Vandaar dit boek. "The approach is largely intuitive and geometric rather than mathematical..." (lc, pag vi). Dit lijkt een wat ongelukkig gekozen tegenstelling. Het volgende citaat is wat minder kontroversieel. "The organization of the book reflects a conviction that understanding can be most readily developed by showing the essential unity and orderly progression of concepts in multivariate statistics. ... The geometric interpretation of the correlation coefficient as the angle between two vectors in a people space is readily generalizable to multiple and canonical correlation." (lc, pag vii). Er staat, zoals beloofd, weinig wiskunde in het boek. "The work of other authors, notable Quinn McNemar, Harry Harman, and Bill Cooley and Paul Lohnes, is frequently cited for the reader who wants or needs the mathematical foundations of the topics discussed." (lc, pag vii). Het staat er echt. Dit boek zou wel eens het MVA boek voor de vijftiger jaren kunnen worden. Thorndike gebruikt het Bartlett-Kendall onderscheid tussen interne (samenhang) analyse en externe (afhankelijkheid) analyse. Hij noemt dit interne en externe faktor analyse, zodat bijvoorbeeld multipele regressie een vorm van externe faktor analyse wordt, terwijl faktor analyse een speciaal geval van interne faktor analyse is. Het is opmerkelijk dat de lezer die op zoek is naar wiskundige grondslagen een lange weg af moet leggen. Thorndike verwijst naar Cooley en Lohnes, die verwijzen hem door naar Tatsuoka, die verwijst hem door naar Morrison, die zich voor de echte grondslagen op Anderson baseert. Ondertussen is de ongelukkige lezer twintig

jaar terug gereisd in de tijd, en tenslotte aangekomen bij een boek dat zich vrijwel uitsluitend met multinormale analyse bezighoudt. Misschien zou het beter zijn de lezer door te verwijzen naar een eenvoudig boek over lineaire algebra.

1.2 Analyse inhoudsopgaven van MVA-boeken

In tabel 1 staat aangegeven hoeveel bladzijden de MVA boeken die we besproken hebben besteden aan de volgende zeven onderwerpen.

A:WISK:

Wiskunde anders dan statistiek, dus lineaire algebra, matriksen, groepen transformaties, verzamelingen, relaties.

B:KORR:

Korrelatie en regressie, omvat ook pad analyse, lineaire strukturele vergelijkingen, funktionele relaties.

C:FAKT:

FA en PCA.

D:KANO:

CA.

E:DISK:

Diskriminant analyse, klassifikatie, kluster analyse.

F:STAT:

Statistiek, omvat zowel verdelingen als toetsings- en schattingsmethoden. Omvat ook statistische analyse van kategorische gegevens.

G:MANO:

Omvat MANOVA, en het algemene multivariate lineaire model.

Op deze indeling is natuurlijk veel aan te merken. Sommige onderwerpen die niet zoveel met elkaar te maken hebben (bijvoorbeeld diskriminant en kluster analyse) zitten in dezelfde kategorie, terwijl bij voorbeeld CA bij Anderson of Kshirsagar iets anders is dan CA bij Van de Geer of Cailliez en Pages. We kozen deze indeling omdat we een zo groot mogelijk gedeelte van de behandelde stof willen onderbrengen, terwijl we toch ook niet te veel nullen in de data matriks kunnen gebruiken. Om die laatste reden was het nodig een aanvankelijk gebruikte fijnere indeling wat in te dikken. Overigens is natuurlijk onze keuze uit MVA boeken ook niet bepaald een willekeurige steekproef. We hebben de meest gebruikte leerboeken opgenomen, plus een aantal die ons persoonlijk na aan het hart liggen, plus een aantal nogal ekstreme, plus een aantal die op dat moment in de bibliotheek aanwezig waren.

Op tabel 1 hebben we korrespondentie analyse toegeapst. Voor het

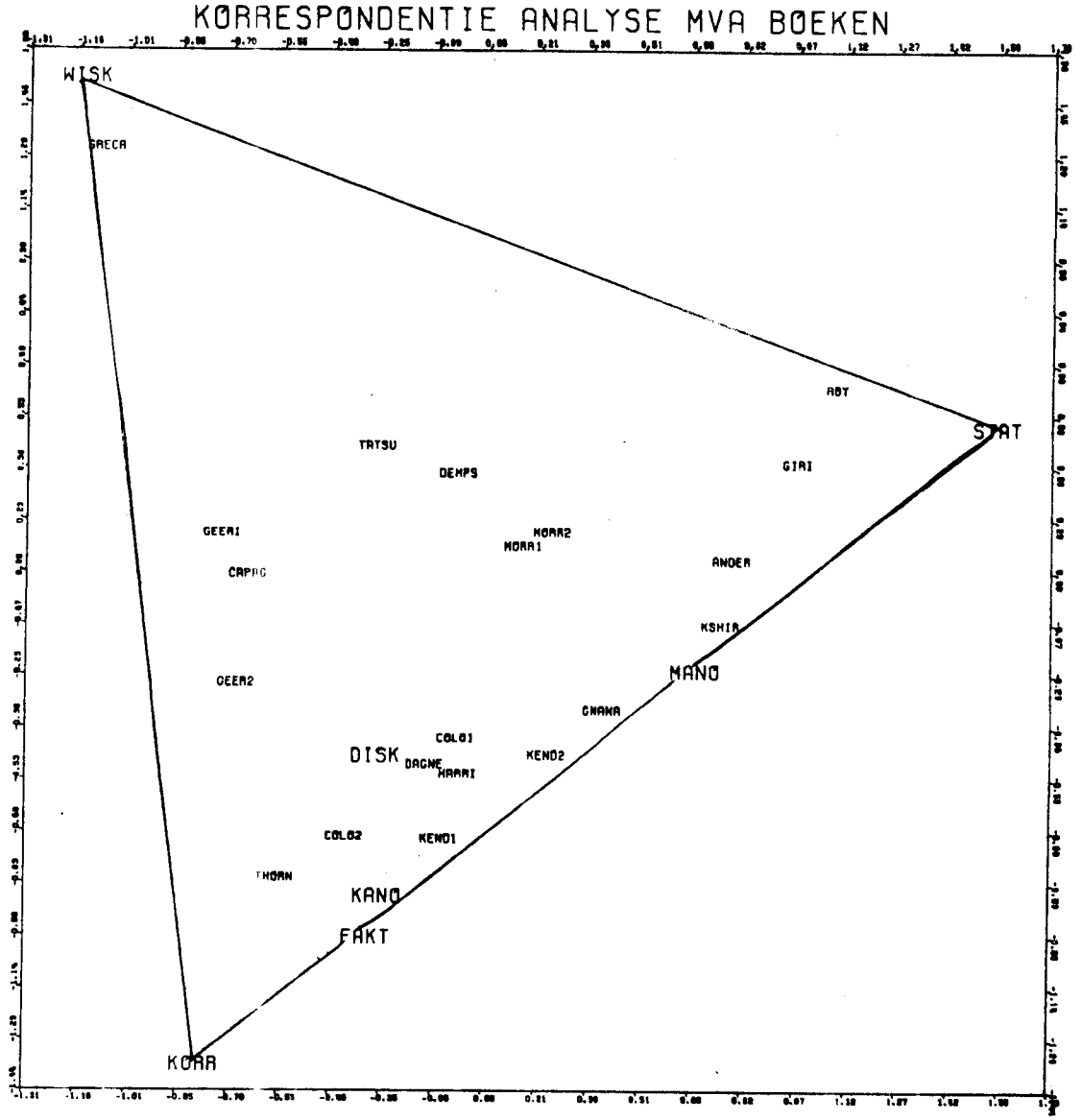
		A	B	C	D	E	F	G
1	Roy (1957)	31	0	0	0	0	164	11
2	Kendall (1957)	0	16	54	18	27	13	14
3	Kendall (1975)	0	40	32	10	42	60	0
4	Anderson (1958)	19	0	35	19	28	163	52
5	Cooley & Lohnes (1962)	14	7	35	22	17	0	56
6	Cooley & Lohnes (1971)	20	69	72	33	55	0	32
7	Morrison (1967)	74	0	86	14	0	84	48
8	Morrison (1976)	78	0	80	5	17	105	60
9	Van de Geer (1967)	74	19	33	12	26	0	0
10	Van de Geer (1971)	80	68	67	15	29	0	0
11	Dempster (1969)	108	48	4	10	46	108	0
12	Tatsuoka (1971)	109	13	5	17	39	32	46
13	Harris (1975)	16	35	69	24	0	26	41
14	Dagnelie (1975)	26	86	60	6	48	48	28
15	Green & Carroll (1976)	290	10	6	0	8	0	2
16	Caillez & Pages (1976)	184	48	82	42	134	0	0
17	Giri (1977)	29	0	0	0	41	211	32
18	Gnanadesikan (1977)	0	19	56	0	39	75	0
19	Kshirsagar (1978)	0	22	45	42	60	230	59
20	Thorndike (1978)	30	128	90	28	48	0	0

tabel 1: aantal bladzijden van MVA boeken besteed aan diverse onderwerpen.

1	ROY	1.1075	0.6141	-0.3444
2	KEND1	-0.0740	-0.7028	0.2523
3	KEND2	0.2384	-0.4584	-0.4785
4	ANDER	0.7766	0.1107	0.1557
5	COLO1	-0.0286	-0.4090	1.0633
6	COLO2	-0.3574	-0.6966	0.0910
7	MORR1	0.1643	0.1536	0.4569
8	MORR2	0.2496	0.1939	0.3870
9	GEER1	-0.7282	0.1929	-0.0469
10	GEER2	-0.6827	-0.2460	-0.1877
11	DEMPS	-0.0282	0.3666	-0.4431
12	TATSU	-0.2698	0.4465	0.2886
13	HARRI	-0.0202	-0.5133	0.5052
14	DAGNE	-0.1200	-0.4852	-0.2003
15	GRECA	-1.0836	1.3213	0.0308
16	CAPAG	-0.6512	0.0719	-0.1241
17	GIRI	0.9811	0.3946	-0.2470
18	GNANA	0.3992	-0.3273	-0.3381
19	KSHIR	0.7459	-0.0801	-0.0040
20	THORN	-0.5634	-0.8163	-0.3634
A	WISK	-1.1446	1.5214	0.0678
B	KORR	-0.7979	-1.3651	-1.2861
C	FAKT	-0.2897	-0.9884	0.6286
D	KANO	-0.2586	-0.8694	0.8394
E	DISK	-0.2656	-0.4599	-0.7342
F	STAT	1.5624	0.4981	-0.6139
G	MANO	0.6722	-0.2137	2.5113
LABDA		0.3650	0.2632	0.1124

tabel 2:
projekties voor boeken, projekties voor onderwerpen en eigenwaarden uit korrespondentie analyse van tabel 1.

moment is het voldoende te weten dat dat een soort meerdimensionele schaalmethode is, die rij- en kolompunten gezamenlijk afbeeldt in een plaatje met een beperkt aantal dimensies. Daarbij liggen twee boeken dicht bij elkaar in het plaatje als ze een vergelijkbare inhoud hebben, twee onderwerpen liggen dicht bij elkaar als ze op dezelfde wijze in de gekozen boeken voorkomen, en een onderwerp ligt dicht bij een boek als in dat boek relatief veel aandacht aan dat onderwerp besteed wordt. In tabel 2 staan de projekties van de boeken en onderwerpen op de drie belangrijkste dimensies. Korrespondentie analyse is een eigenwaarden-eigenvektoren techniek waarin de belangrijkheid van de dimensie gedefinieerd wordt door de grootte van de bijbehorende eigenwaarde. De drie grootste eigenwaarden staan in de laatste rij van tabel 2, de drie overige eigenwaarden zijn .0440, .0305, en .0102. De projekties op de eerste twee dimensies staan afgebeeld in figuur 1. In deze figuur hebben we een driehoek getekend met als hoekpunten KORR, STAT, en WISK. De meest zuivere representanten van deze onderwerpen zijn de boeken THORN, ROY, en GRECA, respektievelijk. De meeste MVA boeken zitten op de zijde tussen KORR en STAT, dit zijn de meer klassieke boeken die voornamelijk varieren in moeilijkheidsgraad. De typisch data analytische boeken zitten op de zijde tussen KORR en WISK, als we tenminste data analyse definiëren als toegepaste lineaire algebra. De grafische data analyse van GNANA past beter in de klassieke schaal. In het binnenste van de driehoek vinden we vier boeken die zowel veel aandacht besteden aan lineaire algebra als aan statistiek. De zijde tussen STAT en WISK is leeg, daardoor is het ook mogelijk het plaatje te interpreteren als een gekromde schaal die loopt van WISK via KORR naar STAT. Het boek COLO1 ligt niet op de juiste plaats, de derde dimensie in tabel 2 laat zien dat dit komt omdat het veel meer aandacht aan MANOVA besteedt dan men op basis van het "technisch niveau" zou verwachten. Maar COLO1 is dan ook een handleiding voor komputergebruik, dat op nogal ontechnische wijze aandacht aan MANOVA besteedt. We hebben de analyse ook herhaald zonder de ekstremisten GRECA en ROY. De positie van de overige punten blijft dan vrijwel gelijk, alleen de eerste twee eigenwaarden worden aanzienlijk kleiner (te weten .3139 en .1490). De derde eigenwaarde is .1080, ongeveer gelijk dus, en de derde dimensie blijft COLO1 met de rest kontrasteren. Het is duidelijk dat we in de figuur dezelfde onderscheidingen aantreffen als in de inhoudsanalyse, ze zijn nu echter wat kompakter en overzichtelijker weergegeven.


KORRESPONDENTIE ANALYSE MVA BOEKEN
WISK
GRECA
ABT
STAT
TATSU
DEMPS
GIRI
GEERA1
MORA2
MORA1
CAPRO
ANDER
KSHIA
MANO
GEER2
GHANA
COLG1
DISK
DAGNE
KEND2
HARRI
COLG2
KEND1
THORN
KANO
FAKT
KORA

Figuur 1.1.

1.3 Korte samenvatting en probleemstelling

In onze diskussie totdusver komen een aantal belangrijke punten naar voren. In de eerste plaats schijnt er een zekere tegenstelling te bestaan tussen de data analytische en de statistische aanpak van MVA. De statistische aanpak van MVA gaat uit van een bepaald statistisch model, over het algemeen gebaseerd op de multinormaal verdeling, binnen dit model worden dan specifieke parametrische hypothesen opgesteld. Vervolgens worden de overbijvende vrije parameters geschat, en worden de hypothesen getoetst. De data analytische aanpak gaat daarentegen niet uit van een bepaald model, maar zoekt naar transformaties en kombinaties van de variabelen met het doel de gegevens op een eenvoudige en overzichtelijke manier weer te geven. Het lijkt zeker nodig om wat nader in te gaan op de rol en de betekenis van het statistische model met de bijbehorende toetsings- en schattingsmethoden. En in het bizonder op de rol van de multinormaal verdeling in MVA.

Een tweede probleem wat uit de besprekingen naar voren komt vormt de rol van kategorische gegevens. Of, in andere termen, van nominale en ordinale variabelen in MVA. Alleen Roy en Kendall besteden redelijk veel aandacht aan dit soort variabelen, bij anderen komen ze uitsluitend terug als koderingen in de kontekst van MANOVA of diskriminant analyse. Ze worden dan wel "dummy variables" genoemd. Over het algemeen worden ze als niet-stochastisch opgevat, ze zijn onderdeel van de "design matrix", oftewel onafhankelijke variabelen. Eigenlijk zelfs geen variabelen, in feite worden dummy variabelen alleen maar gebruikt om parametrische multinormale modellen in kompakte matriks notatie neer te schrijven.

In onze lijst van MVA boeken hebben we ons beperkt tot algemene inleidingen met een behoorlijke mate van overlap. Er zijn echter ook een aantal boeken die speciaal over de analyse van multivariate kategorische gegevens gaan. De belangrijkste zijn Haberman (1974), Bishop, Fienberg, en Holland (1975), en Gokhale en Kullback (1978). De inhoud van dit soort boeken lijkt in weinig opzichten op die van de boeken die wij besproken hebben. Weliswaar zijn de modellen geformuleerd in de traditie van de klassieke statistiek, maar de nadruk ligt veel meer op asymptotische methoden, en lineaire algebra en meetkunde spelen vrijwel geen rol. Het is duidelijk dat er uit een andere traditie gewerkt wordt, en dat is een beetje vreemd omdat beide vormen van MVA direkt voortkomen uit het werk van Pearson en Fisher. In de klassieke algemene handboeken over

statistiek, zoals die van Fisher, Yule en Kendall, Cramér, Kendall en Stuart, Wilks, en Rao, worden de twee vormen van MVA naast elkaar behandeld, en niet of nauwelijks aan elkaar gerelateerd. Er gaapt een kloof tussen het kontinue en het diskrete.

Een gedeeltelijke uitzondering op deze regel zijn de boeken van Kullback (1959) en Lancaster (1969). Kullback baseert zijn statistische procedures op de informatie-theoretische "divergentie" als afstandsmaat tussen parametrische verdelingen, en deze maat is voor diskrete en kontinue verdelingen op precies dezelfde wijze gedefinieerd. Lancaster gaat het om de ontbinding van multivariate verdelingen met behulp van orthogonale funkties op de marginalen, en ook deze techniek is zowel op diskrete als op kontinue verdelingen toepasbaar. Kullback's procedures leiden direkt naar log-lineaire modellen en naar de multiplikatieve definitie van interaktie, Lancaster's methoden leiden naar de additieve definitie van interaktie voor multivariate verdelingen. We zullen in het vervolg van dit hoofdstuk nader in moeten gaan op onze plaats tussen diskrete en kontinue MVA.

1.4 Data analyse en statistiek

De inferentiele statistiek heeft, eigenlijk al sinds zijn begin bij Laplace en Quetelet, te lijden onder wat men gewoonlijk een grondslagen krisis noemt. Het is duidelijk dat er op de waarschijnlijkheidstheorie als zodanig weinig aan te merken is, evenmin zijn er zinnige argumenten in te brengen tegen het gebruik van de deskriptieve statistiek. Maar zodra er sprake is van rationele theorieen van induktie, van statistische wetten, van waarschijnlijkheden van wetenschappelijke hypothesen of theorieen, of van waarschijnlijkheden van toekomstige gebeurtenissen, dan zijn de moeilijkheden niet van de lucht. Er zijn een groot aantal "scholen" ontstaan, die zeer uiteenlopende standpunten hebben over de grondslagen van de statistiek, maar ook over de waarde van specifieke statistische procedures zoals de befaamde nul-hypothese toets. Als men de diskussies leest die over de grondslagen plaatsvinden, of de verslagen van konferenties die zich met dit onderwerp bezighouden, dan is het duidelijk dat er wenig overeenstemming is, en dat er in de toekomst ook weinig overeenstemming te verwachten is. Het is prettig leeswerk, vooral als men van nogal grove polemische taal houdt, maar het zet uitermate weinig zoden aan de dijk. Over één ding zijn de diverse scholen het echter hardgrondig eens. Er wordt altijd geredeneerd binnen een statistisch model, het probleem begint pas wanneer men vanuit het statistisch model uitspraken wil doen

over de tegenwoordige of toekomstige werkelijkheid.

Het is niet verwonderlijk dat vele statistici en vakwetenschappers zich met een licht gevoel van walging van dit soort diskussies afgekeerd hebben. Sommigen van hen zijn tot de konklusie gekomen dat de statistiek helemaal geen grondslagen onderzoek of zelfs geen grondslagen nodig heeft. Deze mensen worden echter ook konsekwent op de symposia uitgenodigd, met het gevolg dat er een nieuwe school gekreeerd is die alle andere scholen uitmaakt voor navelstaarders. En dat gaat al zo'n vijftig jaar door. Voor onze doeleinden is het vooral van belang te konstateren dat, mede als gevolg van de oeverloze diskussies tussen de voorstanders van het statistische model, de rol van het statistische model zelf ter diskussie is komen te staan.

Een probleem dat hier direkt mee te maken heeft is dat volgens sommigen de statistiek voor een groot deel overgenomen is door wiskundigen, die onleesbare artikelen schrijven met vrijwel geen praktische relevantie, behalve dan dat ze er werkgelegenheid mee kreëren voor hun jongere kollegas. Dit standpunt vindt men bij radikale kritici van de inferentiele statistiek, zoals Hogben (1957) en Kempthorne (1971, 1972), maar ook bij meer gematigden als Kendall (1972). Dat er twee geheel verschillende klassieke benaderingen van de mathematische statistiek zijn wordt keurig geillustreerd door het lezen van Kiefer's befaamde bespreking (Kiefer, 1964) van deel II van het nog veel befaamdere handboek van Kendall en Stuart. Men krijgt bij het lezen van dit soort oorlogszuchtig geschrijf al snel de neiging om te denken dat de waarheid wel ergens in het midden zal liggen. Bij nadere beschouwing blijkt het echter eerder om onverenigbare tegenstellingen te gaan, en blijkt dat midden helemaal niet te bestaan. Kiefer heeft, natuurlijk, op alle punten het grootste gelijk van de wereld, alleen is het zo uitermate gemakkelijk op zijn manier heel erg veel keer gelijk te hebben.

We stippen hier ook even het probleem van de <u>aanmatiging</u> aan. Kiefer's bespreking is daar niet eens zo'n sterk voorbeeld van, maar over het algemeen is het al heel lang zo dat wiskundig georienteerde statistici de neiging hebben zich nogal superieur op te stellen. We voegen hier haastig aan toe dat deze tendens zich overigens niet beperkt tot de statistiek. Vakmethodologen stellen zich aanmatigend op tegen vakwetenschappers, theoretici ten opzichte van experimentelen, experimentelen ten opzichte van toegepasten, enzovoorts. Pearson maakte zich er schuldig aan tijdens de hoogtijdagen van het konflikt tussen biometrici en Mendelianen. Hij bleek ongelijk te hebben. Sir Fisher was, ook in dit opzicht, een klasse apart. De gewone stervelingen, die bepaalde aspekten van zijn genetisch en statistisch werk niet konden volgen, werden met hoon overladen. Langzamerhand blijkt steeds duidelijker dat Fisher op kritieke punten echt onbegrijpelijk was.

Ander voorbeeld. Een stuk onschuldiger, maar ook een stuk dichter bij huis. Zelfs Molenaar, toch een gematigd man, kan het niet nalaten zich, in een artikel met de ondeugende titel "Ik word ziek van die statistiek", te vergelijken met een voorganger die tot zijn gemeente preekt (Molenaar, 1974). We noemen die titel ondeugend omdat Molenaar zelf helemaal niet ziek wordt van de statistiek, maar "koude rillingen" krijgt van de manier waarop de statistiek in de sociale wetenschappen dikwijls gebruikt wordt. En we vinden de vergelijking nogal aanmatigend, omdat de voorganger nog altijd beschouwd wordt als een middelaar tussen de zondige gemeente en God, waarbij God in dit geval de Theorie van Neyman en Pearson en het Theorema van Bayes geopenbaard heeft.

Laatste voorbeeld. Nog dichter bij huis, maar weer een stuk minder onschuldig. We bedoelen het onlangs verschenen rapport "Methoden, voetangels en klemmen in de factoranalyse " (Bethlehem e.a., 1977). Om misverstanden te voorkomen moeten we in de eerste plaats zeggen, dat we dit rapport een degelijk en interessant stuk werk vinden. En niet alleen omdat het ons een blik gunt in de keuken van de mathematische statistiek, want in die keuken blijkt weinig eetbaars te halen te zijn, maar ook omdat het wel degelijk een belangwekkend verhaal is over een bepaald statistisch model, dat sommigen (geheel ten onrechte overigens) identificeren met faktor analyse. Het zou ons inziens echter nogal rampzalige, hoewel tegelijkertijd ook nogal komische, gevolgen hebben wanneer dit rapport op enige schaal gebruikt zou gaan worden als praktische handleiding. Het rapport begint, evenals het eerder genoemde verhaal van Molenaar, op een uiterst gematigde toon, en in een sfeer van "sweetness and light". De psychometrici, statistici, en gebruikers worden allen aan de oekomenische tafel genood. Vervolgens wordt faktor analyse uiterst restriktief gedefinieerd, en wordt er veel aandacht besteed aan het voor de hand liggende feit dat aan deze uiterst restriktieve aannamen zelden wordt voldaan. Het meeste aandacht besteden de auteurs aan het probleem van de ongedetermineerdheid van faktor skores, een probleem dat niet specifiek is voor faktor analyse, maar dat een rol speelt in alle (psychometrische of ekonometrische) modellen met "latente" of "niet-observeerbare" variabelen. De auteurs laten, in navolging van Guttman, uitvoerig zien dat "niet-observeerbare" variabelen inderdaad niet observeerbaar zijn. Het rapport is uitgerust met een indrukwekkend ogend blokschema gevuld met gewetensvragen. De ongelukkige onderzoeker die tracht de vragen eerlijk te beantwoorden bevindt zich na hoogstens twee vragen al bij

het advies met de analyse te stoppen, of zijn heil elders te zoeken. Daar zit hij dan aan de oekomenische tafel, met zijn dure databestand in zijn tasje, wreed uit het blokschema gestoten voordat hij zijn gegevens zelfs maar te voorschijn heeft kunnen halen. Hij wordt ondertussen een beetje moe van al die mensen die zo graag willen bemiddelen, en vervolgens onmiddellijk aan het hoofd van de tafel plaats nemen. Thuisgekomen valt zijn oog op een citaat. "Statistique n'est pas probabilité. Sous le nom de statistique mathématique, des auteurs ... ont édifié une pompeuse discipline, riche en hypothèses qui ne sont jamais satisfaites dans la pratique. Ce n'est pas de ces auteurs qu'il faut attendre la solution de nos problèmes typologiques." (Benzécri, 1973, pag 3).

Met dit laatste citaat zitten we opeens midden in de data analyse. Volgens sommigen dateert die data analyse van 1962, omdat toen Tukey zijn befaamde artikel "The future of data analysis" publiceerde, waarin met veel verbaal geweld de nieuwe term aan het publiek gepresenteerd werd. Met bijvoorbeeld als gevolg dat Cooley en Lohnes hun identiteit ontdekten. Maar natuurlijk bestaat data analyse al veel langer. De klassieke Engelse school van statistici (Bartlett, Anscombe, Udny Yule, Kendall) beschouwt data analyse als alleen maar een nieuwe naam voor statistiek zoals zij dat beoefenen. "Whereas the content of Tukey's remarks is always worth pondering, some of his terminology is hard to take. He seems to identify 'statistics' with the grotesque phenomenon generally known as 'mathematical statistics', and finds it necessary to replace 'statistical analysis' by 'data analysis'." (Anscombe, 1967, pag 3). Morlat (voorwoord van Caillez en Pages, 1976) neemt een ander standpunt in. Volgens hem is data analyse, altans de Franse versie "Analyse des données", niets anders dan de oude vertrouwde deskriptieve statistiek, maar dan in een multivariate versie die mogelijk gemaakt wordt door de opkomst van de komputer. En deskriptieve statistiek is natuurlijk al heel oud (zie bv. Kendall, 1972). Ook in de tijd van Pearson en de vroege Fisher was de statistiek data analytisch. "It seems to me that my attitude is nearer to R.A. Fisher's earlier outlook, when he emphasized the reduction of data, the sampling properties of estimates, etc. than to his later attitude, which was closer to that of writers like G. Barnard, Edwards, and Hacking, when they discuss artificial examples which seem to me at times to be in danger of being over-academic and narrow." (Bartlett,

1971, pag 22). Langzamerhand zien we dat het statistische model een steeds grotere rol gaat spelen, met als voorlopige hoogtepunten de theorie van hypothese toetsing van Neyman en Pearson, en de algemene beslissingstheorie van Wald. In deze laatste procedures zijn er nog steeds een aantal subjektieve keuzen die gemaakt moeten worden voor we aan het optimaliseren slaan. Dit leidde tot een opleving van de Baysiaanse School, en tot de bekende diskussies en kontroversen.

Mensen als Hogben (1957) riepen al zeer vroeg dat de Neyman-Pearson theorie in de praktijk nauwelijks toepasbaar was. De toetsingsprocedures zijn misschien geschikt voor kwaliteits-kontrole ("acceptance sampling"), omdat daar de twee beslissingen duidelijk gedefinieerd zijn en er een aanvaardbare interpretatie van waarschijnlijkheid in termen van relatieve frekwentie is. Maar het proces van wetenschappelijke theorievorming lijkt in het geheel niet op "acceptance sampling". En in feite gaat ditzelfde argument op voor de beslissingstheorie. Op een zeer abstrakt nivo kan men wetenschappelijk bezig zijn inderdaad beschrijven als het nemen van beslissingen, maar hetzelfde geldt voor het doen van inkopen voor het week-end of voor het al dan niet opsteken van de volgende sigaret. "The theory talks as though one can anticipate contingencies like the concentration of poisons by the biological tree, and one can attach costs to the loss of a species, and so on and so on. This arrogance, coupled with stupidity, has plagued statistics and science for years, and is no longer tolerable." (Kempthorne, in de diskussie van Rubin, 1971). In een wat meer gematigde vorm komen dezelfde argumenten terug bij Bartlett (1971), Kempthorne (1971, 1972), en bij Hammersley (1976). Niettemin vinden veel van deze auteurs de standaard statistische procedures zoals t-toetsen, variantie analyse en korrelatierekening wel degelijk van groot data analytisch belang. De klassieke statistiek wordt daarbij opgevat als een vorm van sensitiviteitsanalyse of perturbatie-analyse, die laat zien hoe data analytische procedures zich gedragen in geideali-seerde situaties en in hoeverre de resultaten veranderen als men de gegevens verandert.

Aan Tukey danken we dus voornamelijk de naam "data analyse", maar in ieder geval toch ook wel een gedetailleerde analyse van de verhouding van data analyse en mathematische statistiek. Tukey onderscheidt drie aspekten van data analyse, te weten inferentie, eksploratie, en design. Hij ontkent nergens het nut van de

mathematische statistiek, maar hij impliceert wel dat een groot deel ervan niet bijdraagt aan de data analyse, en daarom beoordeeld moet worden als een vorm van zuivere wiskunde. Een aanzienlijk minder hard oordeel dan dat van Anscombe (zie boven). Als het belangrijkste kenmerk van de mathematische statistiek noemt Tukey de nadruk op optimalisatie van een precies kriterium, en die nadruk is het gevolg van de poging om de rol van subjektieve keuzen en oordelen terug te dringen. Bij data analyse zijn er echter drie soorten oordelen van groot belang.

a: Oordelen gebaseerd op wat men weet over het wetenschapsgebied van waaruit de gegevens afkomstig zijn.

b: Oordelen gebaseerd op de ervaring die men heeft met het type techniek dat men toepast.

c: Oordelen gebaseerd op mathematische analyse van de eigenschappen van de techniek die men toepast (inklusief Monte Carlo studies).

Beginnen met een aantrekkelijk kriterium en dan de beste procedure in termen van dat kriterium vinden is één mogelijkheid, maar beginnen met een aantrekkelijke procedure en dan kriteria of situaties vinden waarin deze procedure optimaal is, is even rationeel. "The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: 'Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." (Tukey, 1962, pag 13-14).

De kritiek van Benzécri op de mathematische statistiek is een stuk radikaler als die van Tukey, zoals we in een eerder citaat al gezien hebben. In "Les principes de l'analyse des données" (uit 1969, verschenen in Benzécri e.a. 1973, deel II) geeft hij een aantal data analytische principes, waarvan we de eerste al tegengekomen zijn.

a: Statistiek valt niet samen met waarschijnlijkheidsleer. "... les bases de l'analyse statistique sont plutôt algébriques ou géométriques...que probabilistes...Mais les conceptions probabilistes suggèrent des opérations algébriques et permettent parfois d'en évaluer la portée." (lc, pag 6).

b: Het model moet uit de gegevens volgen, en niet andersom. Eerst de gegevens, dan pas het model.

c: Bij data analyse proberen we informatie te verzamelen over een zo groot mogelijk aantal variabelen.

d: De komputer is onmisbaar.

e: Technieken die niet van de komputer gebruik maken zijn verouderd.

Principe a lijkt ons het belangrijkst, het sluit ook het beste aan op onze analyse van MVA boeken. Principes b en c worden gebruikt in de aanval op de mathematische statistiek, en op de modeltoets in het bizonder. Principe c is aanvechtbaar, omdat het kan leiden tot het waanzinnig soort empirisme dat bijvoorbeeld de psychometrische intelligentietheorie en de Buikhuisense variant van de kriminologie zo aanvechtbaar maakt. Principes d en e lijken ons in hun algemeenheid onjuist, maar ze zijn juist als we ze beperken tot MVA. Het is duidelijk dat Benzécri tot degenen behoort die vinden dat:" ...l'essentiel de la statistique, voire la statistique toute entière, doit se réduire à l'analyse des données, la statistique mathematique classique n'étant qu'un ensemble de jeux de l'esprit, quelque peu arbitraire, destinés à meubler les loisirs des statisticiens qui à l'époque ne disposaient pas d'ordinateurs pour résoudre des problèmes plus réalistes." (Morlat, lc, pag iii). Morlat vindt dit zelf te ver gaan, en wij zijn geneigd het met hem eens te zijn. Er is voor-alsnog geen reden om de aanmatigende houding van sommige statistici over te nemen.

Samenvattend willen we de volgende stellingen poneren, die tegelijkertijd uitgangspunten van deze klapper zijn.

a: Er is geen bevredigende meta-theorie van de inferentiele statistiek, het is zelfs de vraag of zo'n theorie mogelijk en/of nodig is. De aanhangers van de Neyman-Pearson theorie hebben nooit de moeite genomen de tegenwerpingen van Fisher, Hogben, Kempthorne, en vele anderen effektief te weerleggen. In plaats van de relevantie van de theorie aan te tonen, heeft men haar alleen maar uitgebouwd. "What the statistician has done is to reject the experimentalist's question, and to substitute for it a question of his own relating to the long-run reliability of his own performance, to which he has given the mathematically correct answer. In short, the confidence theorist has played a confidence trick and given the right answer to the wrong question." (Hammersley, 1976, pag 28-29).

b: Uit a volgt echter helemaal niet dat de standaard statistische procedures geen nut hebben, of zelfs maar dat de nulhypothese toets geen nut heeft. Met name is het in sommige situaties, waarin het model a priori aannemelijk gemaakt kan worden, nuttig te redeneren vanuit het statistische model. En s het in sommige situaties essentieel om het onderscheid te maken tussen steekproef en populatie. "Mathematical statisticians have long

concentrated their efforts mainly on the supportive side. In reaction to this one-sidedness, and also to the controversial character of the concepts of inference, some data analysts have claimed that a careful exploratory analysis can turn up everything of interest in a data set, rendering supportive techniques unnecessary. In general, however, there are benefits to be drawn from regarding the two modes as complementary and mutually reinforcing." (Dempster, 1971, pag 325).

c: Mathematische statistiek is, ook voor data analytici, de moeite waard om te bestuderen. Uitspraken over de toepasbaarheid van mathematisch-statistische resultaten in de data analyse moeten met de nodige voorzichtigheid gedaan worden, dit wordt al duidelijk bij een oppervlakkige bestudering van de geschiedenis van de toegepaste wiskunde in het algemeen. Als statistiek en data analyse hetzelfde zijn, een standpunt waarmee we sympatiseren, dan is de mathematische statistiek een onderdeel van de data analyse, hoe ver verwijderd van de praktijk de resultaten soms ook lijken.

d: Geometrische en algebraische aspekten van de data analytische methoden zijn minstens even belangrijk als probabilistische.

e: Statistiek kan dikwijls zeer nuttig gebruikt worden als een vorm van sensitiviteits-analyse voor data analytische procedures. Met name de gebruikelijke asymptotische theorie, en technieken als de "Jackknife" (bv Miller, 1974), de "Bootstrap" (Efron, 1979), en kruisvalidatie (bv Stone, 1974) zijn hiervoor uitermate geschikt.

f: Het statistisch model kan, ook in situaties waarin het als beschrijving van de werkelijkheid volkomen faalt, gebruikt worden als test probleem of als normering van data analytische procedures.

g: Vanuit een wetenschap met relatief precieze kriteria is het o zo gemakkelijk arrogante uitspraken te doen over wetenschappen met minder precieze kriteria. Het heeft echter uitermate weinig zin om die precieze kriteria toe te willen passen in situaties waarbij aan de nodige vooronderstellingen in het geheel niet voldaan is. "A common feature of problems of this kind is that the stochastic model so central to classical statistical analysis is either absent altogether or is playing a very subordinate role. This, in my view, is inherent, and nothing could be more misguided than an attempt to force such problems into an ill-fitting classical statistical mould." (Sibson, 1972, pag 311). Dit maakt sommige gedeelten van "Methoden, voetangels en klemmen in de faktor analyse" zo misplaatst en zo "aanmatigend". Monopolie posities horen in de wetenschap niet thuis, en voor zover de statistiek ooit een dergelijke positie

gehad heeft, was die in de eerste plaats gebaseerd op de data analytische en niet op de inferentiele aspekten van de statistische procedures. "Taxonomists must find it infuriating that statisticians, having done so little to help them , laugh at their efforts. I hope taxonomists who have real and, I think, interesting problems find it equally funny that so much statistical work, although logically sound, and often mathematically complicated (and surely done for fun) has little or no relevance to practical problems." (Gower, in de diskussie van Cormack, 1971).

h: Statistiek is niet iets geheel anders dan data reduktie. Sir Fisher heeft tijdens zijn lange karriere vele malen benadrukt dat data reduktie de belangrijkste taak van de statistiek is. Klassieke statististici als Bartlett, Anscombe, Tukey, Mosteller, Kendall, en Kruskal zijn dezelfde opinie toegedaan. Het is daarom op z'n minst wat vreemd dat de auteurs van "Voetangels en klemmen" de term "data reduktie" op een geheel andere, wat neerbuigende, manier gebruiken. "Overigens heeft u voor data redultie niet veel aan dit rapport." (Bethlehem ea, lc pag 4). Inderdaad.

1.5 Kontinue en diskrete MVA:

1.5.1 De multinormaalverdeling

Zoals we in de bespreking van de MVA boeken gezien hebben gaan mathematische statistici er veelal zonder meer vanuit dat de multivariate gegevens multinormaal verdeeld zijn. Op het eerste gezicht lijkt dit een groot verlies in de mate van toepasbaarheid van de technieken tot gevolg te hebben. Is die indruk juist, en, zo ja, waarom wordt deze aanname dan zo gemakkelijk gemaakt ? In 1.1.3 noemden we de redenen die Anderson voor zijn keuze opgaf, we bespreken ze hier wat uitvoeriger.

a: In de praktijk dikwijls een goede beschrijving. Dat is nog maar de vraag. Anderson geeft als voorbeeld Galton's klassieke gegevens over de verdeling van lengte van vaders en zonen, en inderdaad zijn er in de antropometrie nog wel meer van dit soort voorbeelden te vinden. Maar we moeten niet vergeten dat Pearson bij bestudering van de even klassieke garnalen-gegevens van Weldon al direkt ontdekte dat de normaalverdeling in veel biometrisch materiaal helemaal niet opging, en dat hij daarom vervolgens zijn beroemde systeem van scheve verdelingen opstelde. Pearson zelf is wat meer genuanceerd, hoewel ook zijn formulering tegenwoordig als optimistisch beschouwd wordt. "On the basis of a very large experience of frequency curves and surfaces we have no hesitation in saying that up to the present time no distribution has been proposed which roundly represents experience so effectively as the Gaussian frequency. One of the present writers has indicated over and over again how it fails,

and he has measured the significance of its failure, but has always recognized that he must put against this the large percentage of cases in which it gives reasonable results, close enough for all practical purposes." (Pearson and Heron, 1913, pag). Als tweede argument tegen het multinormale optimisme van Anderson kunnen we aanvoeren dat "goodness-of-fit" toetsen voor de multinormaal verdeling weliswaar de laatste tijd wat aandacht krijgen, maar toch nog tamelijk primitief zijn. In ieder geval is normaliteit van de marginalen bij lange na niet voldoende om tot multinormaliteit te besluiten. En tenslotte heeft de aanname helemaal geen zin bij gekategoriseerde gegevens of bij ordinale variabelen. Weliswaar is het soms nuttig de filosofie van Pearson, dat diskrete variabelen altijd gediskretiseerde kontinue variabelen zijn, zo lang mogelijk vol te houden (vergelijk MacKenzie, 1978), maar zelfs als men dat gelooft zijn er voor gekategoriseerde multinormale gegevens nog altijd andere technieken nodig dan voor kontinu gemeten normaal-verdelingen.

b: De centrale grenswaarde stelling. In de eerste plaats is de aanname dat de variabelen als sommen van een groot aantal onafhankelijke kleine effekten opgevat kunnen worden in veel situaties niet erg voor de hand liggend. Misschien nog wel in de biometrische genetika, maar veel minder in de sociologie en de ekonomie. In de tweede plaats weten we dat konvergentie naar de normaalverdeling erg langzaam kan zijn, met name als de op te tellen komponenten fink scheef verdeeld zijn. En tenslotte is de normaalverdeling helemaal niet de enige verdeling die als gevolg van het centrale grenswaarde effekt als asymptotische verdeling kan optreden, er is een hele klasse van zogenaamde stabiele verdelingen die in aanmerking komt.

c: Eenvoudige formules en komplete theorie. Dit is in feite het belangrijkste argument. Dat de theorie voor de multinormaalverdeling zo kompleet is, komt natuurlijk niet alleen door de eenvoudige formules, maar ook door de zeer grote hoeveelheid aandacht die er in de loop der tijd om de redenen (a) en (b) aan de multinormaal-verdeling gespandeerd is. Eenvoudigheid van formules is maar betrekkelijk, berekenen van een tetrachorische korrelatie is bijvoorbeeld bepaald niet eenvoudig, berekenen van de eksakte verdeling van de eigenwaarden van een multinormale kovariantiestruktuur evenmin. Maar omdat iedereen met die multinormaal verdeling bezig is, is er ook al veel van het gekompliceerde rekenwerk gedaan. En dan inklusief Monte Carlo studies, die natuurlijk in principe even zo

goed in andere situaties toegepast kunnen worden.

Niettemin is het wel degelijk waar dat de multinormaalverdeling een groot aantal uiterst aantrekkelijke theoretische eigenschappen heeft, die de taak van de mathematische statistikus aanzienlijk vereenvoudigen. We noemen de eigenschappen die voor onze doeleinden het belangrijkste zijn.

c1: Als $\underline{x}$ multinormaal is met gemiddelde μ en kovariantiematriks Σ, en A is een arbitraire matriks, dan is $A\underline{x}$ multinormaal met gemiddelde $A\mu$ en kovariantiematriks $A\Sigma A'$. Omdat MVA veel met lineaire kombinaties van variabelen werkt, is deze eigenschap natuurlijk van onschatbare waarde.

c2: Stel $(\underline{x},\underline{y})$ is multinormaal met gemiddelden (μ_x,μ_y) en een kovariantiematriks

$$\begin{vmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{vmatrix}.$$

Dan is de konditionele verdeling van $\underline{x}$, gegeven dat $\underline{y} = y$, multinormaal met gemiddelde $\mu_{\underline{x}_1} - \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$ en met kovariantiematriks $\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$. De eerste eigenschap van de gemiddelden noemen we <u>lineäire regressie</u>, de tweede eigenschap (de kovariantiematriks is onafhankelijk van de waarde van y) noemen we <u>homoscedasticiteit</u>.

c3: Als voor i=1,...,n de vektoren $\underline{x}_i$ onafhankelijk zijn, met allemaal dezelfde multinormaalverdeling, gemiddelde μ en kovariantiematriks Σ, dan zijn $\underline{m} = \frac{1}{n}\Sigma\underline{x}_i$ en $\underline{S} = \frac{1}{n}\Sigma(\underline{x}_i - \underline{m})(\underline{x}_i - \underline{m})'$ <u>onafhankelijke</u> variabelen. Bovendien zijn $\underline{m}$ en $\underline{S}$ <u>voldoende statistieken</u> voor μ en Σ, en de schatters met <u>maksimale aannemelijkheid</u> van μ en Σ. Of, om het anders te zeggen, bij veel verdelingen zijn momenten en produkt-momenten gekompliceerde funkties van de parameters, zijn "optimale" schatters moeilijk te berekenen, en zijn de parameters lastig te interpreteren. Bij de multinormaalverdeling kan het als het ware niet eenvoudiger wat dit betreft, en bevatten de eerste orde momenten $\underline{m}$ en de tweede orde produkt-momenten $\underline{S}$ in de technische zin van het woord alle informatie die in de steekproef betreffende μ en Σ aanwezig is.

c4: Multinormaal verdeelde variabelen $\underline{x}$ zijn <u>onafhankelijk</u> als en alleen als ze <u>ongekorreleerd</u> zijn, dat wil zeggen als en alleen als de kovariantiematriks diagonaal is. In het algemeen is ongekorreleerdheid alleen maar een <u>noodzakelijke</u> voorwaarde voor onafhankelijkheid, en alweer vinden we dus dat de afhankelijkheid en samenhang van multinormale variabelen geheel beschreven kan worden in termen van de kovarianties.

c5: De normaalverdeling heeft een nauwe band met de Euklidische meetkunde. Punten met gelijke waarschijnlijkheidsdichtheid liggen bijvoorbeeld op ellipsen met de vektor van gemiddelden als middelpunt, zodat waarschijnlijkheidsdichtheid en (gewogen) Euklidische afstand direkt in elkaar te vertalen begrippen zijn. Als er twee multinormaalverdelingen met gelijke dispersiematriks gedefinieerd zijn, dan worden de punten waar de eerste dichtheid groter is dan de tweede van de overige punten gescheiden door de gewogen middelloodlijn tussen de vektoren van gemiddelden.

De onder c1-c5 genoemde eigenschappen zijn niet alleen statistisch van belang, het is duidelijk dat dit soort eenvoudige eigenschappen ook data-analytisch uiterst interessant zijn. Niettemin blijft in veel situaties de aanname van multivariate normaliteit verdacht, en bovendien moeilijk te toetsen. In niet geringe mate hangt dit ook samen met de eigenschappen van sommige veel gebruikte statistische toetsen. We toetsen over het algemeen in MVA een parametrische hypotese H_0 binnen een meer algemene hypothese H_1, en H_1 bevat bijna altijd de aanname van multivariate normaliteit. De gebruikelijke toetsprocedures zorgen ervoor dat H_1 buiten schot blijft. We toetsen alleen of we de nadere specifikatie H_0 toe kunnen laten, zonder ons het hoofd te breken over de vraag of H_1 eigenzelf zelf wel waar is. Samenvattend moeten we het zeker eens zijn met de volgende uitspraak. "Theorists of multivariate analysis clearly need to venture away from multivariate normal models." (Dempster, 1971, pag 317).

In de univariate statistiek heeft men de eens zo oppermachtige normaalverdeling al veel eerder losgelaten. Hierbij zijn twee verschillende wegen bewandeld. Men heeft voor sommige procedures (zoals de t-toets) laten zien dat ze behoorlijk robuust zijn, en dus bestand tegen flinke afwijkingen van normaliteit. En in de tweede plaats heeft men een groot aantal niet-parametrische procedures ontwikkeld, die in het geheel geen normaliteit aannemen. In MVA zijn beide wegen vrijwel onbewandeld. Men weet weinig over de robuustheid van MVA procedures, en wat men weet is niet altijd bemoedigend. De gebruikelijke kovariantieschatters zijn bijvoorbeeld uitermate gevoelig voor uitbijters. En zoals we gezien hebben, bijvoorbeeld bij de bespreking van het boek van Kendall, is niet-parametrische MVA onvoldoende ontwikkeld. Weliswaar zijn er een aantal varianten van de meer eenvoudige

multivariate toetsen ontwikkeld (zie bijvoorbeeld Puri en Sen, 1971), maar de eigenschappen van deze procedures zijn minder bevredigend dan in het univariate geval, en (bovenal) hun data analytische waarde is gering (omdat ze, per definitie, geen geometrische interpretaties met zich meebrengen).

1.5.2 Tabellaire analyse

In de sociologie, politikologie, en aanverwante wetenschappen speelt de enquete een grote rol. Aan een groot aantal mensen stelt men een groot aantal vragen, de antwoorden vormen, eventueel samen met achtergrondsgegevens van de ondervraagden, het data-bestand. Variaties op dit thema zijn vragenlijsten voor klinische doeleinden, attitude studies, panel studies, multiple choice tests, enzovoorts. Over het algemeen zijn kontinu varierende numerieke variabelen bij dit soort onderzoek een uitzondering. Numerieke variabelen worden in brede kategorieen verdeeld, antwoordkatego-rieen van vragen zijn dikwijls ordinaal, en achtergrondsgegevens (zoals religie, politieke keuze) kunnen heel goed nominaal zijn.

Men was al vroeg doordrongen van het feit dat het aannemen van normaliteit in dit soort situaties geen enkele zin had; in plaats van de methoden van Pearson om associatie te meten koos men systematisch voor de methoden van Yule (vergelijk MacKenzie, 1978). Zoals we eerder gezien hebben geloofde Pearson dat alles in deze wereld kontinue op de een of andere schaal varieerde, en dat associatiematen bedoeld zijn om de korrelatie tussen de onder-liggende kontinue variabelen te schatten. Dit is, in laatste instantie, een metafysisch uitgangspunt, wat ook verklaart waarom Pearson nooit de "diskrete" leer van Mendel heeft kunnen aanvaar-den (Norton, 1976, 1978). Pearson's uitgangspunt heeft grote invloed gehad in de psychometrie, met name omdat veel mensen dachten dat alleen kontinue variatie "meetbaar" was, en dat dus alleen aannemen van kontinue onderliggende variatie de psychologie kon verheffen tot het nivo van de echte wetenschappen zoals de fysika. Tetrachorische korrelatie, bijvoorbeeld, wordt bijna uitsluitend in de psychometrie gebruikt, en nog steeds verschijnen er artikelen met rekenmethoden die het mogelijk maken deze koefficient sneller of beter te berekenen. Yule, daarentegen, had geen last van dit soort metafysische vooroordelen. Voor hem was iets een associatiemaat als het +1 was bij perfekte positieve samenhang, -1 bij perfekte negatieve samenhang, en 0 bij onaf-hankelijkheid. Op basis van dit meer aksiomatische uitgangspunt stelde hij een aantal maten voor, die aan deze eisen voldeden.

Ze worden opgesomd, met diverse uitbreidingen en een lawine van interessante details, in de befaamde artikelen van Goodman en Kruskal (1954, 1959, 1963, 1972).

De sociologen hebben nooit in dezelfde mate als de psychologen het kompleks gehad dat hun wetenschap naar het model en met de methoden van de meer eksakte wetenschappen opgebouwd moest worden. Ze hebben ook nooit het kompromisloze empirisme en de bijbehorende korrelatiomanie van de Pearson-Spearman school overgenomen. De techniek die in de sociologische data analyse populair werd was die van de kruistabel met bijbehorende associatiematen. Aanvankelijk werkte dit bevredigend, om voor de hand liggende redenen. "Many tried and tested techniques of multivariate data analysis were invented at a time when ten was a typical number of variables in an ambitious data collection program." (Dempster, 1971, pag 336). Met tien variabelen zijn er 45 kruistabellen, en dat is allemaal nog te overzien. Maar in de tegenwoordige surveys zijn 100 vragen heel normaal, de MMPI heeft ongeveer 700 vragen, en in longitudinale studies kan dat aantal zelfs nog overtroffen worden. De komputer kwam te hulp, er werden pakketten zoals CROSSTABS gekonstrueerd, die alle kruistabellen keurig afdrukten, vergezeld van een hele lijst van associatiematen. Dat zijn die hele dikke pakken output, die je soms nog wel ziet liggen. Ze zien er over het algemeen nogal beduimeld uit, en zijn vaak echte stofnesten. "Si l'on voit aujourd'hui sortir des imprimantes de longues listes de ces petits tableaux, c'est trop de chercheurs, notamment dans les sciences humaines, n'ont pas accordé leurs méthodes à la puissance des nouveaux outils de calcul: ils sont semblables à un ingenieur qui, pour en bâtir un pont, dessinerait des blocks d'acier ayant la forme de pierres de taille." (Benzecri e.a. 1973, pag 11).

In de eerste plaats kunnen 5000 kruistabellen niet opgenomen worden in een onderzoeksrapport, en dus moet men selekteren. Men selekteert wat men interessant vindt, en dat is nogal subjektief. Het gevolg is dan ook, dat anderen met geheel andere (en misschien tegengestelde) interesses geheel andere verbanden in het materiaal kunnen vinden (als ze het tenminste te pakken kunnen krijgen). En in de tweede plaats gaat men zich bij een ellenlange lijst kruistabellen min of meer automatisch afvragen wat die tabellen nu eigenlijk met elkaar te maken hebben. Men krijgt het idee dat rapporteren van geselekteerde kruistabellen vaak een overdreven

indruk geeft van de verbanden, en op zijn minst een uiterst onsystematische indruk (vergelijk Hirshi en Selvin, 1973). Even een vergelijking er tussen door. Als men heel veel numerieke variabelen heeft kan men een hele grote korrelatiematriks uitrekenen. Volgens de methoden van de tabellaire analyse is de volgende stap dan dat de individuele korrelaties, die men interessant vindt, stuk voor stuk besproken worden. Het ligt echter meer voor de hand om naar methoden te zoeken om de korrelatiematriks op een kompakte manier te beschrijven, en dan bij voorkeur ook nog methoden die er niet van uitgaan dat de geobserveerde korrelaties onafhankelijke statistieken zijn die geheel los van elkaar beschreven kunnen worden. Een andere gewoonte is om alleen de signifikante verbanden te beschrijven. Daarbij moet men er dan natuurlijk wel rekening mee houden dat bij 5000 tabellen er, op basis van het toeval alleen, al gauw 250 een signifikant verband op 5% nivo vertonen.

De sociologische methodologen hebben twee uitwegen gevonden uit de puinhopen van de tabellaire analyse. De eerste is de analyse van meer-dimensionale kruistabellen, meestal door middel van de zogenaamde log-lineaire modellen. Boeken die die technieken beschrijven zijn Haberman (1974), Bishop e.a. (1975), en Gokhale en Kullback (1978). We komen hier terecht bij de eerder genoemde diskrete MVA. Een tweede uitweg is de zogenaamde "kausale analyse", uitvoerig besproken in bijvoorbeeld Blalock (1964) en Boudon (1967). We bespreken deze twee vrij recente ontwikkelingen in aparte paragrafen.

1.5.3 Diskrete MVA

Een van de nadelen van tabellaire analyse was dat het niet duidelijk werd wat de diverse kruistabellen nu eigenlijk met elkaar te maken hadden. Hierdoor was het mogelijk dat er allerlei verbanden van het bekende soort, zoals het positieve verband tussen het aantal geimporteerde bananen en het aantal onwettige geboortes, gevonden werden. De oplossing die hiervoor in diskrete MVA bedacht is, richt zich op de analyse van meerdimensionale kruistabellen, dat wil zeggen we bekijken drie of meer variabelen tegelijkertijd en bestuderen de bijbehorende meer-dimensionale array van frekwenties. We noemen eerst even de in het oog springende nadelen van deze benadering.

Als we een groot aantal variabelen hebben, dan kunnen we daaruit een zeer groot aantal meer-dimensionale kruistabellen vormen. Het probleem van de selektie van tabellen wordt daardoor alleen maar

ernstiger. Bij tien variabelen zijn er 45 twee-dimensionale, 120 drie-dimensionale, 210 vier-dimensionale tabellen, enzovoorts. Er is natuurlijk maar één tien-dimensionale tabel, en het is dus in principe mogelijk om alleen die tabel zo volledig mogelijk te analyseren. Maar helaas, in dat geval krijgen we te maken met het tweede nadeel van diskrete MVA. Als iedere variabele vier kategorieen heeft, dan heeft de tien-dimensionale tabel 4^{10}, dus meer dan één miljoen, cellen. Het aantal observaties zal over het algemeen veel kleiner zijn, en dus zal veruit het grootste deel van de cellen leeg zijn. Omdat de inferentiele kant van diskrete MVA gebaseerd is op de asymptotische normaalverdeling van de frekwenties is het echter nodig dat de tabel redelijk gevuld is. Volgens de klassieke, hoewel wat arbitraire, regel van Cochran willen we gemiddeld vijf observaties per cel, dat wil zeggen we willen in ons kleine voorbeeld meer dan vijf miljoen observaties. Als er veel variabelen zijn is de analyse van de totale tabel niet mogelijk, terwijl selekteren van kleinere tabellen uit de totale tabel op ongelooflijk veel manieren kan gebeuren. In die zin is diskrete MVA vooral nuttig wanneer we het verband tussen drie of vier variabelen willen analyseren, hetzij in situaties waarin we maar drie of vier variabelen hebben, hetzij wanneer we redenen hebben om drie of vier variabelen uit het totaalonderzoek uitermate belangrijk te vinden. Voor een gelijktijdige analyse van een groot aantal kategorische variabelen is diskrete MVA niet geschikt.

We hebben gezien dat Roy (1957) een van de eersten was die de diskrete MVA besprak in zijn handboek, en dat hij tevens deze benadering veel realistischer vond dat multinormale MVA. Wat zijn de belangrijkste verschillen ? In MVA interesseren we ons in het algemeen voor de afhankelijkheid en samenhang van variabelen. Afhankelijkheid en samenhang zijn uiteindelijk eigenschappen van de waarschijnlijkheidsverdeling van de variabelen. Ze kunnen op verschillende manieren gedefinieerd worden. Roy gebruikt, in navolging van Bartlett en Fisher, definities in termen van konditionele waarschijnlijkheden en van de produktregel voor onafhankelijke waarschijnlijkheden. Dit is de <u>multiplikatieve</u> benadering. Er bestaat ook een <u>additieve</u> benadering van samenhang en afhankelijkheid, voornamelijk gepropageerd door Lancaster. De relatie tussen de twee benaderingen is eenvoudig te omschrijven: multiplicatieve analyse is hetzelfde als additieve analyse van de logarithmes van de waarschijnlijk-

heidsverdelingen.

We bespreken kort een klein voorbeeld om te laten zien hoe diskrete MVA te werk gaat. Mensen die op dit moment geen zin hebben in formules en zo kunnen dit gerust overslaan, en verder gaan bij 1.5.4. De rest veronderstelt met ons dat o_{ijk} een diskrete drie-dimensionele verdeling is, met dimensies k, ℓ, m en met marginalen p_i, q_j, r_k. Stel x_{is}, y_{jt}, en z_{ku} zijn ortonormale funkties op de marginalen, dat wil zeggen X, Y, en Z zijn matrices van de orde k, ℓ, m met kolommen die voldoen aan

$$\sum x_{is}x_{is'}p_i = \delta^{ss'},$$
$$\sum y_{jt}y_{jt'}q_j = \delta^{tt'},$$
$$\sum z_{ku}z_{ku'}r_k = \delta^{uu'}.$$

We veronderstellen bovendien dat $x_{i1} = y_{j1} = z_{k1} = 1$, alle eerste kolommen van X, Y, en Z zijn dus konstant, en gelijk aan één. Bij additieve diskrete multivariate analyse vormen we nu de sommen

$$\alpha_{stu} = \sum\sum\sum x_{is}y_{jt}z_{ku}o_{ijk},$$

en bij multiplikatieve diskrete MVA berekenen we de sommen

$$\beta_{stu} = \sum\sum\sum x_{is}y_{jt}z_{ku}p_iq_jr_k.\ln o_{ijk}.$$

De hypothesen in diskrete MVA zijn nu dat bepaalde, over het algemeen sistematisch gekozen, groepen van de α_{stu} of β_{stu} gelijk aan nul zijn. Als eerste voorbeeld bekijken we de hypothese dat i onafhankelijk is van j en k. Gebruik makend van konditionele waarschijnlijkheden kunnen we schrijven $o_{i|jk} = p_i$, of nog anders $o_{ijk} = p_io_{jk}$. Een ekwievalente formulering is dat $\alpha_{stu} = 0$ wanneer $s \neq 1$, een andere is dat $\beta_{stu} = 0$ wanneer $s \neq 1$. Als tweede voorbeeld nemen we konditionele onafhankelijkheid van i en j gegeven k, dus $o_{ij|k} = o_{i|k}o_{j|k}$ oftewel $o_{ijk} = o_{ik}o_{jk}/r_k$. Een ekwievalente formulering is dat $\beta_{stu} = 0$ wanneer zowel $s \neq 1$ als $t \neq 1$. Het is niet mogelijk deze hypothese op een eenvoudige manier in termen van de α_{stu} te formuleren.

In het algemeen is het zo dat de gebruikeljke hypothesen eenvoudiger in termen van de β's dan in termen van de α's vertaald kunnen worden. Additieve analyse heeft weer andere voordelen, de twee benaderingen worden vergeleken in Darroch (1974), Lancaster (1971, 1975). Ook voor de kontinue normaalverdeling is de multiplikatieve benadering het meest informatief. In de voor de hand liggende generalisatie van diskrete MVA gebruiken we de drie-dimensionale dichtheid p(xyz), en orthogonale polynomen op de normaalverdeling $\phi_s(x)$ van de graad s=0,1,2,...

In dat geval vinden we bijvoorbeeld $\beta_{stu} = 0$ als $s \geq 1$, $t \geq 1$, en $u \geq 1$. Dit noemt men wel het ontbreken van tweede-orde interaktie. Op dezelfde manier vinden we dat in meer dan drie dimensies in de multinormaalverdeling alle hogere orde interakties gelijk aan nul zijn, wat keurig overeenkomt met onze eerdere konstatering dat alle afhankelijkheid in de multinormale verdeling beschreven wordt door de eerste-orde interakties, dat wil zeggen door de kovarianties. We zien dus dat in het algemeen diskrete MVA niet-lineaire funkties op de marginalen definieert om samenhang en afhankelijkheid te onderzoeken, terwijl klassieke lineaire MVA volstaat met lineaire funkties. Zoals we aan het voorbeeld van de multinormaalverdeling hebben kunnen zien bestaat er een kontinue versie van diskrete MVA, die we niet-lineaire MVA kunnen noemen. We bekijken daarbij een basis voor de niet-lineaire funkties op de marginalen, en definieren de reeksen α_{stu} en/of β_{stu}. In het kontinue geval gebruiken we een basis met oneindig veel elementen. Doordat we een basis gebruiken bekijken we in feite alle mogelijke funkties tegelijkertijd, niet-lineaire MVA gebruikt daarom globale chi-kwadraat toetsen die voor groepen α's of β's bekijken of ze allemaal gelijk aan nul zijn. Verderop zullen we laten zien hoe de in deze klapper besproken vorm van niet-lineaire MVA samenhangt met de diskrete MVA uit deze paragraaf.

1.5.4 Kausale analyse

Kausale analyse komt uit een geheel andere historische hoek dan tabellaire analyse. In de biometrische genetika overheersten lange tijd de zuiver deskriptieve methoden van Pearson. De basis hiervoor was de wetenschapsfilosofie van Pearson, die aannam dat korrelatie een meer fundamentele kategorie was dan kausatie, en dat kausatie het grensgeval van perfekte korrelatie was. Het is niet nodig om naar kausale verbanden te zoeken, je hoeft alleen maar korrelaties te berekenen. De theorie volgt dan vanzelf, want een wetenschappelijke theorie is alleen maar een korte, overzichtelijke samenvatting van een groot aantal empirische gegevens, in dit geval van een groot aantal korrelaties. Dit standpunt is inmiddels verlaten, en wel om drie redenen. In de eerste plaats werkt het niet, zoals psychometrische 'theorieen' over intelligentie overduidelijk aangetoond hebben. Het aantal korrelaties dat berekent is loopt ondertussen ongetwijfeld in de miljarden, maar uit deze ongelofelijke berg getallen is nog geen theorie van enige algemeenheid te voorschijn gekomen. In de tweede plaats toonde met name Yule

aan dat korrelatierekening bep erkt toepasbaar was. Als we tijdreeksen korreleren vinden we dikwijls nonsens-korrelaties. En tegelijkertijd schoten de bekende voorbeelden, die aan willen tonen dat korrelatie geen oorzakelijk verband impliseert, als paddestoelen uit de grond. De derde reden is van meer filosofische aard. Kausaliteit is asymmetrisch, impliseert een richting, en dikwijls een temporele volgorde. Korrelatie daarentegen is symmetrisch. De nonsens korrelaties deden sommige mensen er toe overgaan om te beweren dat korrelatie alleen geinterpreteerd kan worden binnen een kausaal model. Met name sociologen vinden dit een zeer aanvaardbaar standpunt, omdat ze van oudsher niets willen weten van het rabiate empirisme van Pearson en de psychometrici. Niettemin kwamen de eerste aanzetten tot kausale analyse zoals we het nu kennen uit de biometrische genetika, en meer in het bizonder uit het werk van Sewall Wright.

Kausale analyse werd in eerste instantie voor kontinu varierende multivariate gegevens opgesteld. De gepostuleerde afhankelijkheidsrelaties worden samengevat in een pijldiagram, de pijlen in het diagram worden geinterpreteerd als lineaire relaties tussen variabelen. We geven in figuur 2a een klein voorbeeld. Er zijn drie eksogene variabelen (daaruit vertrekken alleen pijlen) $\underline{v}$, $\underline{z}$, en $\underline{w}$, en twee endogene variabelen $\underline{x}$ en $\underline{y}$. De vier paden in de figuur hebben waarden a, b, c, d. Van de eksogene variabelen nemen we hier aan dat ze onafhankelijk zijn. Een mogelijke interpretatie is dat $\underline{x}$ en $\underline{y}$ allebei metingen van een onderliggende variabele $\underline{z}$ zijn, met meetfouten respektievelijk $\underline{v}$ en $\underline{w}$. De lineaire relaties zijn

$\underline{x} = b\underline{z} + a\underline{v}$,

$\underline{y} = c\underline{z} + d\underline{w}$.

Als we aannemen dat $\underline{v}$, $\underline{z}$, en $\underline{w}$ allemaal verwachte waarde nul en variantie één hebben, dan geldt

$E(\underline{xy}) = bc$,

$E(\underline{xz}) = b$,

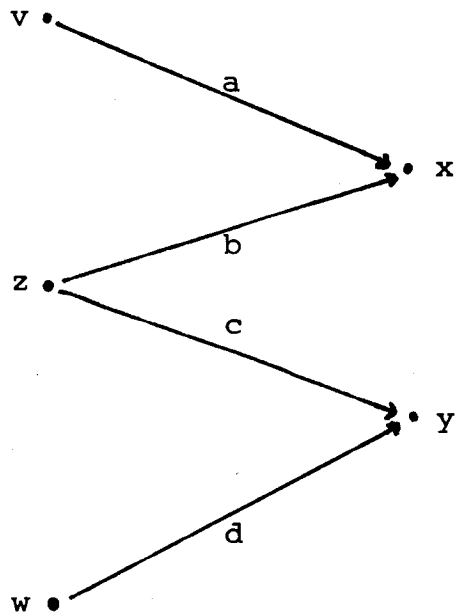
$E(\underline{yz}) = c$,

$E(\underline{x}^2) = b^2 + a^2$,

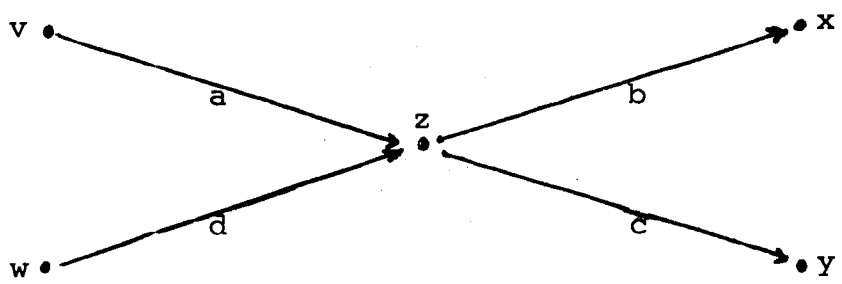
$E(\underline{y}^2) = c^2 + d^2$.

Dit impliseert ondermeer dat de korrelatie tussen $\underline{x}$ en $\underline{y}$, met $\underline{z}$ uitgepartialiseerd, gelijk is aan nul. Het is duidelijk dat er met dezelfde vijf variabelen een groot aantal andere pijldiagrammen getekend kan worden. Eén daarvan staat in figuur 2b.

Dit maakt het grote probleem van de kausale analyse duidelijk. In



Figuur 2a: Een eenvoudig kausaal model.



Figuur 2b: alweer een eenvoudig kausaal model.

plaats van het tabelselektieprobleem hebben we nu het probleem van de modelselektie. En dit probleem is natuurlijk weer ernstiger, naarmate we meer variabelen hebben. We moeten bovendien niet uit het oog verliezen dat een model min of meer automatisch oorzakelijke interpretaties met zich meebrengt (in figuur 2a veroorzaakt z zowel x als y), en dat de korrelaties en regressiekoefficienten altijd binnen het gekozen model worden geinterpreteerd. Als we een ander model gekozen zouden hebben, zouden dezelfde statistieken anders geinterpreteerd worden. Het interpretatieprobleem, in de tabellaire analyse dus het relateren van de tabellen aan elkaar, is naar een a priori nivo verschoven, maar daardoor geenszins opgelost. Weliswaar hebben we, met wat additionele assumpties, een modeltoets tot onze beschikking, maar die additionele assumpties zijn zelden realistisch, en het onderscheidend vermogen van dit soort toetsen is bij veel variabelen uiterst gering. In de biometrische genetika legt de Mendeliaanse theorie beperkingen op aan de keuze van modellen, tenminste in relatief eenvoudige situaties die gedeeltelijk onder eksperimentele kontrole staan. Bij onderwerpen als bijvoorbeeld over erving van intelligentie (wat dat ook mag zijn) hebben we niets aan genetische theorie en is de modelkeuze daardoor vrij arbitrair, met het katastrofale gevolg dat dezelfde basisgegevens tot de meest uiteenlopende interpretaties kunnen leiden. Op dezelfde manier is er ekonomische theorie, die beperkingen oplegt aan de modelkeuze bij ekonometrische kausale modellen. Maar in beide situaties omvatten de kausale modellen relatief weinig variabelen, er zijn er desondanks nog genoeg kontroverses over de modelkeuze.

In de moderne versies van kausale modellen, die zich zowel op de biometrische genetika als op de psychometrie inspireren, gebruiken we zelfs "latente" of "onmeetbare" variabelen om het model uit te breiden. Genotype is bijvoorbeeld vrijwel altijd niet meetbaar, evenals algemene intelligentie. In figuur 2a spelen v en w de rol van onmeetbare meetfouten, en speelt z de rol van de algemene intelligentie; we kunnen alleen veronderstellen dat deze variabelen ongekorreleerd zijn, maar we kunnen ze nooit direkt meten. Met deze interpretatie zegt figuur 2a dat x en y in een één-faktor model passen. Invoeren van latente variabelen maakt het probleem van de modelkeuze nog groter dan het al was. De modeltoets wordt nog meer overbelast, om redenen die we al eerder genoemd hebben, maar ook omdat modeltoetsen

nauwelijks kunnen helpen bij het kiezen tussen modellen. Met name kunnen er altijd "betere" modellen blijven bestaan, die we nog niet onderzocht hebben, te meer omdat veel onderzoekers zich beperken tot kleine variaties binnen één model. Interpreteerbaarheid van de resultaten is een uiterst onbetrouwbaar kriterium, hier nog meer dan ergens anders, omdat de interpretatie al in het stadium van de modelkeuze is vastgelegd.

Samenvattend beschouwen we kausale analyse als een lofwaardige poging om a priori informatie over de variabelen (bijvoorbeeld hun volgorde in de tijd) te gebruiken in MVA. Als zodanig is het een nuttige generalisatie van het simpele onderscheid tussen analyse van afhankelijkheid en analyse van samenhang. Wij vinden het terecht dat rationalistische overwegingen gebruikt worden om het onhoudbare empiristische optimisme van de psychometrici aan te vullen. Niettemin zijn er in de sociale wetenschappen zeer veel situaties waarin de keuze van het kausale model zeer willekeurig is, omdat de nodige a priori kennis ten enenmale ontbreekt of hoogstens zeer fragmentair is. Dit is met name het geval wanneer er veel variabelen zijn. Vergelijk ook onze diskussie van de beperkte toepasbaarheid van het statistische model in het algemeen. In dit soort situaties is het misleidend om de resultaten te interpreteren in termen van de parameters van het kausale model, de modelkeuze was arbitrair en dus is de interpretatie dat ook. Het heeft weinig zin om in ongestruktureerde situaties hoog gestruktureerde modellen te gebruiken. De subjektieve keuzes uit de tabellaire analyse worden verschoven naar het modelkeuze stadium, en blijven verder buiten schot. Een programma pakket als LISREL is zeker een stap vooruit vergeleken met CROSSTABS, de stapels output worden tenminste aanzienlijk dunner, dat wil zeggen er is meer <u>data-reduktie</u> toegepast. Maar onkritische interpretatie van LISREL-resultaten is eerder mogelijk dan onkritische interpretatie van CROSSTABS resultaten. Tenslotte vermelden we nog dat de lineariteit van kausale analyse geen serieuze beperking is. We kunnen de verworvenheden van diskrete MVA ook in de kausale analyse toepassen, en er is niets wat ons verhindert de pijlen in figuur 2a niet-lineair op te vatten. We krijgen dan $\underline{x} = f(\underline{z},\underline{v})$ en $\underline{y} = g(\underline{z},\underline{w})$, met dezelfde aanname over onhankelijkheid van de eksogene variabelen.

1.6 Definitie van MVA

Op basis van de diskussie in de voorafgaande paragrafen kunnen we nu proberen een definitie van MVA te geven. In de oorspronkelijke en meest beperkte omschrijving is MVA de analyse van willekeurige steekproeven uit een multinormaalverdeling. De gegevens worden verzameld in een matriks met n rijen en m kolommen. De rijen van de matriks zijn onafhankelijke replikaties van dezelfde multinormale m-vektor. We hebben gezien dat zowel de aanname van onafhankelijkheid tussen rijen als de aanname van multinormaliteit binnen rijen te restriktief zijn om een algemene theorie van MVA op te zetten.

Een eerste generalisatie (gesuggereerd door het werk van Van de Geer, Caillez en Pages, en Green en Carroll) zou dus kunnen zijn dat MVA de analyse is van een arbitraire rechthoekige matriks, waarbij het eksplisiete doel van de analyse het beschrijven van de matriks in termen van een kleiner aantal parameters is. Over de oorsprong van de matriksen doen we geen uitspraak. Merk op dat meerdimensionele schaalmethoden en de meeste vormen van kluster analyse onder deze definitie van MVA vallen. Daar is op zichzelf niets tegen, en de groep rond Krishnaiah en het Journal of Multivariate Aanalysis besteedt inderdaad ruim aandacht aan deze niet-statistische vormen van MVA. Niettemin is deze definitie misschien wat te algemeen. De naam MVA impliseert in ieder geval dat we te maken hebben met een aantal dingen, die we variabelen noemen. In de matriks-definitie worden rijen en kolommen van de matriks symmetrisch behandeld, en is er geen sprake van variabelen. Misschien zou hiervoor de term meer-dimensionale data analyse meer op zijn plaats zijn.

Het is daarom nuttig om wat meer elementen uit de klassieke definitie over te nemen, met name de asymmetrische rol van rijen en kolommen. We willen daarbij tevens af van de beperking dat de gegevens uit reele getallen bestaan, omdat deze beperking bij diskrete MVA niet op hoeft te gaan. In plaats van een arbitraire n × m matriks nemen we als uitgangspunt ~~een~~ waarschijnlijkheidsruimte ~~en~~ m meetbare funkties op die ruimte met waarden in m andere waarschijnlijkheidsruimten. De variabelen zijn dus stochastische variabelen. In het eenvoudigste geval heeft de ruimte waarop de variabelen gedefinieerd zijn n elementen, en is de waarschijnlijkheidsmaat van een aantal elementen gelijk aan het aantal elementen gedeeld door n. De funkties kunnen worden gedefinieerd door alle n elementen op te schrijven, met voor ieder element de bijbehorende m funktiewaarden. Deze lijst kan ondergebracht worden in een (niet noodzakelijk numerieke) n × m matriks. Het waarschijnlijkheids-

element in de definitie speelt in feite geen rol, we kunnen nog steeds arbitraire rechthoekige matriksen bekijken, maar de benadering in termen van m funkties op eenzelfde ruimte heeft de nodige asymmetrie gebracht. In statistische terminologie bekijken we in dit geval de hele populatie, en die populatie is eindig en volledig bestudeerd. Het spreekt vanzelf dat we ook een eindige populatie met een arbitraire diskrete waarschijnlijkheidsverdeling over de n elementen kunnen bekijken, maar dan is de waarschijnlijkheids-komponent van de gegevens niet meer triviaal, en hebben we dus informatie die niet in de data matriks weergegeven is.

Het kan echter ook zo zijn dat de ruimte waarop de variabelen gedefinieerd zijn oneindig veel elementen heeft, en dat we onze funkties niet kunnen definieren door het geven van een eksplisiete lijst van waarden. We blijven de populatie bestuderen, maar de populatie is nu oneindig. In plaats van de funktie eksplisiet weer te geven definieren we hem nu door zijn eigenschappen te noemen, bijvoorbeeld door aan te nemen dat de variabelen multi-normaal verdeeld zijn. Deze situatie is vanzelfsprekend uitsluitend van theoretisch belang, er is niets geobserveerd, er is geen data matriks. We doen geen data analyse, we doen ook geen statistiek.

In de derde situatie is er weer een n × m data matriks, en we veronderstellen dat de rijen een willekeurige steekproef van de grootte n vormen, dat wil zeggen de rijen van n zijn onafhankelijke realisaties van hetzelfde populatiemodel. De situatie is geheel anders als in de vorige twee gevallen, daar waren de stochastische variabelen gedefinieerd, maar niet geobserveerd. We kunnen ook zeggen dat de waarschijnlijkheidsmaat op de ruimte waarop de variabelen gedefinieerd zijn nu geobserveerd is, dat wil zeggen de waarschijnlijkheidsmaat is zelf een stochastische variabele geworden. We werken met de empirische waarschijnlijkheidsmaat in plaats van met de theoretische. Aannamen over de theoretische verdeling (dus over de populatie) hebben vanzelfsprekend wel konsekwenties voor de mogelijke waarschijnlijkheidsverdelingen die we kunnen observeren, op dezelfde manier als aannamen over de manier waarop de steekproef getrokken is konsekwenties hebben.

Op basis van onze analyse kunnen we nu de volgende definitie van MVA geven: MVA bestudeert systemen van gekorreleerde stochastische variabelen of willekeurige steekproeven uit dergelijke systemen. Merk op dat we bij deze definitie niet gespecificeerd hebben dat

het gaat om een eindig aantal variabelen. In deze klapper is dit steeds wel het geval, maar bij analyse van stochastische processen is het gebruikelijk om oneindig veel variabelen, korresponderend het oneindig veel tijdstippen, toe te laten. Het stochastische element is in onze definitie steeds aanwezig, maar we hebben gezien dat het in het geval van een eindige populatie triviaal gemaakt kan worden. Daardoor valt MVA in dat geval in feite samen met de analyse van arbitraire n × m matriksen.

We definieren nu specifieke vormen van MVA als volgt. Een MVA is lineair wanneer de resultaten van de analyse invariant zijn onder één-éénduidige lineaire transformaties van de stochastische variabelen. Een MVA is monotoon wanneer de resultaten van de analyse invariant zijn onder één-éénduidige monotone transformaties van de variabelen, en niet-lineair als de resultaten invariant zijn onder één-éénduidige meetbare transformaties. Deze definities zijn noodzakelijkerwijs wat vaag, omdat we niet specificeren wat we bedoelen met "de resultaten". Het is mogelijk dat een deel van de resultaten (de output van een komputerprogramma) bijvoorbeeld wel verandert en een ander deel niet, of zelfs dat alle resultaten veranderen maar op een eenvoudig aan te geven manier. Bij bespreking van de diverse technieken in deze klapper zal steeds duidelijk worden wat er nu invariant is, en wat niet. In deze zelfde weinig specifieke zin kunnen we nu al vast meedelen dat in deze klapper niet-lineaire varianten besproken worden van enige bekende lineaire MVA technieken.

We moeten ook nog een opmerking maken over de vorm van de data matriks. Tot nu toe hebben we steeds de n × m matriks gebruikt, geheel konform het gebruik in klassieke MVA. In diskrete MVA gebruikt men daarentegen de meerdimensionale kruistabel. Het verband tussen de twee representaties is echter nogal eenvoudig, het is niets anders dan het verband tussen de stochastische variabele en zijn verdelingsfunktie. In het diskrete geval, waarin de m variabelen de waarschijnlijkheidsruimte afbeelden in m eindige verzamelingen, kunnen we iedere mogelijke waarde van alle variabelen tezamen zien als een cel van een m-dimensionale kruistabel. Als deze tabel gegeven is, dan kunnen we de gegevens kompleet beschrijven door in iedere cel weer te geven hoe vaak het korresponderende profiel in de data matriks voorkomt. Dit geldt in het geval van een populatie, de celwaarden worden dan waarschijnlijkheden, maar ook in een steekproef, de celwaarden worden dan geobserveerde relatieve frekwenties. Als de variabelen waarden aannemen in een

bereik met oneindig veel elementen, dan vervangen we de kruistabel door het produkt van de m bereiken, en we vervangen de waarschijnlijkheden van de cellen door de multivariate waarschijnlijkheidsverdeling. Of, in het steekproefgeval, door de empirische waarschijnlijkheidsverdeling. Als in het populatiegeval de waarschijnlijkheidsverdeling opgevat kan worden als de integraal van een meer-dimensionale dichtheidsfunktie, dan kunnen we ook die dichtheidsfunktie als kontinu analogon van de kruistabel gebruiken.

Merk overigens op dat een steekproef altijd leidt tot een eindig aantal observaties, en dat we daarom zonder verlies van algemeenheid kunnen aannemen dat het bereik van de variabelen eindig is. Immers kontinue variabelen worden altijd met een zekere precisie gemeten, die leidt tot een representatie met een eindig aantal decimalen. Een grondidee achter de opzet van deze klapper is dat alle gegevens kategorisch (oftewel diskreet) zijn, en dat kontinue modellen voornamelijk gebruikt worden om de berekeningen te vereenvoudigen. Dit is in feite ook de manier waarop de normaalverdeling in de geschiedenis van de waarschijnlijkheidsleer het eerst te voorschijn komt. Pas veel later, als Galton en vooral Pearson kontinue variatie tot norm verheffen, gaat het standpunt overheersen dat diskrete variabelen een soort gedegenereerde of afgeronde kontinue variabelen zijn. En dit laatste standpunt speelt in de klassieke MVA, blijkens de inhoudsopgaven van de besproken boeken, nog steeds een grote rol.

De m stochastische variabelen definieren een m-dimensionale waarschijnlijkheidsverdeling. Deze verdeling heeft univariate marginalen, bivariate marginalen, enzovoorts. Er zijn nog steeds mensen die MVA univariaat aanpakken, dat wil zeggen dat ze technieken toepassen die dezelfde resultaten opleveren als ze op een andere multivariate verdeling met dezelfde univariate marginalen worden toegepast. Men is het er tegenwoordig wel over eens dat een dergelijke benadering uitermate misleidend kan zijn (Rao, 1960). We hebben gezien dat multinormale MVA typisch bivariaat is. Bij de multinormaalverdeling leidt dit niet tot verlies van informatie, bij meer algemene multivariate verdelingen echter wel. Ook tabellaire analyse is bivariaat, alleen tabellaire analyse is niet-lineair en multinormale analyse is lineair. De technieken die in deze klapper besproken worden zijn voor het grootste deel ook bivariaat, hoewel mogelijke uitbreidingen naar niet-bivariate analyse voor de hand liggen. In deze zin zijn onze

technieken een tussenvorm tussen multinormale en algemene multivariate analyse, we kombineren niet-lineariteit en bivariaatheid, dikwijls in de hoop dat de bivariate marginalen een groot deel van de informatie uit de multivariate verdeling bevatten. Door ons te konsentreren op bivariate marginalen ondervangen we het probleem van de ontelbare lege cellen, en het probleem van de moeizame interpretatie van hogere-orde interakties (ook bekend uit de variantie analyse). Door niet-lineaire transformaties toe te laten komen we los van veel van de beperkingen van de multinormale analyse. En door alle bivariate verdelingen in één enkele analyse op te nemen zijn we grotendeels af van het probleem van de model selektie. Door zo veel mogelijk data reduktie toe te passen zijn we af van de grote stapels onoverzichtelijke output. De nadruk ligt in deze klapper daarom op grote datasets met veel variabelen, efficiente rekenmethoden, op niet-lineaire transformaties en waar mogelijk op geometrische interpretaties. Het begrip "willekeurige steekproef" speelt een relatief ondergeschikte rol, en de multinormaalverdeling wordt nergens als uitgangspunt genomen. Niettemin interesseert het ons wel degelijk wat onze technieken precies doen als we ze loslaten op multinormaal verdeelde variabelen, of op steekproeven uit multinormaal verdelingen. En waar mogelijk gebruiken we statistische technieken, voornamelijk als asymptotische perturbatiemethoden. Het is moeilijk om de plaats van deze klapper in de driehoek van figuur 1 aan te geven, omdat er bij ons veel overeenkomsten zijn met tabellaire analyse en diskrete MVA. We zullen wel ergens in de buurt van het zwaartepunt terechtkomen.

1.0 ~~Overzicht~~ Some important predecessors

De in deze klapper besproken technieken vallen in twee groepen uiteen. In de eerste plaats zijn er diverse generalisaties van PCA, en in de tweede plaats overeenkomstige generalisaties van CA. Dit onderscheid komt overeen met het befaamde onderscheid tussen interne en eksterne MVA, of tussen de analyse van samenhang en de analyse van afhankelijkheid. We formaliseren een generalisatie van dit onderscheid. In woorden komt het ongeveer op het volgende neer. MVA technieken als PCA proberen een deelruimte van de ruimte van variabelen te vinden van zo klein mogelijke dimensie waarin alle variabelen passen, MVA technieken als CA proberen een deelruimte te vinden van zo groot mogelijke dimensie die past in alle groepen variabelen. PCA benadert van buiten af en zoekt naar het "kleinste gemene veelvoud", CA benadert van

binnenuit en zoekt naar de "grootste gemene deler". Hoewel onze technieken niet-lineair zijn in de zin dat ze invariante resultaten geven onder niet-lineaire transformaties van de variabelen, gebruiken we toch technieken uit de lineaire analyse en algebra. Dit komt omdat de niet-lineaire transformaties zelf weer een lineaire ruimte definieren, de lineaire transformaties vormen hier een deelruimte van. We werken dus in een grotere ruimte, die echter nog steeds een lineaire ruimte is. Algemene niet-bivariate MVA werkt in een nog grotere ruimte.

Onze aanpak van PCA valt uiteen in twee gedeelten, omdat meer-dimensionale transformaties op twee manieren gedefinieerd kunnen worden. De eerste groep technieken gebruikt meervoudige kwantifikatie, voor iedere dimensie van de oplossing worden nieuwe niet-lineaire transformaties berekend. Het bijbehorende komputerprogramma is HOMALS, het is een vorm van PCA die eerder besproken is door Guttman, Fisher, Burt, Hayashi, en Benzécri. We benaderen het probleem analytisch, uit het oogpunt van het vinden van de optimale transformaties, en geometrisch, als een methode voor het oplossen van meer-dimensionale schaalproblemen. De tweede groep PCA methoden gebruikt enkelvoudige kwantifikatie, waarin de diverse transformaties op een bepaalde manier aan elkaar gerelateerd moeten zijn. Het bijbehorende programma is PRINCALS, het is een vorm van PCA die eerder gepresenteerd is door Kruskal en Shepard, Roskam, en Takane, Young, en De Leeuw. Ook hier onderscheiden we de analytische en de geometrische aanpak. De twee groepen PCA technieken worden aan elkaar gerelateerd, zowel in geometrisch als in algorithmisch opzicht. In dit raamwerk past ook onze generalisatie van CA die we CANALS noemen. Daarnaast is er OVERALS, een kanonische analyse voor meerdere groepen variabelen, waarvan HOMALS, PRINCALS, en CANALS speciale gevallen zijn.

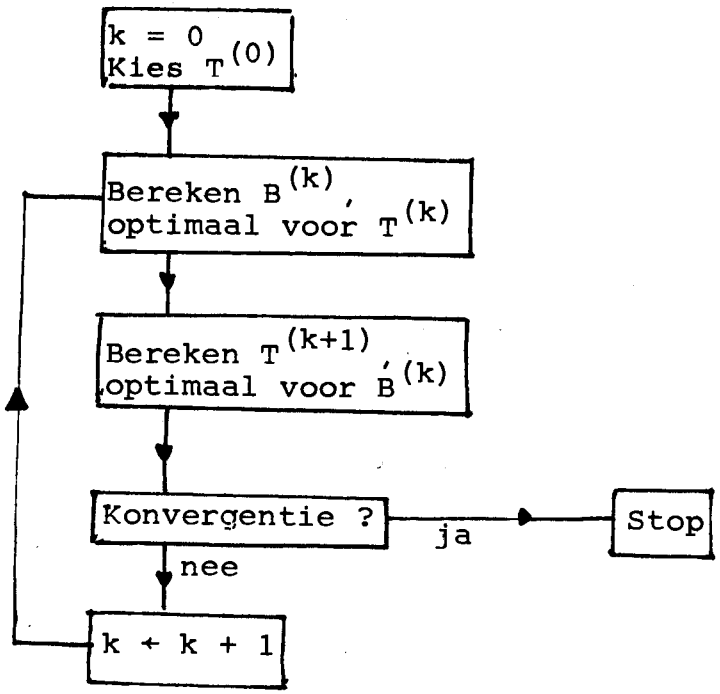
Al deze problemen leiden tot algorithmes die zeer veel op elkaar lijken. Ze worden namelijk aangepakt door een kleinste kwadraten verliesfunktie te definieren, die door een uniform type algorithme geminimaliseerd wordt. We gebruiken het principe van de alternerende kleinste kwadraten, wat gebaseerd is op het onderverdelen van de te berekenen parameters in groepen. We doen dat op zo'n manier dat de optimale waarden voor een groep parameters eenvoudig te berekenen zijn als de waarden van de parameters in de andere groepen bekend zijn. Het algorithme wisselt dan dit soort eenvoudige deel-problemen af, en vervangt de waarden van de parameters steeds door de nieuw berekende optimale waarden uit de deelproblemen. Dit

levert een eenvoudig en redelijk efficient algoritme op, dat snel konvergeert, en dat daarom toegepast kan worden op grote data sets.

Het is eenvoudig in te zien welke soorten deelproblemen door de alternerende kleinste kwadraten methode opgelost moeten worden. We willen in het algemeen twee soorten onbekenden vinden. In de eerste plaats de ruimte die alle variabelen omspant of die alle groepen variabelen gemeenschappelijk hebben. En in de tweede plaats de optimale transformaties van de variabelen, oftewel de optimale schalingen. Voor gegeven schalingen is het vinden van de optimale ruimte een probleem dat ook in klassieke lineaire MVA opgelost wordt. Zowel in PCA als in CA vinden we een basis voor de optimale ruimte door een of ander eigenwaarde-eigenvektor probleem op te lossen. Als omgekeerd de basis tijdelijk gegeven is en we de optimale schalingen moeten vinden, dan betekent dat dat we voor iedere variabele een regressieprobleem op moeten lossen. Dit kan lineaire, monotone, of niet-lineaire regressie zijn, afhankelijk van het schaalnivo van de variabele. We bespreken in deze klapper ook een uitgebreid systeem van de diverse soorten schaalnivo's die in de praktijk het meest voorkomen. Iedere aanname over schaalnivo definieert een klasse van toegestane transformaties, waaruit we onze optimale schalingen moeten kiezen. Niet alle besproken schaalnivo's zijn overigens in onze programmaas geimplementeerd. In figuur 3 zien we de abstrakte struktuur van onze algorithmes. We beginnen met een arbitrair stel transformaties $T^{(0)}$, het algorithme genereert reeksen van transformaties $T^{(0)},T^{(1)},\ldots$ en reeksen van bases $B^{(0)},B^{(1)},\ldots$. De verliesfunktie $\sigma(T,B)$ neemt voortdurend af:
$\sigma(T^0,B^{(0)}) > \sigma(T^{(1)},B^{(0)}) > \sigma(T^{(1)},B^{(1)}) > \sigma(T^{(2)},B^{(1)}) > \ldots$
Omdat $\sigma(T,B) \geq 0$ konvergeert de reeks van verliesfunktiewaarden, we stoppen zodra een iteratie de waarde van de verliesfunktie nauwelijks meer verlaagt. De in deze klapper besproken technieken zijn allemaal variaties op het algorithme in figuur 3.

1.8 Notatie en terminologie

We bespreken eerst de notatie en terminologie in het geval waarin we een eindige populatie of steekproef hebben, terwijl de variabelen de n individuen afbeelden in eindige ruimtes. Aanvankelijk is alles dus eindig. Aan het eind van deze paragraaf bespreken we de meer algemene notatie die het mogelijk maakt oneindige bereiken en oneindige populaties te bestuderen.



figuur 3: Struktuur van onze alternerende kleinste kwadraten algoritmes.

1.8.1 Basisbegrippen, de gegevens

We hebben in MVA dus te maken met een eindige verzameling $\mathbb{N}$ van individuen (ook wel objekten), en met een eindig aantal variabelen. We schrijven $\mathbb{N} = \{\nu_1, \nu_2, \ldots, \nu_n\}$, er zijn dus n individuen, en ze zijn genummerd. Er zijn m variabelen, voor variabelen gebruiken we als regel de indeks j. Een variabele is een funktie η_j, die de verzameling $\mathbb{N}$ afbeeldt in de verzameling $\mathbb{K}_j$. De verzameling $\mathbb{K}_j$ heet het bereik van de variabele η_j, de elementen van $\mathbb{K}_j$ zijn de kategorieen van de variabele. We schrijven $\mathbb{K}_j = \{\kappa_1^j, \kappa_2^j, \ldots, \kappa_{k_j}^j\}$, dus $\mathbb{K}_j$ heeft k_j elementen, en die elementen zijn ook genummerd. Voor kategorieen gebruiken we de indeks r, een typisch element van $\mathbb{K}_j$ is dus κ_r^j.

Het Cartesiaans produkt van de $\mathbb{K}_j$ noemen we het multivariate bereik, de elementen van het multivariate bereik heten profielen. Een profiel is dus een geordend m-voud, met als eerste element een element van $\mathbb{K}_1$, als tweede element een element van $\mathbb{K}_2$, enzovoorts. Omdat de $\mathbb{K}_j$ eindig zijn, is er ook maar een eindig aantal profielen. De data matriks H is een n × m matriks met als elementen $h_{ij} = \eta_j(\nu_i)$, de h_{ij} zijn niet noodzakelijk getallen, we weten alleen dat h_{ij} een element is van $\mathbb{K}_j$. In veel gevallen geldt evenwel dat $\mathbb{K}_j = \{1, 2, \ldots, k_j\}$, de matriks H bevat dan kategorienummers.

Het totale aantal profielen is gelijk aan $\prod k_j$, het produkt van alle k_j. Het is mogelijk dat dit veel kleiner is dan n. In dat geval is de data matriks niet de efficientste manier om de gegevens te koderen. We gebruiken dan liever profiel frekwenties, waarbij de frekwentie van een profiel gedefinieerd wordt als het aantal malen dat een profiel als rij in de data matriks voorkomt. Profiel frekwenties kunnen natuurlijk gelijk aan nul zijn, en dat zal met name het geval zijn wanneer $\prod k_j$ veel groter is dan n. Er zijn twee manieren om profiel frekwenties numeriek weer te geven. We kunnen een matriks konstrueren met $\prod k_j$ rijen en m kolommen waarin alle mogelijke profielen staan, en daar een ekstra kolom met profielfrekwenties aan toevoegen. We noemen een dergelijke matriks de profiel-frekwentie matriks. Deze methode heeft het voordeel dat in de verdere analyse alle rijen korresponderend met profiel frekwentie nul gevoeglijk kunnen worden weggelaten. Dit definieert de gereduceerde profiel-frekwentie matriks, waarvan het aantal rijen gelijk is aan het totaal aantal voorkomende verschillende profielen. Een tweede manier van koderen is wellicht ietwat bekender.

De profielen korresponderen met de cellen van een $k_1 \times k_2 \times \ldots \times k_m$ array, als we de profiel frekwenties in de korresponderende cellen zetten krijgen we een meer-dimensionale kruistabel. De begrippen worden toegelicht in tabel 3a, 3b, 3c, 3e. We hebben daar tien individuen, drie variabelen, iedere variabele heeft drie kategorieen.

1.8.2 Indikator matriksen

We bespreken nu een andere wijze van koderen, die voor ons verdere werk van fundamenteel belang is. De variabele $\eta_j: \mathbb{N} \to \mathbb{K}_j$ definieert een $n \times k_j$ matriks G_j op de volgende manier. G_j is een binaire matriks, al zijn elementen zijn of nul of één, en $g^j_{i\nu} = 1$ als $\eta_j(\nu_i) = \kappa^j_\nu$. In het geval dat $\eta_j(\nu_i) \neq \kappa^j_\nu$ is dus $g^j_{i\nu}$ gelijk aan nul. De matriks G_j wordt de indikator matriks van de variabele η_j genoemd, een element van G_j is gelijk aan één als het individu korresponderend met de rij tuishoort in de kategorie van variabele j korresponderend met de kolom.

In iedere rij van G_j komt altijd precies één element gelijk aan één voor, dus alle rijen van G_j tellen op tot één. We gebruiken u voor de vektor met alle elementen gelijk aan één. Over het algemeen volgt de lengte van u uit de kontekst, we schrijven dus bijvoorbeeld zonder blikken of blozen $G_j u = u$, hoewel de eerste u van lengte k_j is en de tweede u van lengte n is. Er zijn nog een aantal van dit soort gebruiken. De eenheidsmatriks bijvoorbeeld is altijd I, onafhankelijk van zijn orde. De eenheidsvektoren (de kolommen van de eenheidsmatriks) zijn altijd f_i, onafhankelijk van hun lengte. We zorgen er steeds voor dat deze gewoonte niet tot verwarring kan leiden.

De kolomsommen van G_j, dat wil zeggen de elementen van $d_j \triangleq G_j'u$, zijn ook eenvoudig te interpreteren. Element d^j is gelijk aan het aantal individuen ν_i waarvoor $\eta_j(\nu_i)$ gelijk is aan κ^j_ν, dat wil zeggen aan het aantal individuen in kategorie κ^j_ν. Vanzelfprekend geldt $u'd_j = n$. Zo nu en dan is het handig om de d_j neer te schrijven op de diagonaal van een diagonale matriks D_j, van de orde k_j. Dus $D_j u = d_j$, en bovendien $G_j'G_j = D_j$, dus de kolommen van G_j zijn ortogonaal.

Het produkt $C_{j\ell} \triangleq G_j'G_\ell$ heeft ook een eenvoudige betekenis. Het element in rij ν en kolom τ van deze matriks is het aantal individuen ω_i waarvoor zowel $\eta_j(\nu_i) = \kappa^j_\nu$ als $\eta_\ell(\nu_i) = \kappa^\ell_\tau$. En dus is $C_{j\ell}$ de kruistabel van variabelen j en ℓ. Merk op dat $C_{j\ell}u = G_j'G_\ell u = G_j'u = d_j$. Technisch is het handig alle G_j samen te brengen in één enkele matriks G, van afmetingen $n \times (\Sigma k_j)$.

Allemaal achter elkaar dus. Deze G heet de indikator supermatriks. De matrix $C \triangleq G'G$, van afmetingen $(\Sigma k_j) \times (\Sigma k_j)$, bestaat uit m^2 blokken, waarbij blok (j,ℓ) van afmetingen $k_j \times k_\ell$ is, en gelijk is aan $C_{j\ell}$. De diagonale blokken C_{jj} zijn gelijk aan D_j, we verzamelen deze diagonale blokken ook in een diagonale supermatriks D, van dezelfde afmetingen als C. We noemen C de bivariate marginalen, en D de univariate marginalen. De matriksen G, C, en D staan voor ons kleine voorbeeld in tabel 3d, 3f, 3g.

In sommige van onze technieken wordt een matriks G geanalyseerd die niet noodzakelijk uit indikator matriksen is opgebouwd. In dat geval gebruiken we ook de notatie $d = G'u$ voor de kolomtotalen, eventueel ondergebracht in een diagonale matriks D, en we gebruiken een apart symbool $e \triangleq Gu$ voor de rijtotalen, eventueel ondergebracht in een diagonale matriks E (van de orde n). Als regel geldt dus $e = mu$ en $E = mI$, maar op deze regel komen uitzonderingen voor. In dit soort toepassingen is het dikwijls ook niet erg voor de hand liggend om van individuen en variabelen te spreken, we gebruiken dan ook wel de meer neutrale termen rij-objekten en kolom-objekten.

1.8.3 Kwantifikatie, skoring

Het uiteindelijke doel van de meeste in deze klapper besproken technieken is het afbeelden van individuen en/of variabelen en/of kategorieen in een Euclidische ruimte met niet al te veel dimensies. Onze werkwijze kan in het kort als volgt weergegeven worden. We formuleren een verliesfunktie, die een funktie is van zowel de data als de representatie, en die de eigenschap heeft dat hij naar beneden begrensd is. De benedengrens wordt aangenomen als en alleen als de representatie bepaalde eigenschappen heeft die we als een soort van ideaal beschouwen. We spreken dan van een perfekte representatie. Hoe meer onze representatie dit ideaal benadert, hoe lager de waarde van de verliesfunktie zal zijn. We vinden nu de optimale representatie doordat we de verliesfunktie voor gegeven data over representaties minimaliseren. Voorlopig laten we de precieze definitie van de verliesfunktie achterwege, en onthouden we alleen dit algemene idee.

De dimensionaliteit van de ruimte waarin we afbeelden noemen we p, de indeks s wordt gewoonlijk gebruikt voor dimensies. De matriks X, van afmetingen $n \times p$, bevat de scores voor individuen. De matriks Y_j, van afmetingen $k_j \times p$, bevat de kwantifikaties van de kategorieen van variabele j. De matriksen

a	p	u
b	q	v
a	r	v
a	p	u
b	p	v
c	p	v
a	p	u
a	p	v
c	p	v
a	p	v

tabel 3a: data matriks

a	p	u	3
a	p	v	2
a	p	w	0
a	q	u	0
a	q	v	0
a	q	w	0
a	r	u	0
a	r	v	1
a	r	w	0
b	p	u	0
b	p	v	1
b	p	w	0
b	q	u	0
b	q	v	1
b	q	w	0
b	r	u	0
b	r	v	0
b	r	w	0
c	p	u	0
c	p	v	2
c	p	w	0
c	q	u	0
c	q	v	0
c	q	w	0
c	r	u	0
c	r	v	0
c	r	w	0

tabel 3b: profiel-frekwentie matriks

a	p	u	3
a	p	v	2
a	r	v	1
b	p	v	1
b	q	v	1
c	p	v	2

tabel 3c: gereduceerde profiel-frekwentie matriks.

a	b	c	p	q	r	u	v	w
1	0	0	1	0	0	1	0	0
0	1	0	0	1	0	0	1	0
1	0	0	0	0	1	0	1	0
1	0	0	1	0	0	1	0	0
0	1	0	1	0	0	0	1	0
0	0	1	1	0	0	0	1	0
1	0	0	1	0	0	1	0	0
1	0	0	1	0	0	0	1	0
0	0	1	1	0	0	0	1	0
1	0	0	1	0	0	0	1	0

tabel 3d: indikator matriksen.

	p	q	r
a	3	0	0
b	0	0	0
c	0	0	0

u

	p	q	r
a	2	0	1
b	1	1	0
c	2	0	0

v

	p	q	r
a	0	0	0
b	0	0	0
c	0	0	0

w

tabel 3e: meer-dimensionale kruistabel

	a	b	c	p	q	r	u	v	w
a	6	0	0	5	0	1	3	3	0
b	0	2	0	1	1	0	0	2	0
c	0	0	2	2	0	0	0	2	0
p	5	1	2	8	0	0	3	5	0
q	0	1	0	0	1	0	0	1	0
r	2	0	0	0	0	1	0	1	0
u	3	0	0	3	0	0	3	0	0
v	3	2	2	5	1	1	0	7	0
w	0	0	0	0	0	0	0	0	0

tabel 3f: bivariate marginalen (matriks C)

	a	b	c	p	q	r	u	v	w
a	6	0	0	0	0	0	0	0	0
b	0	2	0	0	0	0	0	0	0
c	0	0	2	0	0	0	0	0	0
p	0	0	0	8	0	0	0	0	0
q	0	0	0	0	1	0	0	0	0
r	0	0	0	0	0	1	0	0	0
u	0	0	0	0	0	0	3	0	0
v	0	0	0	0	0	0	0	7	0
w	0	0	0	0	0	0	0	0	0

tabel 3g: univariate marginalen (matriks D)

Y_j worden dikwijls samengebracht in een supermatriks Y, van afmetingen $(\Sigma k_j) \times p$, allemaal onder elkaar dus. Het algemene idee achter onze procedures is dat we voor iedere keuze van X en/of Y de waarde van de verliesfunktie kunnen berekenen, en dat we uiteindelijk X en/of Y willen kiezen op zo'n manier dat het verlies zo klein mogelijk wordt.

We geven nog wat meer terminologie en notatie. Een bepaalde keuze van de kwantifikaties Y_j leidt op een natuurlijke manier ook tot een skoring van de individuen. We definieren $V_j \triangleq G_j Y_j$ als de <u>geinduceerde skores</u> op variabele j, individuen die in dezelfde kategorie van variabele j vallen hebben ook dezelfde geinduceerde skore op die variabele. Alle m variabelen tesamen definieren geinduceerde skores volgens $V_\bullet \triangleq \frac{1}{m} \sum V_j$. We zien hier dat we, zoals gebruikelijk, een stip gebruiken om aan te geven dat er over een indeks gemiddeld is. Op dezelfde manier als <u>direkte kwantifikaties</u> tot <u>geinduceerde skores</u> leiden, kunnen we ook van <u>direkte skores</u> naar <u>geinduceerde kwantifikaties</u> gaan. De geinduceerde kwantifikatie van de kategorieen van variabele j is dan $U_j \triangleq D_j^{-1} G_j' X$. We nemen hier aan dat de inverse van D_j bestaat, dat wil zeggen dat alle kategorieen tenminste eenmaal gebruikt worden. In woorden is de geinduceerde kwantifikatie van een kategorie het gemiddelde van de skores van alle individuen in die kategorie, en is de geinduceerde skore van een individu het gemiddelde van de kwantifikaties van alle kategorieen waarin dat individu zit.

In onze eerdere diskussie van MVA hebben we gezien dat in veel situaties de variabelen onderverdeeld zijn in <u>groepen</u> variabelen. We hebben daar ook wat notatie voor. De onderverdeling in groepen partioneert de indeks set $\mathbb{M}=\{1,2,\ldots,m\}$. We nemen aan dat er M groepen zijn, we gebruiken de indeks K voor groepen, groep K bestaat uit m_K variabelen. We kunnen nu ook de geinduceerde skores van groep K definieren als $V_K = m_K^{-1} \{\sum V_j \mid j \in \mathbb{M}_K\}$, waarbij $\mathbb{M}_K$ de deelverzameling van $\mathbb{M}$ is die korrespondeert met groep K. Het is duidelijk dat $V_\bullet = m^{-1} \sum m_K V_K$.

Een belangrijk onderscheid dat we maken is tussen <u>enkelvoudige</u> en <u>meervoudige</u> meer-dimensionale kwantifikaties. Bij enkelvoudige kwantifikatie eisen we dat de Y_j van de rang één zijn, dat wil zeggen Y_j moet van de vorm zijn $Y_j = z_j a_j'$. Nog anders gezegd: in enkelvoudige kwantifikatie eisen we dat de kolommen van Y_j proportioneel zijn, bij meervoudige kwantifikatie is

er niet een dergelijke eis. Hoewel beide kwantifikaties in het algemeen van dimensie p zijn, is het duidelijk dat enkelvoudige kwantifikatie in zekere zin altijd één-dimensionaal is. Vanzelfsprekend zijn enkelvoudige en meervoudige kwantifikatie precies hetzelfde in het echte één-dimensionale geval (p = 1). We noemen de z_j de <u>enkelvoudige kwantifikaties</u> van variabele j, en we noemen a_j de <u>gewichten</u> voor variabele j. Voor de geinduceerde skores bij enkelvoudige kwantifikatie vinden we $V_j = G_j z_j a_j'$, we definieren daarom $q_j \triangleq G_j z_j$ als de <u>geinduceerde enkelvoudige skores</u> op variabele j. Op de verschillen en overeenkomsten tussen meervoudige en enkelvoudige kwantifikatie gaan we verderop in dit hoofdstuk nog uitvoerig in.

1.8.4 <u>Meer algemene notatie</u>

Stel dat $\mathbb{N}$ nu niet noodzakelijk eindig is, en dat er op $\mathbb{N}$ een waarschijnlijkheidsmaat gedefinieerd is. De variabelen η_j worden daardoor stochastische veranderlijken met waarden in $\mathbb{K}_j$, waarvan we hier aannemen dat het de een of andere deelverzameling van $\mathbb{R}$ is. Voor de funktie $\eta_j(\nu)$ schrijven we nu kortweg $\underline{h}_j$, we nemen dus de konventie van Van Dantzig over en schrijven een streep onder stochastische variabelen (vergelijk Hemelrijk, 1966). We gebruiken bovendien de notatie $E(\underline{x})$ voor de verwachte waarde van $\underline{x}$, $V(\underline{x})$ voor de variantie van $\underline{x}$, en $C(\underline{x},\underline{y})$ voor de kovariantie van $\underline{x}$ en $\underline{y}$. De korrelatie tussen $\underline{x}$ en $\underline{y}$ is tenslotte $R(\underline{x},\underline{y})$.

Het begrip indikator matriks moet ook gegeneraliseerd worden. We gebruikte de indikator matriks tot nu toe als een handige orthogonale basis voor de ruimte van alle funkties of $\underline{h}_j$. Algemener is dus het gebruik van een arbitraire basis $g_r^j(\underline{h}_j)$, r=1,2,... , waarbij k_j, het aantal elementen in de basis mogelijk oneindig is. In deze klapper zijn we altijd alleen geinteresseerd in funkties van $\underline{h}_j$ met eindige variantie (kwadratisch integreerbaar dus). De ruimte van funkties waarin we geintereseerd zijn wordt daardoor een separabele Hilbert ruimte, met een basis die ten hoogste aftelbaar veel elementen bevat. Een speciaal geval is dus wanneer we de ruimte geinduceerd door $\underline{h}_j$ partitioneren in een ten hoogste aftelbaar aantal meetbare verzamelingen, en we definieren de g_r^j als de indikator funkties van die verzamelingen. Voor een bepaalde keuze van de basis definieren we $c_{rt}^{j\ell} = E(g_r^j(\underline{h}_j) g_t^{\ell}(\underline{h}_\ell))$ en $d_{rt}^j = c_{rt}^{jj}$. Als de basis orthogonaal is, zoals bij de indikator funkties, dan geldt $d_{rt}^j = 0$ als $r \neq t$. Als we indikator funkties gebruiken

en de partitionering steeds fijner maken kunnen we gebruik maken van de gebruikelijke grenswaarden argumenten om te laten zien dat we d_r^j kunnen vervangen door $p_j(r)$, de univariate dichtheid van variabele j bij r, en $c_{rt}^{j\ell}$ door $p_{j\ell}(r,t)$, de bivariate dichtheid van variabelen j en ℓ bij r en t. We nemen dan natuurlijk aan dat deze waarschijnlijkheidsdichtheden inderdaad bestaan.

De skores voor individuen X worden vervangen door p stochastische variabelen $\underline{x}_1,\ldots,\underline{x}_p$ op $\mathbb{N}$, we blijven Y_j gebruiken voor de kwantifikaties van de kategorieen, alleen zijn het nu koefficienten voor de vektoren in de gekozen basis, die niet noodzakelijk uit indikator funkties bestaat. De geinduceerde scores V_j worden de stochastische veranderlijken $\underline{v}_s^j = \sum_r y_{rs}^j g_r^j(\underline{h}_j)$, de geinduceerde kwantificaties worden getallen u_{rs}^j die oplossingen zijn van het kleinste kwadraten probleem

$$E(\underline{x}_s - \sum_r u_{rs}^j g_r^j(\underline{h}_j))^2 \quad \min !$$

Hierbij minimaliseren we dus over de u_{rs}^j voor gegeven $\underline{x}_s$. Als de gebruikte basis orthogonaal is dan geldt

$$u_{rs}^j = (d_r^j)^{-1} E(\underline{x}_s g_r^j(\underline{h}_j)).$$

Als de g_r^j indikator funkties zijn kunnen we deze grootheden interpreteren als voorwaardelijke verwachtingen.

1.8.5 Nagekomen notaties

In dit laatste paragraafje nemen we nog wat dingen op die we eigenlijk vergeten zijn. De getransponeerde van een matriks of vektor wordt met een aksent aangeduid. Als A_k een matriks is, dan schrijven we voor de elementen dikwijls a_{ij}^k, de indeks verhuist dus naar boven. Het symbool $\triangleq$ wordt gebruikt voor definities, het symbool { : } voor verzamelingen. Als dit verzamelingen reele getallen zijn dan definieren we ook inf{ : } en min{ : }. Als f(x) een funktie is, dan schrijven we $f(*) = \min\{f(x) : x \in \Omega\}$, en ook $f(*,y) = \min\{f(x,y) : x \in \Omega\}$. We zorgen er daarbij vanzelfsprekend voor dat er geen ambivalentie mogelijk is over de verzamelingen waarover we minimaliseren, of over het al dan niet bestaan van het minimum. Dezelfde soort konventies gelden natuurlijk voor maxima en suprema.

2.0 Inleiding

In dit hoofdstuk bespreken we de analytische theorie van HOMALS, onze eerste vorm van niet-lineaire PCA. De geometrische theorie komt in het volgende hoofdstuk aan de orde. We bespreken eerst HOMALS in één dimensie als een wegings- en transformatieprobleem, en we generaliseren vervolgens naar meerdere dimensies. Speciale aandacht wordt besteed aan binaire gegevens, aan het geval waarin we maar twee variabelen hebben, en aan de resultaten van HOMALS toegepast op de multinormaalverdeling. Het hoofdstuk wordt besloten met een aantal echte voorbeelden, die laten zien in welke situaties we zoal HOMALS toepassen, en wat voor resultaten we daar gewoonlijk uitkrijgen.

2.1 Wegen en transformeren

2.1.1 Iets over wegen

Van het begin af aan zijn mensen erin geintesseerd geweest hoe multivariate gegevens tot univariate gegevens gereduceerd kunnen worden. Ook de kontekst waarin dit nuttig leek was van het begin af aan duidelijk. In 1888 al verscheen er een artikel van Edgeworth over "The Statistics of Examinations", waarin het wegen van de diverse onderdelen aan de orde kwam. In 1913 verscheen Spearman's "Correlations of sums or differences", waar de invloed van het varieren van de gewichten systematisch onderzocht wordt, en waarin onder andere de vergelijkingen van multipele regressie en kanonische analyse afgeleid worden. Rond 1930 kwamen de attitude schalen erbij door het werk van Thurstone en Likert, ook hier worden multivariate gegevens tot univariate gereduceerd. De redenen liggen voor de hand. We hebben schalen nodig omdat we een respektabele wetenschap willen worden, dit speelt met name een rol in de diskussie over Spearman's theorie over algemene intelligentie. Maar bovendien willen we onze onderzoeksobjekten kunnen ordenen, omdat dat voor een groot aantal praktische beslissingen noodzakelijk is (slagen of zakken, behandelen of niet behandelen). Helaas kunnen we niet, zoals in de natuurwetenschappen, het nulpunt gebruiken als norm. Onze norm wordt dus het gemiddelde, en dat verklaart de grote invloed van Quetelet's "homme moyen" in de sociale wetenschappen. Multivariate gegevens kunnen niet direkt geordend worden en leiden niet tot een duidelijke norm, voor zowel de psychometrische theorie van "norm-referenced measurement" als ook voor de hierop gebaseerde praktische beslissingen hebben we één-dimensionale schalen nodig.

Het probleem of we items en tests moeten wegen of gewoon optellen, en het probleem hoe we moeten wegen, bleven dus van groot praktisch

en teoretisch belang. Empirische studies in met name de test theorie wezen in eerste instantie uit dat wegen van items in een test weinig verschil maakte. Deze konklusies werden later wiskundig wat onderbouwd. Een overzicht van de literatuur kan men vinden in het boek van Gulliksen (1950, hoofdstuk 20), maar veel leuker is het overzichtartikel van Burt (1950), en de nauw verwante artikelen Burt (1948, 1951). Als sommige formules in deze artikelen wat gekompliceerd aandoen, dan komt dat waarschijnlijk omdat het formules uit de klassieke griekse psychometrie zijn. We laten hier eerst zin in welke zin de uitspraak van Guilford dat "weighting is not worth the trouble" opgevat moet worden. Bewijzen laten we achterwege.

Stel $\underline{h}_1,\ldots,\underline{h}_m$ zijn stochastische variabelen met $E(\underline{h}_j) = 0$ en $V(\underline{h}_j) = 1$. Definieer $r_{j\ell} \triangleq E(\underline{h}_j\underline{h}_\ell)$, en stel $\underline{u} \triangleq \sum a_j\underline{h}_j$ en $v \triangleq \sum b_j\underline{h}_j$ zijn lineaire kombinaties van de $\underline{h}_j$. Dan geldt $V(\underline{u}) = a'Ra$, $V(\underline{v}) = b'Rb$, en $C(\underline{u},\underline{v}) = a'Rb$. Het is duidelijk dat $R(\underline{u},\underline{v})$ alle waarden tussen -1 en +1 kan aannemen, mar als we eisen dat de gewichten a en b niet-negatief zijn kunnen we al een interessante eerste benedengrens afleiden.

Als $a \geq 0$ en $b \geq 0$ dan $R(\underline{u},\underline{v}) \geq \min_{j=1}^{m} \min_{\ell=1}^{m} r_{j\ell}$.

In de literatuur vindt men ook dikwijls het idee dat kiezen van alle gewichten gelijk aan +1 altijd een behoorlijke keuze is. In de kontekst van multipele korrelatie is dit standpunt recentelijk verdedigd door Wainer (1976, en andere publikaties) onder het hippe motto "Estimating coefficients in linear models: It don't make no nevermind." Het werk van Wainer is terecht bekritiseerd (bv Rozeboom, 1979). In onze kontekst kunnen we echter wel aantonen dat gebruik van $b = b_0$, met alle elementen van b_0 gelijk aan +1, tot een betere benedengrens leidt.

Als $a \geq 0$ en $b = b_0$ dan $R^2(\underline{u},\underline{v}) \geq \min_{j=1}^{m} r_{j.}$, waarbij $r_{j.}$ het gemiddelde van rij j van de matriks $R = \{r_{j\ell}\}$ is.

De konklusie tot zover is dat als de korrelaties tussen de $\underline{h}_j$ hoog zijn, dan is de korrelatie tussen niet-negatieve lineaire kombinaties van de $\underline{h}_j$ nog hoger. De korrelatie tussen de som van de $\underline{h}_j$ en een andere niet-negatieve lineaire kombinatie is nog weer een stuk hoger. Voor items in een test zullen de korrelaties behoorlijk hoog zijn, zo is de test immers gekonstrueerd, en daarom maakt daar wegen inderdaad weinig verschil. Als we een meer algemene batterij variabelen hebben kan wegen nog steeds een groot verschil

maken.

Een tweede argument is gebaseerd op willekeurig gekozen gewichten. We vinden het bij Wilks (1938), bij Burt (1950), en bij Gulliksen (1950). We bespreken hier kort een wat vereenvoudige versie, alweer zonder bewijzen. Stel $\underline{h}_1,\underline{h}_2,\ldots$ is een reeks genormaliseerde stochastische variabelen, en definieer weer $r_{j\ell} \triangleq E(\underline{h}_j\underline{h}_\ell)$. Stel $\underline{a}_1,\underline{a}_2,\ldots$ en $\underline{b}_1,\underline{b}_2,\ldots$ zijn twee andere reeksen stochastische variabelen, onafhankelijk van elkaar, onafhankelijk van de $\underline{h}_j$. Veronderstel dat $E(\underline{a}_j) = E(\underline{b}_j) = \mu \neq 0$ voor alle j, en $V(\underline{a}_j) = V(\underline{b}_j) = \sigma^2$ voor alle j. Definieer $\kappa \triangleq \sigma/\mu$, de variatiekoefficient van de $\underline{a}_j$ en de $\underline{b}_j$. Definieer bovendien

$$\underline{u}_m \triangleq \sum_{j=1}^{m} \underline{a}_j\underline{h}_j,$$

$$\underline{v}_m \triangleq \sum_{j=1}^{m} \underline{b}_j\underline{h}_j,$$

en

$$r_m \triangleq m^{-2} \sum_{j=1}^{m}\sum_{\ell=1}^{m} r_{j\ell}.$$

We veronderstellen tenslotte dat r_m, de gemiddelde korrelatie van de eerste m variabelen, konvergeert naar een getal r als $m \to \infty$. Uit de aannamen tot nu toe volgt dat $R(\underline{u}_m,\underline{v}_m)$ in waarschijnlijkheid naar één konvergeert. De volgende formule geeft wat preciesere informatie.

$$E(R(\underline{u}_m,\underline{v}_m)) = 1 - \frac{\kappa^2}{mr} + O(m^{-2}).$$

We zien dus dat de korrelatie dichter bij één ligt naarmate de variatie van de gewichten klein is, naarmate de gemiddelde interkorrelatie hoog is, en naarmate er meer variabelen zijn. Een heel nette en eenvoudig te interpreteren formule dus. Maar zoals alle asymptotische argumenten is dit wat onbevredigend omdat het niet precies vertelt wat er bij een eindig en niet al te groot aantal variabelen gebeurt, en zoals de meeste probabilistische argumenten is het ook wat onbevredigend omdat het op een geidealiseerde en onrealistische situatie berust. In de werkelijkheid kiezen we onze gewichten niet willekeurig, maar volgens het een of andere optimaliteitskriterium. Het lijkt daarom zinvol om in diverse in de praktijk voorkomende situaties na te gaan in hoeverre wegen zinvol is.

Voor binaire tests zijn er probabilistische modellen zoals dat van Birnbaum en Rasch die het mogelijk maken optimale gewichten

af te leiden uit theoretische overwegingen. Hetzelfde geldt wanneer we aannemen dat de variabelen voldoen aan het één-faktor model van Spearman of aan het schaalmodel van Guttman. In deze klapper volgen we ook hier de omgekeerde weg. We ontwikkelen wegingstechnieken die geen bepaald model veronderstellen, en we gaan na hoe die technieken werken bij bepaalde modellen. Bovendien bespreken we wegingstechnieken die gebruik maken van niet-lineaire wegingen. Voor dit soort technieken gaan de hier besproken argumenten tegen wegen helemaal niet meer op.

2.1.2 Homogeniteit zonder wegen

Stel $\underline{h}_1,\ldots,\underline{h}_m$ zijn gestandariseerde stochastische variabelen. We zijn geinteresseerd in de vraag in hoeverre we deze variabelen kunnen vervangen door één enkele variabele, en in hoeverre dat mogelijk is zonder al te veel verlies van informatie. Deze vraag kwam al aan de orde in het klassieke artikel uit 1888 waarin Galton de korrelatiekoefficient introduceerde. Aan het eind van dat artikel lezen we: "Neither is it necessary to give examples of a method by which the degree may be measured, in which the variables in a series each member of which is the summed effect of n variables, may be modified by their partial co-relation. After transmuting the separate measures as above, and them summing them, we should find the probable error of any one of them to be $\sqrt{n}$ if the variables were perfectly independent, and n if they were rigidly and perfectly co-related." (1888, pag 144-145). Als we dit citaat wat vrij interpreteren (met Burt), dan stelt Galton hier als maat voor de homogeniteit van een aantal variabelen de gemiddelde korrelatiekoefficient voor. We formaliseren dit argument hier.

Stel $\underline{z}$ is een kandidaat variabele die we willen gebruiken om alle $\underline{h}_j$ te vervangen. We kunnen vaststellen hoe suksesvol onze poging is door de verliesfunktie

$$\sigma(\underline{z}) \triangleq \frac{1}{m} \sum_{j=1}^{m} V(\underline{z} - \underline{h}_j)$$

te evalueren. Het is duidelijk dat $\sigma(\underline{z}) \geq 0$, en dat $\sigma(\underline{z}) = 0$ als en alleen als $\underline{z} = \underline{h}_j$ voor alle j, dat wil zeggen als en alleen als alle $\underline{h}_j$ dezelfde stochastische variabele zijn (behalve mogelijker wijs op een verzameling met waarschijnlijkheid nul). We laten nu zien hoe we $\underline{z}$ het beste kunnen kiezen. Definieer ook

$$\sigma(*) \triangleq \min \{\sigma(\underline{z}) \mid \underline{z}\}.$$

Het minimum wordt aangenomen voor $\underline{z}$ gelijk aan $\underline{h}_.$ en het minimum is gelijk aan

$$\sigma(*) = 1 - V(\underline{h}_.) = 1 - r_{..}.$$

We vinden dus de door Galton voorgestelde maat terug.
Nog iets over de notatie. We gebruiken hier de algemene notatie, die ervan uitgaat dat $\underline{h}_j$ en $\underline{z}$ stochastische variabelen zijn. In het speciale geval waarin we een eindige populatie hebben met n elementen en een diskrete waarschijnlijkheidsverdeling op die populatie geldt dat $\underline{h}_j$ en $\underline{z}$ opgevat kunnen worden als vektoren van de lengte n. We noemen ze dan h_j en z, zonder streep, omdat we moeten onderscheiden tussen de variabele en de waarden die de variabele aanneemt. Als we de waarschijnlijkheidsverdeling in een vektor p zetten dan geldt dus $E(\underline{z}) = p'z$, en als we de elementen van p in een diagonale matriks P onderbrengen dan geldt $E(\underline{z}^2) = z'Pz$. Als alle elementen van p gelijk zijn aan 1/n, dan $E(\underline{z}) = z_.$, en als $E(\underline{z}) = 0$ dan $V(\underline{z}) = \frac{1}{n} SSQ(z)$. We zien dus dat we in deze speciale gevallen de formules steeds kunnen herinterpreteren door de variabelen te vervangen door de waarden die ze aannemen, en door V(.) te vervangen door SSQ(.). In het geval van een willekeurige steekproef geldt dat $\underline{p}$ en $\underline{P}$ zelf stochastische veranderlijken zijn, en hetzelfde geldt daardoor voor $E(\underline{z}) = \underline{p}'z$ en dergelijke.

2.1.3 Lineair wegen

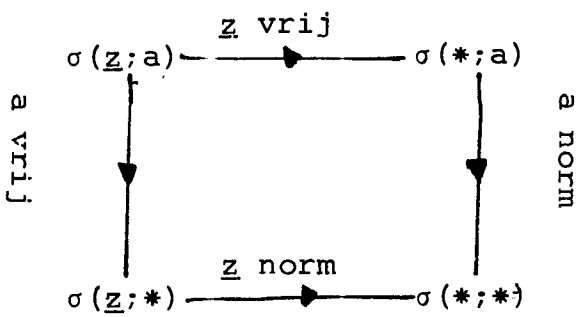
Stel nu dat we de $\underline{h}_j$ lineair mogen transformeren in een poging een hogere homogeniteit te krijgen. Lineair transformeren verandert natuurlijk de korrelaties niet, en dus hebben we niets aan de maat $r_{..}$ uit de vorige paragraaf. We moeten de analyse op zo'n manier generaliseren dat lineaire transformaties wel degelijk verschil maken. Om te beginnen nemen we nog steeds aan dat $E(\underline{h}_j) = 0$, maar niet meer dat $V(\underline{h}_j) = 1$. We definieren ook de matriks C met elementen $c_{j\ell} = C(\underline{h}_j, \underline{h}_\ell)$ en de diagonale matriks D met diagonaal elementen $d_{jj} = V(\underline{h}_j)$. Tenslotte $R \triangleq D^{-\frac{1}{2}}CD^{-\frac{1}{2}}$, zodat $r_{j\ell} = R(\underline{h}_j, \underline{h}_\ell)$.

Voor een gegeven keuze van de kandidaatschaal $\underline{z}$, met $E(\underline{z}) = 0$, en voor gegeven gewichten a_j is het verlies gelijk aan

$$\sigma(\underline{z};a) \triangleq \frac{1}{m} \sum_{j=1}^{m} V(\underline{z} - a_j\underline{h}_j).$$

We willen nu weten hoe klein we het verlies kunnen maken door een slimme keuze van $\underline{z}$ en a. Omdat $\underline{z} = 0$ en $a = 0$ onvermijdelijk leidt tot $\sigma(\underline{z};a) = 0$ heeft het geen zin om $\sigma(\underline{z};a)$ over $\underline{z}$ en a te minimaliseren. Altans niet over <u>alle</u> $\underline{z}$ en <u>alle</u> a. We moeten $\underline{z}$ en/of a <u>normaliseren</u>.

De keuze van de normalisatie definieert twee verschillende minimalisatieproblemen. In het eerste probleem is $\underline{z}$ vrij, en eisen we dat a genormaliseerd is, in het tweede probleem is a vrij en eisen we dat $\underline{z}$ genormaliseerd is. Het mooie is nu dat als we de normalisatiekondities op een slimme manier kiezen beide problemen tot dezelfde oplossing leiden. We hebben dit weergegeven in figuur 4, waarin we de konventie gebruiken dat minimalisatie van een funktie over een variabele weergegeven wordt met het vervangen van die variabele door een ster. Het minimum dat we uiteindelijk zoeken is dus $\sigma(*;*)$, en er zijn twee wegen met ieder twee stappen van $\sigma(\underline{z};a)$ naar $\sigma(*;*)$. We gebruiken deze twee wegen omdat ze in direkt verband staan met het alternerende kleinste kwadraten principe dat we voor onze berekeningen gebruiken, we kunnen de twee wegen zien als een soort van konseptuele variant van alterende kleinste kwadraten. Bovendien liggen er langs de wegen allerlei interpretaties van onze procedures.



Figuur 2.1: Wegen der minimalisatie.

De eerste weg begint met de definitie

$$\sigma(*;a) = \min \{\sigma(\underline{z};a) : \underline{z}\}.$$

In figuur 4 betekent dat dat we van links boven naar rechts boven stappen. Het minimum $\sigma(*;a)$ wordt aangenomen voor $\underline{z} = \underline{y}_.$, waarbij $\underline{y}_j = a_j \underline{h}_j$. En het minimum is gelijk aan

$$\sigma(*;a) = \frac{1}{m} \sum_{j=1}^{m} V(\underline{y}_j) - V(\underline{y}_.)$$

We geven een interpretatie van $\sigma(*;a)$ in variantie-analytische termen. In de enkelvoudige klassifikatie gebruiken we partioneringen van de variantie in een komponent <u>tussen</u> klassifikatienivo's en een komponent <u>binnen</u> klassifikatienivo's. Op dezelfde manier definieren we hier variantie tussen variabelen en variantie binnen variabelen. Daartoe schrijven we

$$\underline{y}_j = \underline{y}_\cdot + (\underline{y}_j - \underline{y}_\cdot).$$

Als we beide kanten kwadrateren en verwachte waarden nemen, dan vinden we

$$V(\underline{y}_j) = V(\underline{y}_\cdot) + V(\underline{y}_j - \underline{y}_\cdot) + 2C(\underline{y}_\cdot, \underline{y}_j - \underline{y}_\cdot).$$

Als we dit tenslotte over j optellen dan valt de kovariantieterm weg, en vinden we uiteindelijk

$$\sum_{j=1}^{m} V(\underline{y}_j) = mV(\underline{y}_\cdot) + \sum_{j=1}^{m} V(\underline{y}_j - \underline{y}_\cdot).$$

Deze laatste vergelijking kunnen we lezen als Totaal = Tussen + Binnen, en we kunnen hem herschrijven als

$$T(\underline{y}_1,\ldots,\underline{y}_m) = B(\underline{y}_1,\ldots,\underline{y}_m) + W(\underline{y}_1,\ldots,\underline{y}_m).$$

Om verwarring te voorkomen: B is de afkorting van Between, dus van Tussen, en W is de afkorting van Within, dus van Binnen. Als we nu $\sigma(*;a)$ in deze termen herschrijven, dan vinden we

$$\sigma(*;a) = \frac{1}{m} T(\underline{y}_1,\ldots,\underline{y}_m)\{1 - \frac{B(\underline{y}_1,\ldots,\underline{y}_m)}{T(\underline{y}_1,\ldots,\underline{y}_m)}\}.$$

Natuurlijk is $\sigma(*;a) = \frac{1}{m} W(\underline{y}_1,\ldots,\underline{y}_m)$ een eenvoudiger formule, maar we hebben de bovenstaande gekozen om vergelijking met de vorige paragraaf eenvoudiger te maken. Uit deze formule blijkt dat een natuurlijke normalisatie konditie is $T(\underline{y}_1,\ldots,\underline{y}_m) = m$, en dat we de ratio van variantie tussen en totale variantie van de getransformeerde variabelen maksimaliseren. Als we a_j gelijk aan $V^{-\frac{1}{2}}(\underline{h}_j)$ kiezen dan geldt inderdaad dat $T(\underline{y}_1,\ldots,\underline{y}_m) = m$, en bovendien geldt $\sigma(*;a) = 1 - r_{\cdot\cdot}$, een oude bekende.

Voor we verder gaan op de eerste weg maken we eerst even de eerste stap op de tweede weg. Deze gaat in figuur 4 van links boven naar links beneden. We definieren

$$\sigma(\underline{z};*) \triangleq \min_a\{\sigma(\underline{z};a) : a\}.$$

Het minimum wordt aangenomen voor $a_j = C(\underline{z},\underline{h}_j)/V(\underline{h}_j)$, de lineaire regressiekoefficient van $\underline{z}$ op $\underline{h}_j$. Het minimum is gelijk aan

$$\sigma(\underline{z};*) = V(\underline{z})\{1 - \frac{1}{m}\sum_{j=1}^{m} R^2(\underline{z},\underline{h}_j)\}.$$

Een voor de hand liggende normalizatie is dus $V(\underline{z}) = 1$, we zien dat we $\underline{z}$ zo kiezen dat de gemiddelde gekwadrateerde korrelatie van $\underline{z}$ met de $\underline{h}_j$ zo groot mogelijk is.

Voor de tweede stap op de eerste weg (van rechts boven naar rechts beneden) hebben we eerst wat matriks notatie nodig. We vinden
$T(\underline{y}_1,\ldots,\underline{y}_m) = a'Da$,
$B(\underline{y}_1,\ldots,\underline{y}_m) = \frac{1}{m}\, a'Ca$.
Definieer
$\sigma(*;*) \triangleq \min\{\sigma(*;a) : a'Da = m\}$.
Dan

$$\sigma(*;*) = 1 - \max\{\frac{a'Ca}{m\ a'Da} : a\}$$

Hieruit volgt (zie de matriks algebra appendiks) dat
$\sigma(*;*) = 1 - \lambda_+$,
waarbij λ_+ de grootste eigenwaarde van $\frac{1}{m}\, R$ is. Als t de bijbehorende genormaliseerde eigenvektor is, dus $Rt = m\lambda_+ t$ en $t't = 1$, dan vinden we als optimale a voor het eerste probleem $a = m^{\frac{1}{2}}D^{-\frac{1}{2}}t$. De bijbehorende optimale $\underline{z}$ voor het eerste probleem is

$$\underline{z} = \frac{1}{m}\sum_{j=1}^{m} a_j\underline{h}_j = m^{-\frac{1}{2}}\sum_{j=1}^{m} t_j\underline{h}_j^o,$$

waarbij $\underline{h}_j^o = \underline{h}_j/V^{\frac{1}{2}}(\underline{h}_j)$. Hieruit volgt dat voor de optimale $\underline{z}$ uit het eerste probleem geldt $V(\underline{z}) = \lambda_+$ en $R(\underline{z},\underline{h}_j) = m^{\frac{1}{2}}\lambda_+^{\frac{1}{2}}t_j$, zodat tevens de gemiddelde gekwadrateerde korrelatie tussen $\underline{z}$ en de $\underline{h}_j$ gelijk is aan λ_+. De eerste weg is nu klaar. We hebben gezien dat de interpretatie is dat we tussen/totaal maksimaliseren, of, anders gezegd, dat we $V(\sum a_j h_j)$ zo groot mogelijk maken op voorwaarde $\sum V(a_j h_j) = m$. Een derde manier om deze interpretatie te formuleren is dat we de som van de kovarianties $\sum\sum C(\underline{y}_j,\underline{y}_\ell)$ maksimaliseren, terwijl we de som van de varianties $\sum V(\underline{y}_j)$ konstant houden.

De tweede stap op de tweede weg (van links onder naar rechts onder) gebruikt de definitie
$\sigma(*;*) \triangleq \min\{\sigma(\underline{z};*) : V(\underline{z}) = 1\}$.
Uit de resultaten in de eerste stap volgt

$$\sigma(*;*) = 1 - \max\{\frac{1}{m}\sum_{j=1}^{m} R^2(\underline{z},\underline{h}_j) : \underline{z}\}.$$

Om de maksimalisatie uit te voeren schrijven we $\underline{z}$ als $\underline{z} = \sum b_j\underline{h}_j + \underline{u}$, waarbij $C(\underline{u},\underline{h}_j) = 0$ voor alle j. Stel $V(\underline{u}) = \varepsilon^2$. Dan

$$\sigma(*;*) = 1 - \max\{\frac{b'CD^{-1}Cb}{m(b'Cb + \varepsilon^2)} : b,\ \varepsilon^2\} = 1 - \lambda_+.$$

Het maksimum wordt bereikt voor $\varepsilon^2 = 0$ en $b = m^{-\frac{1}{2}}\lambda_+^{-\frac{1}{2}}D^{-\frac{1}{2}}t$. Dit geeft

voor het tweede probleem een optimale $\underline{z}$ gelijk aan $\underline{z} = m^{-\frac{1}{2}}\lambda_+^{-\frac{1}{2}}\Sigma t_j \underline{h}_j^o$, met een bijbehorende optimale a gelijk aan $a = m^{\frac{1}{2}}\lambda_+^{\frac{1}{2}}D^{-\frac{1}{2}}t$. Dus $a'Da = m\lambda_+$, en de gemiddelde kwadratische korrelatie is λ_+. We zien dus dat beide wegen tot een zelfde $\sigma(*;*)$ leiden, en dat de bijbehorende optimale a en $\underline{z}$ slechts een eenvoudige schaalfaktor verschillen. Een alternatieve mogelijkheid is dat we zowel eisen dat $V(\underline{z}) = 1$ als dat $a'Da = m$. Er zijn dan weer twee wegen, die leiden naar $\sigma(*;*) = 2(1 - \lambda_+^{\frac{1}{2}})$, maar de oplossingen zijn dezelfde als die we eerder gevonden hebben. We gaan niet verder op de details in, het is allemaal al vervelend genoeg. Wat we in deze paragraaf op een onortodokse manier hebben laten zien, is dat het optimaal is om te wegen met de eerste principale komponent van de korrelatiematriks. We hebben dit gedaan voor de populatiekorrelatiematriks, maar we hadden het ook voor de steekproef of voor niet-stochastische vektoren van lengte n kunnen doen.

2.1.4 Terug naar de geschiedenis

Het argument dat wegen niet noodzakelijk is, betekent dat we veronderstellen dat λ_+ ongeveer gelijk is aan $r_{..}$. Een noodzakelijke voorwaarde voor $\lambda_+ = r_{..}$ is dat de rijgemiddelden $r_{j.}$ allemaal gelijk zijn. Als alle korrelaties positief zijn is deze voorwaarde ook voldoende voor $\lambda_+ = r_{..}$. In het geval dat alle korrelaties positief zijn kunnen we uit de Perron-Frobenius theorie afleiden dat $\min r_{j.} \leq \lambda_+ \leq \max r_{j.}$. Bij test bestaande uit items die op dit soort kriteria geselekteerd zijn is het dus best mogelijk dat "weighting is not worth the trouble", in andere situaties zullen de $r_{j.}$ soms flink verschillen. En zelfs voor het wegen van items denkt men tegenwoordig iets genuanceerder. Na een uitvoerige analyse van het wegingsdilemma zegt Rozeboom: "To put it bluntly, second digit precision in item weighting is generally a waste of effort." (1979, pag 296).

Het is duidelijk dat onze techniek lineair is in de zin van het eerste hoofdstuk. Als we de $\underline{h}_j$ lineair transformeren, en dan op nieuw starten, dan vinden we dezelfde korrelatiematriks. We vinden daarom dezelfde t en dezelfde $\underline{z}$, alleen a is een beetje anders omdat a berekend wordt door t aan te passen met de variantie van de variabelen. Een andere manier om dit te beschrijven is dat onze techniek schaal-onafhankelijk is, we hebben geen nulpunt en geen eenheid nodig om de wegingstechniek toe te kunnen passen, of we nu Celsius of Fahrenheit gebruiken maakt niets uit.

Iets over de geschiedenis van deze lineaire wegingstechniek. Zoals Burt opmerkt wordt de techniek het eerst genoemd in MacDonell (1901), en wel in de volgende bewoordingen. "Prof. Pearson has pointed out to me that the ideal index characters would be given if we calculated the seven directions of uncorrelated variables, that is, the principal axes of the correlation 'ellipsoid'" (1901, pag

209). Op dezelfde pagina vinden we: "I propose to return in a later paper to this calculation." Maar dat latere paper is er naar wij weten nooit gekomen. Als uitvinder van de techniek, en in feite van de meer-dimensionale versie ervan, kunnen we Pearson dus beschouwen. Hij wordt inderdaad vaak in dit verband genoemd, maar dan refereert men gewoonlijk aan Pearson (1901). Daarin wordt echter een 'duale' benadering gebruikt, dat wil zeggen Pearson kiest de a_j op zo'n manier dat $V(\sum a_j \underline{h}_j)$ zo klein mogelijk is. Hij beschouwt dit als een alternatieve manier om multipele regressie te doen, in situaties waarin zowel de onafhankelijke als de afhankelijke variabelen meetfouten hebben, of in gevallen waarin het onderscheid tussen onafhankelijke en afhankelijke variabelen niet zo duidelijk ligt. Immers ook bij multipele regressie minimaliseren we $V(\sum a_j \underline{h}_j)$, alleen gebruiken we daar als normalisatiekonditie (wanneer we $\underline{h}_1$ uit de rest willen voorspellen) dat $a_1 = 1$. We krijgen daardoor andere waarden voor a_j wanneer we bijvoorbeeld $\underline{h}_2$ willen voorspellen uit de rest. Bij Pearson's techniek normaliseren we met $\sum V(a_j \underline{h}_j) = 1$, en vinden we dezelfde oplossing voor al die verschillende multipele regressie problemen. Het is overigens nuttig om te onthouden dat multipele regressie te maken heeft met de kleinste eigenwaarde van de korrelatiematriks, en dat onze lineaire wegingstechnieken te maken hebben met de grootste eigenwaarde.

De wegingstechniek wordt ook wel toegeschreven aan Hotelling (1933), maar dat berust op een misverstand. Pearson was veel eerder. In het artikel van Hotelling wordt PCA bovendien geintroduceerd als een vorm van FA, het eksplisiete doel is om de verliesfunktie $\sum_{j=1}^{m} V(\underline{h}_j - \sum_{s=1}^{p} a_{js} \underline{z}_s)$ te minimaliseren. Hotelling laat zien dat de principale komponenten de oplossing van dit benaderingsprobleem geven. En vervolgens zegt hij: "An easily verified property of the method is that the first of our principal components has a greater mean square correlation with the tests than does any other variable; ..." (1933, pag 422). Dit is echter meer een soort terloopse opmerking. Eksplisiet gebruik van principale komponenten voor lineair wegen en voor het konstrueren van één-dimensionale schalen vinden we bij Horst (1936), Edgerton en Kolbe (1936), en bij Wilks (1938). Zij hanteren allemaal het kriterium $V(\sum a_j \underline{h}_j)$, Wilks bespreekt nog andere mogelijkheden maar vindt uiteindelijk zijn eerste methode de meest aanbevelenswaardige. En dat is gewoon de eerste principale komponent van de korrelatiematriks.

2.1.5 Niet-lineair wegen

Het is niet erg ingewikkeld om in te zien hoe onze verliesfunktie gegeneraliseerd wordt voor het geval we niet-lineaire wegingen toestaan. Definieer

$$\sigma(\underline{z};\Phi) = \frac{1}{m} \sum_{j=1}^{m} V(\underline{z} - \phi_j(\underline{h}_j)),$$

hierbij nemen we in eerste instantie niets aan over ϕ_j, behalve dat $V(\phi_j(\underline{h}_j)) < \infty$. Ons probleem is weer hoe we $\underline{z}$ en Φ moeten kiezen om $\sigma(\underline{z};\Phi)$ zo klein mogelijk te maken.

Evenals in de vorige paragraaf bewandelen we bij het berekenen van het minimum verlies twee verschillende wegen, die op hetzelfde punt uitkomen. En we doen dat weer omdat die twee wegen direkt in verband staan met het alternerende kleinste kwadratenprincipe dat we voor onze berekeningen gebruiken, en omdat er langs de wegen allerlei interpretaties liggen. In de eerste plaats lossen we het probleem op om $\sigma(\underline{z};\Phi)$ te minimaliseren over alle $\underline{z}$, en over de Φ die voldoen aan $\sum V(\phi_j(\underline{h}_j)) = m$. De eerste stap in de oplossing van het eerste probleem is berekenen van

$$\sigma(*;\Phi) = \min \{\sigma(\underline{z};\Phi) : \underline{z}\}.$$

Als $\underline{y}_j = \phi_j(\underline{h}_j)$ dan wordt het minimum aangenomen voor $\underline{z} = \underline{y}_.$, zoals gewoonlijk. Met deze aangepaste definitie van $\underline{y}_j$ kunnen we alle formules en interpretaties uit de vorige paragraaf gewoon overnemen, voor zover ze betrekking hebben op $\sigma(*;a)$. Met name is het zo dat we nu onze procedure kunnen interpreteren als het vinden van niet-lineaire transformaties $\underline{y}_j = \phi_j(\underline{h}_j)$ op zo'n manier dat de som van de kovarianties van de getransformeerde variabelen maksimaal is, op voorwaarde dat de som van de varianties gelijk is aan m. Een eenvoudige truuk kan gebruikt worden om een alternatieve interpretatie te bewijzen. We kunnen $\underline{y}_j = \phi_j(\underline{h}_j)$ schrijven als $\underline{y}_j = a_j\psi_j(\underline{h}_j)$, en doordat we de a_j ingevoerd hebben kunnen we nu zonder verlies van algemeenheid eisen dat $V(\psi_j(\underline{h}_j)) = 1$. Minimaliseren van $\sigma(*;\Phi)$ over alle Φ zodat $\sum V(\phi_j(\underline{h}_j)) = m$ kan nu ook geformuleerd worden als minimaliseren van $\sigma(*;a;\Psi)$ over alle a zodanig dat $a'a = m$ en over alle Ψ zodanig dat $V(\psi_j(\underline{h}_j)) = 1$. Als we nu definieren

$$\sigma(*;*;\Psi) = \min \{\sigma(*;a;\Psi) \mid a'a = m\},$$

dan weten we uit de vorige paragraaf dat $\sigma(*;*;\Psi) = 1 - \lambda_+$, waarbij λ_+ de grootste eigenwaarde is van $\frac{1}{m}R(\Psi)$, waarbij $R(\Psi)$ de korrelatiematriks van de getransformeerde variabelen is. De alternatieve interpretatie is dus dat we onze variabelen zo transformeren dat

de grootste eigenwaarde van de korrelatiematriks van de getransformeerde variabelen zo groot mogelijk is.

Ondertussen hebben we nog niet aangetoond hoe we $\sigma(*;\Phi)$ moeten minimaliseren over Φ, of hoe we $\sigma(*;*;\Psi)$ moeten minimaliseren over Ψ. Over het algemeen is dat ook niet zo eenvoudig. Voor we dit probleem aanpakken bekijken we eerst even het begin van de tweede weg. Bij dit tweede probleem minimaliseren we $\sigma(\underline{z};\Phi)$ over $\underline{z}$ met $V(\underline{z}) = 1$ en over alle mogelijke Φ. Als eerste stap in de richting van de oplossing definieren we, zoals gewoonlijk,

$$\sigma(\underline{z};*) \triangleq \min \{\sigma(\underline{z};\Phi) : \Phi\}.$$

Nu geldt (bv Whittle, 1970, pag 81; Neveu, 1964, pag 117) dat $V(\underline{z} - \phi_j(\underline{h}_j))$ geminimaliseerd wordt over ϕ_j door $\phi_j(\underline{h}_j)$ te kiezen als $E(\underline{z}|\underline{h}_j)$, de voorwaardelijke verwachting van $\underline{z}$ gegeven $\underline{h}_j$. Het minimum is gelijk aan

$$V(\underline{z} - E(\underline{z}|\underline{h}_j)) = 1 - V(\underline{z}|\underline{h}_j),$$

waarbij $V(\underline{z}|\underline{h}_j) \triangleq V(E(\underline{z}|\underline{h}_j))$ de voorwaardelijke variantie van $\underline{z}$ gegeven $\underline{h}_j$ is. Dus

$$\sigma(\underline{z};*) = 1 - \frac{1}{m} \sum_{j=1}^{m} V(\underline{z}|\underline{h}_j),$$

en deze formule levert een keurige nieuwe interpretatie op. We zoeken een stochastische variabele $\underline{z}$ zodanig dat de gemiddelde korrelatie ratio van $\underline{z}$ met de $\underline{h}_j$ zo groot mogelijk is. Voor mensen die het vergeten zijn: de korrelatie ratio van $\underline{x}$ op $\underline{y}$ is de voorwaardelijke variantie van $\underline{x}$ gegeven $\underline{y}$, gedeeld door de variantie van $\underline{x}$. De korrelatie ratio ligt tussen nul en één, hij is gelijk aan nul als $\underline{x}$ en $\underline{y}$ onafhankelijk zijn, en hij is gelijk aan één als $\underline{x}$ een funktie is van $\underline{y}$. Als de regressie lineair is, dan is de korrelatie ratio gelijk aan het kwadraat van de korrelatiekoefficient. In het algemeen is het alweer niet eenvoudig om $\sigma(\underline{z};*)$ te minimaliseren over $\underline{z}$, we gaan dus even terug naar de eerste weg.

De tweede stap op de eerste weg die we moeten uitvoeren is het minimaliseren van $\sigma(*;\Phi)$ over die Φ waarvoor $\sum V(\phi_j(\underline{h}_j)) = m$. De eenvoudigste manier om dit probleem aan te pakken is gebruik van een komplete ortonormale basis voor ieder van de ruimtes

$$\mathcal{L}_j \triangleq \{\phi_j \mid E(\phi_j(\underline{h}_j)) = 0 \ \& \ V(\phi_j(\underline{h}_j)) < \infty\}.$$

Stel g_{js} is zo'n basis, s=1,2,..., en voor alle s n t geldt

dus $C(g_{js}(\underline{h}_j), g_{jt}(\underline{h}_j)) = \delta^{st}$. Iedere ϕ_j in $\mathcal{L}_j$ kunnen we schrijven in de vorm

$$\phi_j(\underline{h}_j) = \sum_{s=1}^{\infty} a_{js} g_{js}(\underline{h}_j).$$

Uit deze representatie volgt dat

$$C(\phi_j(\underline{h}_j), \phi_\ell(\underline{h}_\ell)) = \sum_{s=1}^{\infty} \sum_{t=1}^{\infty} a_{js} a_{\ell t} C(g_{js}(\underline{h}_j), g_{\ell t}(\underline{h}_\ell)).$$

We kunnen de a_{js} voor gegeven s verzamelen in een m-vektor a_s, en voor gegeven s en t kunnen we de kovarianties tussen de funkties in de basis verzamelen in een m × m matriks C_{st}. We vinden dan voor de som van alle kovarianties tussen de getransformeerde variabelen de ekspressie

$$B(\underline{y}_1, \ldots, \underline{y}_m) = \frac{1}{m} \sum_{s=1}^{\infty} \sum_{t=1}^{\infty} a_s' C_{st} a_t,$$

terwijl de som van de varianties gelijk is aan

$$T(\underline{y}_1, \ldots, \underline{y}_m) = \sum_{s=1}^{\infty} a_s' a_s.$$

De stationaire vergelijkingen zijn dus $\sum_{t=1}^{\infty} C_{st} a_t = m\lambda a_t$, en dit is een eigenwaarde probleem.

Aan het eind van de eerste weg vinden we dus $\sigma(*;*) = 1 - \lambda_+$, zoals eerder, alleen nu is λ_+ de grootste eigenwaarde van een veel grotere korrelatiematriks. Op dezelfde manier als in de vorige paragraaf, met alleen wat gekompliceerdere notatie, kunnen we aantonen dat de tweede weg naar hetzelfde eigenwaardenprobleem leidt, zodat de situatie volkomen analoog is aan de situatie uit de vorige paragraaf.

In de praktijk is het veelal niet goed mogelijk om eigenwaarde problemen met oneindig grote matriksen op te lossen. We gebruiken dan een orthonormale basis voor een eindig-dimensionale deelruimte van de $\mathcal{L}_j$. We kunnen daarvoor een eindig aantal ortogonale polynomen of een basis voor de splines van een bepaalde graad nemen, maar voor onze doeleinden is het handiger $\phi_j(\underline{h}_j)$ te benaderen met stap funkties. Dit betekent dat we voor iedere j een basis van indikator funkties gebruiken, met een eindig aantal elementen in de basis. Dit komt op hetzelfde neer als diskrete stochastische variabelen die maar een eindig aantal waarden aannemen, en die gekodeerd worden als "dummy variabelen" of indikator matriksen.

In dit speciale geval vereenvoudigt er van alles, zowel wat interpretatie als wat berekeningen betreft. Als we k_j kategorieen hebben voor variabele j, dan is het handig om de matriks C weer te geven als in tabel 3f op bladzijde 52. We veranderen de indeksen, en we beseffen dat indikatoren wel ortogonaal maar niet noodzakelijk ortonormaal zijn. Daardoor vinden we

$$B(\underline{y}_1,\ldots,\underline{y}_m) = \frac{1}{m}\sum_{j=1}^{m}\sum_{\ell=1}^{m} a_j' C_{j\ell} a_\ell,$$

$$T(\underline{y}_1,\ldots,\underline{y}_m) = \sum_{j=1}^{m} a_j' D_j a_j.$$

Het is hier dus eenvoudiger om C en a te partioneren volgens variabelen in plaats van volgens dimensies in de basis. In ieder geval zijn de stationaire vergelijkingen aan het einde van de eerste weg $Cy = m\lambda Dy$, en is het eigenwaarde probleem aam het einde van de tweede weg $\frac{1}{m}\sum P_j(\underline{z}) = \lambda\underline{z}$, waarbij P_j de orthogonale projektor op de deelruimte omspannen door de k_j vektoren $g_{jr}(\underline{h}_j)$ is. Geometrische interpretaties van deze procedure worden in een ander deel van deze klapper gegeven. We beperken ons hier tot de analytische benadering waarin we <u>of</u> de optimale $\phi_j(\underline{h}_j)$ benaderen met behulp van een stapfunktie met k_j stappen, <u>of</u> de optimale $\phi_j(\underline{h}_j)$ vinden in het geval $\underline{h}_j$ diskreet is en slechts k_j verschillende waarden aanneemt.

De hier besproken techniek is (één-dimensionale) HOMALS. Het korresponderende komputerprogramma geeft als belangrijkste output de (benaderingen van de) optimale transformaties, en de optimale $\underline{z}$. HOMALS bewandelt de tweede weg, dat wil zeggen dat $\underline{z}$ genormaliseerd wordt met $V(\underline{z}) = 1$ en dat a vrij is. In het programma is z natuurlijk een n-vektor, met $SSQ(z) = n$, en wordt de optimale transformatie gegeven door $G_j a_j$, met G_j een $n \times k_j$ indikator matriks, en met $a_j = D_j^{-1} G_j' z$. HOMALS berekent ook de zogenaamde <u>diskriminatiematen</u> $a_j' D_j a_j = z' P_j z$. Deze maten generaliseren de komponent ladingen van variabelen uit het lineaire wegen, of liever gezegd de kwadraten van de ladingen. Evenals bij lineaire weging is de diskriminatiemaat een getal tussen nul en één, en evenals bij lineaire weging is de homogeniteit gelijk aan de gemiddelde diskriminatiemaat. Een andere interpretatie is dat de diskriminatiemaat gelijk is aan het kwadraat van de item-totaal korrelatie, berekend tussen getransformeerde variabelen.

In een aantal theoretische gevallen is het niet nodig om te benaderen met een eindige basis omdat we ook zo wel kunnen zeggen wat er uit komt. Dit is met name eenvoudig wanneer we onze basis zo kunnen kiezen dat

$$R(g_{js}(\underline{h}_j), g_{\ell t}(\underline{h}_\ell)) = \delta^{st} r^{s}_{j\ell}.$$

Dit betekent dat alle bivariate verdelingen gediagonaliseerd worden, en de kondities waaronder dit mogelijk is zijn bekend. Merk op dat s bij $r_{j\ell}^s$ een index is, en geen macht. Als aan deze konditie voldaan is, dan geldt natuurlijk

$$B(\underline{y}_1,\ldots,\underline{y}_m) = \frac{1}{m}\sum_{s=1}^{\infty} a_s' R_s a_s,$$

en dus

$$\sigma(*;*) = 1 - \max_{s=1}^{\infty} \lambda_+(R_s).$$

Het grote voordeel is dus, dat we al de kleine matriksjes R_s apart kunnen bekijken, we kunnen a_s altijd kiezen als een eigenvektor van één van de a_s, en dan geldt $a_t = 0$ voor alle $t \neq s$. De optimale transformaties zijn dus altijd van de vorm $a_{js}g_{js}(\underline{h}_j)$, en als de g_{js} ortogonale polynomen zijn, dan zijn de optimale transformaties dus allemaal polynomen van dezelfde graad.

Het meest bekende voorbeeld van deze vereenvoudigde theorie is de multinormaalverdeling. Bivariate normaalverdelingen kunnen gelijktijdig gediagonaliseerd worden als we voor onze bases de Hermite-Chebyshev polynomen gebruiken. We gaan niet in op de definitie van deze polynomen, daarvoor verwijzen we naar ieder willekeurig boek over orthogonale polynomen. Tricomi (1955) is bijvoorbeeld handig, daar vinden op bladzijde 254 ook de formule van Mehler, die aantoont dat voor Hermite-Chebyshev polynomen op de bivariate normaalverdeling geldt $r_{ij}^s = (r_{ij})^s$. We gebruiken hier dus eerst s als indeks, en daarna s als macht. Stel $R^{(s)}$ is de matriks met s-de machten van korrelaties. We weten dat $\lambda_+(R^{(1)}) \geq \lambda_+(R^{(2)}) \geq \ldots$, zie bijvoorbeeld Styan (1973, corollary 3.1, pag 224). Voor de normaalverdeling vinden we dus $\sigma(*;*) = 1 - \lambda_+(R)$, en het minimum wordt aangenomen voor $\phi_j(\underline{h}_j) = a_j\underline{h}_j$, met andere woorden: lineair en niet-lineair wegen geeft hetzelfde resultaat. Maar misschien belangrijker is het volgende: als $\underline{h}_j = \psi_j(\underline{e}_j)$, waarbij $\underline{e}_1,\ldots,\underline{e}_m$ multinormaal, en ψ_j één-éénduidig, dan vindt onze procedure optimale transformaties $\phi_j(\underline{h}_j) = a_j\psi_j^{-1}(\underline{h}_j) = a_j\underline{e}_j$. Met andere woorden: als we de multinormaal-verdeling verknoeien met bijvoorbeeld niet-lineaire maar wel monotone transformaties, dan vind onze niet-lineaire wegingstechniek de oorspronkelijke multinormale variabelen terug. Dit is een bizonder geval van de stelling dat onze techniek inderdaad niet-lineair is, en dus dezelfde resultaten geeft onder één-éénduidige transformaties van de variabelen.

2.1.6 Meer geschiedenis

De geschiedenis van de niet-lineaire wegingstechniek is niet eenvoudig weer te geven. In het bivariate geval is er al veel vroeg werk van Pearson dat direkt relevant is, maar de introduktie van deze techniek als schaalkonstruktie methode moeten we toeschrijven aan Richardson. Hij stelde de "method of reciprocal averages" voor, waarin de individuen als schaalwaarde de gemiddelden van de schaalwaarden van de antwoordkategorieen die ze aangestreept hebben krijgen, en waarbij de antwoordkategorieen als schaalwaarde de gemiddelden van de schaalwaarden van de individuen waardoor ze aangestreept zijn krijgen. Lees deze laatste zin nog maar eens over. Niet alleen wordt er hier een vorm van "interne konsistentie" gedefinieerd, maar bovendien kan men de zin ook lezen als een definitie van een iteratief algoritme. Hetzelfde geldt voor de twee stappen van onze minimalisatiemethoden, waar ook gemiddelden en konditionele verwachtingen afgewisseld worden. Als de gebruikte basis uit een eindig aantal indikatoren bestaat zijn de voorwaardelijke verwachtingen heel eenvoudig te berekenen, en definieren onze twee stappen een generalizatie van de "method of reciprocal averages". Richardson heeft zelf overigens niets hierover gepubliceerd, we ontlenen onze informatie aan Horst (1935). Deze zelfde Horst (1935, 1936) stelde voor om $V(\sum_j \phi_j(h_j))$ te maksimaliseren met behulp van een basis van indikatoren. Dat wil zeggen: zoiets zou je kunnen denken, maar in feite stelt Horst voor om zijn lineaire wegingstechniek toe te passen op dummy variabelen, omdat dit leidt tot iets wat veel lijkt op Richardson's methode. Er wordt in het geheel niet gesproken over benaderen van kontinue niet-lineaire transformaties. Ook niet bij Guttman (1941), hoewel Guttman in dit terecht befaamde artikel een stap vooruit doet door niet-lineaire PCA te relateren aan chi kwadraat analyse. Ook rekenkundig is het 1941 artikel bijna volmaakt, Guttman onderscheidt keurig de drie wegen met hun stappen. Latere bijdragen in dezelfde sfeer zijn Johnson (1950), Lord (1958), Bock (1960), De Leeuw (1973), en Nishisato (1978).

De nadruk op niet-lineaire transformaties komt uit de Franse school, en wordt bijvoorbeeld uitvoerig uiteengezet in Dauxois en Pousse (1976), en in Lafaye de Michaux (1978). Het resultaat voor de multivariate normaalverdeling is gebaseerd op grote hoeveelheden klassiek werk, maar in deze vorm komt het het eerst voor bij De Leeuw (1973), zie ook Hill (1974). In het laatstgenoemde artikel van Hill zien we trouwens de "method of reciprocal averages" plotseling weer populair worden door toepassingen in de ekologie en archeologie.

2.1.7. Multidimensionale transformaties

Over meerdimensionale uitbreidingen van de HOMALS theorie kunnen we kort zijn, om meerdere redenen. In de eerste plaats komen ze uitvoerig aan de orde in het geometrische deel van de klapper. En in de tweede plaats zijn de generalizaties niet erg ingewikkeld. We bespreken alleen het niet-lineaire geval, en wel de zogenaamde meervoudige kwantifikatie. Enkelvoudige kwantifikatie komt elders aan de orde. Het verlies is

$$\sigma(\underline{z};\Phi) = \frac{1}{m}\sum_{j=1}^{m} V(\underline{z} - \phi_j(\underline{h}_j)),$$

zoals gewoonlijk, alleen zijn $\underline{z}$ en $\phi_j(\underline{h}_j)$ nu vektor variabelen, en gebruiken we $V(\underline{x})$ voor de dispersie matriks. Onze verliesfunktie heeft nu als waarden dus symmetrische, positief semi-definiete matriksen in plaats van niet-negatieve getallen. Dat doet misschien wat vreemd aan. In de gebruikelijke HOMALS en PRINCALS theorie, zoals uiteengezet in hoofdstuk 3 en hoofdstuk 5 van deze klapper, nemen we als verlies dan ook het spoor van de verlies-matriks uit deze paragraaf. Maar in principe is dat spoor een nogal willekeurige keuze, en we kunnen laten zien dat we de meeste resultaten af kunnen leiden voor een veel algemenere klasse van verliesfunkties. We doen dat hier niet in detail, maar we komen op het onderwerp terug als we HOMALS en PRINCALS vergelijken in hoofdstuk 5.

Als eerste stap in onze meervoudige kwantifikatie gebruiken we de ongelijkheid

$$\sigma(z;\Phi) \geqslant \frac{1}{m}\sum_{j=1}^{m} V(\underline{y}_j - \underline{y}_\cdot) =$$

$$= \frac{1}{m} T(\underline{y}_1,\ldots,\underline{y}_m)\{I - T^{-1}(\underline{y}_1,\ldots,\underline{y}_m)B(\underline{y}_1,\ldots,\underline{y}_m)\},$$

waarbij $\underline{y}_j = \phi_j(\underline{h}_j)$, zoals gebruikelijk. Strikt genomen hebben we niet gedefinieerd wat we met $\geqslant$ bedoelen, maar we kunnen altijd de interpretatie in termen van het spoor gebruiken. We kunnen nu dus $\sigma(*;\Phi)$ definieren als de matriks aan de rechterkant van de ongelijkheid, en we maken de tweede stap op deze weg door $B(\underline{y}_1,\ldots,\underline{y}_m)$ zo groot mogelijk te maken op voorwaarde dat $T(\underline{y}_1,\ldots,\underline{y}_m) = mI$.

Het spreekt min of meer vanzelf dat wanneer we een basis gebruiken we een formule krijgen van de vorm

$$B(\underline{y}_1,\ldots,\underline{y}_m) = \frac{1}{m}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty} A_s' C_{st} A_t,$$

$$T(\underline{y}_1,\ldots,\underline{y}_m) = \sum_{s=1}^{\infty} A_s' A_s.$$

De tweede stap komt dan neer op het vinden van de p grootste eigenwaarden van C, en de p bijbehorende eigenvektoren, waarbij p de gekozen dimensionaliteit is. De taak wordt aanzienlijk eenvoudiger wanneer $C_{st} = 0$ voor $s \neq t$. In dat geval

berekenen we alle eigenwaarden van alle C_{ss} en kiezen daar de p grootsten uit. In de multinormaalverdeling zijn er bij p = 2 bijvoorbeeld twee mogelijkheden (die geillustreerd worden met een voorbeeld in hoofdstuk 5). Het is altijd zo dat $\lambda_+(R)$ de grootste van alle eigenwaarden is, maar de tweede grootste kan zowel de tweede grootste eigenwaarde van R als wel de grootste eigenwaarde van $R^{(2)}$ zijn. In het eerste geval zijn beide transformaties lineair, in het tweede geval is de eerste dimensie lineair en de tweede kwadratisch. De eerste twee-dimensionale schaal bestaat uit twee onafhankelijk normaal verdeelde variabelen, en bij de tweede twee-dimensionale schaal is de regressie van de tweede op de eerste dimensie kwadratisch. Het tweede geval blijkt in de praktijk aanzienlijk vaker voor te komen, en het leidt tot de zogenaamde "hoefijzers".

Als we bases gebruiken met een eindig aantal elementen, dan heeft het zin de elementen van C anders te organiseren. We vinden dan

$$B(\underline{Y}_1,\ldots,\underline{Y}_m) = \frac{1}{m}\sum_{j=1}^{m}\sum_{\ell=1}^{m} A_j'C_{j\ell}A_\ell,$$

$$T(\underline{Y}_1,\ldots\underline{Y}_m) = \sum_{j=1}^{m} A_j'A_j.$$

In dit geval kunnen we eenvoudig laten zien dat de grootheid

$$\omega^2 = \sum(m\,\lambda_s - 1)^2,$$

waarbij gesommeerd wordt over <u>alle</u> eigenwaarden ongelijk aan nul gelijk is aan de som van de chi kwadraten van alle $\binom{m}{2}$ kruistabellen in de supermatriks C. We gaan hierbij uit van een basis van indikatoren, het bewijs staat in De Leeuw, 1973, pag 32. Dit laat nogmaals zien dat HOMALS de totale <u>bivariate</u> afhankelijkheid onderzoekt. Op dezelfde manier geldt in het oneindige geval dat ω^2 gelijk is aan de som van de $\binom{m}{2}$ gemiddelde kwadratische kontingenties van de bivariate verdelingen.

Over de geschiedenis van de meerdimensionale generalisaties kunnen we kort zijn. Natuurlijk was het iedereen die de eendimensionale variant ontdekte duidelijk dat er nog meer eigenwaarden met bijbehorende oplossingen bestonden. Guttman (1941) benadrukte al dat men met die overige oplossingen voorzichtig moet zijn, ze moeten anders behandeld worden als in het geval van lineaire PCA. Dit wordt bijvoorbeeld geillustreerd door het voorbeeld van de multinormaalverdeling, maar Guttman zelf denkt natuurlijk eerder aan de perfekte schaal (zie 2.2.3). Hierover ontstond een kleine polemische uitwisseling met Burt (Guttman, 1953, Burt, 1953), dit naar aanleiding van Burt (1950), waarin met de gebruikelijke hoeveelheid retoriek, eruditie, en irrelevantie aangetoond wordt dat de matriks C in een PCA programma gestopt kan worden. Dat is juist, maar niet erg interessant. De waarschuwingen van Guttman zijn ook juist, maar wel interessant.

2.2 Binaire gegevens

2.2.1 Inleiding

Het getal twee speelt een bizondere rol in deze klapper. Onze verliesfunkties zijn gebaseerd op het principe van de kleinste kwadraten, en onze plaatjes zijn meestal twee dimensionaal. Bovendien zullen we zien dat onze PCA technieken vereenvoudigen wanneer $m = 2$, en ook wanneer $k_j = 2$ voor alle j. In de volgende paragrafen gaan we in op dit laatste geval, PCA van binaire variabelen. We konsentreren ons op toepassingen die op de test-theorie geinspireerd zijn.

Waarom zijn binaire variabelen zo bizonder. Omdat er als $k_j = 2$ geen verschil is tussen lineaire en niet-lineaire transformaties Alle transformaties zijn lineair, want door twee punten kunnen we altijd een rechte lijn trekken. Onze homogeniteitsmaat is dus, evenals in het lineaire geval, λ_+, waarbij λ_+ de grootste eigenwaarde is van $\frac{1}{m} R$, en waarbij R de matriks van phi-koeffi-cienten is. Voor de volledigheid noemen we hier even de definitie en belangrijkste eigenschappen van de phi-koefficient.

Als $\underline{h}_j$ en $\underline{h}_\ell$ de waarden 0 en 1 aan kunnen nemen, dan definieren we $\pi_j \triangleq E(\underline{h}_j)$, $\pi_\ell \triangleq E(\underline{h}_\ell)$, en $\pi_{j\ell} \triangleq E(\underline{h}_j\underline{h}_\ell)$. Dan geldt $V(\underline{h}_j) = \pi_j(1 - \pi_j)$, $V(\underline{h}_\ell) = \pi_\ell(1 - \pi_\ell)$, en $C(\underline{h}_j, \underline{h}_\ell) = \pi_{j\ell} - \pi_j\pi_\ell$. De phi-koefficient is dan natuurlijk $\phi_{j\ell} = R(\underline{h}_j, \underline{h}_\ell) = C(\underline{h}_j, \underline{h}_\ell) / V^{\frac{1}{2}}(\underline{h}_j)V^{\frac{1}{2}}(\underline{h}_\ell)$. Een belangrijk verschil tussen $\phi_{j\ell}$ en de meer gebruikelijke produkt moment korrelatiekoefficienten is dat $\phi_{j\ell}$ boven- en benedengrenzen heeft die afhangen van de univariate marginalen π_j en π_ℓ. Stel $\pi_j \leq \pi_\ell$, en definieer voor iedere j de grootheid $\upsilon_j \triangleq \{\pi_j/(1 - \pi_j)\}^{\frac{1}{2}}$. Omdat $0 \leq \pi_{j\ell} \leq \pi_j$ geldt ook

$$-\upsilon_j\upsilon_\ell \leq \phi_{j\ell} \leq \upsilon_j/\upsilon_\ell.$$

Omdat psychometrici van nature nogal bang zijn voor lage korrelaties wordt in de literatuur als associatiemaat wel ϕ/ϕ_{max} voorgesteld, waarbij we zojuist gezien hebben dat $\phi_{max} = \upsilon_j/\upsilon_\ell$. Er zijn overigens ook nog genoeg andere redenen om ϕ te vergelijken met ϕ_{max}, die komen in het verdere verloop van dit hoofdstuk aan de orde wanneer we bekijken wat voor matriksen van ϕ-koefficienten uit verschillende interessante populatiemodellen komen rollen.

2.2.2 Monotone latente trek modellen

Stel dat er een niet direkt observeerbare, een-dimensionale stochastische variabele $\underline{x}$ is, met verdelingsfunktie $F(x)$, zodanig dat $p_j(x) \triangleq E(\underline{h}_j \mid \underline{x} = x)$ een stijgende funktie van x is. Natuurlijk geld

$$\pi_j = \int p_j(x)\,dF(x).$$

We nemen bovendien voor paren variabelen een vorm van voorwaardelijke onafhankelijkheid aan. Voor alle $j \neq \ell$ geldt

$$E(\underline{h}_j \underline{h}_\ell \mid \underline{x} = x) = p_j(x)p_\ell(x),$$

en dit impliceert

$$\pi_{j\ell} = \int p_j(x)p_\ell(x)\,dF(x).$$

Het is duidelijk dat

$$C(\underline{h}_j, \underline{h}_\ell) = \iint (p_j(x) - p_j(y))(p_\ell(x) - p_\ell(y))\,dF(x)\,dF(y),$$

en dus geldt $C(\underline{h}_j, \underline{h}_\ell) \geq 0$ en $\phi_{j\ell} \geq 0$. In monotone latente trek modellen zijn de korrelaties dus nooit negatief. Dit verbetert de benedengrens uit 2.2.1.

Stel nu bovendien dat de funkties $p_j(x)$ elkaar niet kruisen. Dus $p_j(x) \lesseqgtr p_\ell(x)$ voor één enkele x betekent $p_j(x) \lesseqgtr p_\ell(x)$ voor alle x. Een stel variabelen met een dergelijke eigenschap wordt door Mokken <u>holomorf</u> genoemd (Mokken, 1970). In holomorfe systemen kan men laten zien dat $p_j(x)$ in de vorm $p(x - \theta_j)$ geschreven kan worden, waarbij θ_j een reele parameter is. We verwijzen de geinteresseerde lezer naar werk van Levine (1970, 1972, 1975) voor een interessante analyse van holomorfe systemen. We volstaan hier met de opmerking dat $\theta_\ell \geq \theta_{\ell'}$ impliceert dat voor alle x geldt $p(x - \theta_\ell) \leq p(x - \theta_{\ell'})$, en dus zowel $\pi_\ell \leq \pi_{\ell'}$ als ook $\pi_{j\ell} \leq \pi_{j\ell'}$ voor alle j. Als $\pi_{j\underline{\ell}} \triangleq \pi_j - \pi_{j\ell}$, $\pi_{\underline{j}\ell} \triangleq \pi_\ell - \pi_{j\ell}$, en $\pi_{\underline{j}\underline{\ell}} \triangleq 1 - \pi_{j\ell} - \pi_{\underline{j}\ell} - \pi_{j\underline{\ell}}$, dan geldt bovendien dat $\pi_{j\underline{\ell}} \geq \pi_{j\underline{\ell}'}$, $\pi_{\underline{j}\ell} \leq \pi_{\underline{j}\ell'}$, en $\pi_{\underline{j}\underline{\ell}} \geq \pi_{\underline{j}\underline{\ell}'}$. Het bewijs voor deze eenvoudige relaties kan men bijvoorbeeld in het boek van Mokken vinden. Zonder verdere aannamen over de $p(x - \theta)$ te maken kunnen we nu niet veel meer zeggen. En daarom gaan we die ekstra aannamen nu maar maken.

2.2.3 <u>De Guttman schaal</u>

Bij het door Guttman (1944, 1950a,b) geintroduceerde model geldt

$$p_j(x) = \begin{cases} 0 \text{ als } x < \theta_j, \\ 1 \text{ als } x \geq \theta_j. \end{cases}$$

Hieruit volgt dat $\pi_{j\ell} = \min(\pi_j, \pi_\ell)$, en dus geldt altijd dat of $\pi_{j\underline{\ell}} = 0$ of $\pi_{\underline{j}\ell} = 0$. Stel we ordenen de variabelen zo dat $\theta_1 \geq \ldots \geq \theta_m$, dan $\pi_1 \leq \ldots \leq \pi_m$ en ook $v_j \leq \ldots \leq v_m$. Voor $j \leq \ell$ geldt dus $\phi_{j\ell} = v_j/v_\ell$, en voor $j \geq \ell$ geldt $\phi_{j\ell} = v_\ell/v_j$. Het is daardoor noodzakelijk (en ook voldoende) voor het bestaan van een perfekte schaal dat $\phi = \phi_{max}$ voor alle paren variabelen, o. dat $H \triangleq \phi/\phi_{max}$ gelijk is aan één voor alle paren. Gebruik van H voor Guttman

schaalbaarheid stamt van Loevinger (1947, 1948). Mokken (1970) lijkt $H_{j\ell}$ te gebruiken als een maat voor holomorfie van items. We zullen verderop zien dat die vlieger niet opgaat, er zijn perfekt holomorfe tests met arbitrair lage $H_{j\ell}$. Het MOKKEN SCALE programma moet dan ook gezien worden als een programma om Guttman schalen te konstrueren, niet als een programma dat meer algemene holomorfe verzamelingen selekteert. Maar dit terzijde.

We zijn nu geinteresseerd in de eigenwaarden en eigenvektoren van de matriks van ϕ-koefficienten. Guttman (1950b) geeft een briljante analyse, hoewel hij de benodigde resultaten eigenlijk beter uit de al lang bestaande mathematische literatuur over had kunnen schrijven. In Guttman (1954) wordt nader ingegaan op de interpretatie van de eigenvektoren. Dit is een wat esoterisch probleem, de eigenvektoren zijn immers allemaal welomschreven funkties van υ, en de vraag is of men dit soort funkties nog moet gaan zitten interpreteren ook.

We geven hier de belangrijkste resultaten over de korrelatiematriks en zijn eigenvektoren weer. Voor bewijzen verwijzen we naar Guttman (1950b), naar Gantmacher en Krein (1950), en Karlin (1964, 1968). In de eerste plaats is de inverse van de korrelatiematriks tri-diagonaal. In de tweede plaats dat als we de eigenvektoren uitzetten tegen υ, en we verbinden opeenvolgende punten, dan ontstaan er polygonale funkties met interessante eigenschappen. Funktie s, behorende bij eigenvektor s, dus bij de s-de eigenwaarde, heeft s - 1 nulpunten. Bovendien is het zo dat de nulpunten van de verschillende funkties verwoven zijn: tussen ieder paar naast elkaar liggende nulpunten van funktie s ligt precies één nulpunt van funktie s - 1. Maar dat is nog niet alles: als we ervoor zorgen dat de eigenvektor elementen behorend bij de kleinste υ voor alle eigenvektoren niet-negatief zijn, dan gelden dezelfde verwevings-eigenschappen voor de getransponeerde matriks van eigenvektoren. Het is duidelijk dat op basis van de eigenwaarden-eigenvektoren struktuur een Guttman korrelatiematriks goed herkenbaar is. Voor de volledigheid vermelden we ook nog even dat als alle υ_j verschillend zijn ook alle eigenwaarden verschillend zijn. En dat de door onze lineaire wegingsprocedure gevonden a_j (de voor variantie gekorrigeerde eerste eigenvektor) monotoon zijn met de υ_j, en dus met de θ_j. Omdat we bij gebrek aan informatie over F alleen de volgorde van de θ_j kunnen bepalen is dit voldoende. Niettemin is het wanneer we weten dat er een Guttman schaal in de gegevens

zit, aanzienlijk handiger om gebruik te maken van $-\ln \phi_{j\ell} = |\alpha_j - \alpha_\ell|$, met $\alpha_j \triangleq \ln \upsilon_j = \frac{1}{2}\,\text{logit}\,\pi_j$.
Een interessant speciaal geval is $\upsilon_j = \upsilon^j$, waarbij υ een getal tussen nul en één is. Dit impliceert dat $\phi_{j\ell} = \upsilon^{|j-\ell|}$. In dit geval zijn de eigenwaarden

$$\lambda_s = 2\sin^2\frac{s\pi}{2(m+1)},$$

en de eigenvektoren

$$t_{js} = \sin\frac{\pi sj}{m+1}.$$

Ook in andere speciale gevallen kunnen we eksplisiete oplossingen geven, die over het algemeen uit diskrete ortogonale polynomen bestaan.

Er wordt dikwijls gezegd dat de Guttman schaal niet realistisch, en daarom niet interessant is. Het lijkt ons dat deze konstatering berust op een verkeerd begrip van de rol van een model. Het model van Guttman is, evenals de Normaalverdeling en de Spearman hierarchie en de Staat van Plato en het Leven van Jezus, een norm. Zo zou het moeten gaan als het perfekt was. De tegenwerping dat het in werkelijkheid niet perfekt is, is helemaal geen tegenwerping.

2.2.4 De Spearman hierarchie.

Stel $p_j(x) = \alpha_j x + \beta_j$. Zonder verlies van algemeenheid kunnen we aannemen dat $E(\underline{x}) = 0$ en $V(\underline{x}) = 1$. Dan geldt dus $\pi_j = \beta_j$, en $C(\underline{h}_j, \underline{h}_\ell) = \alpha_j\alpha_\ell$. Als $\mu_j = \alpha_j / \{\beta_j(1-\beta_j)\}^{\frac{1}{2}}$, dan $\phi_{j\ell} = \mu_j\mu_\ell$. De Spearman hierarchie geeft dus een korrelatiematriks van de vorm $R = \mu\mu' + \Delta$ met Δ diagonaal en niet negatief. Als δ_j^2 het j-de diagonaalelement van Δ is, dan $\delta_j^2 = 1 - \mu_j^2$.

Op dit moment zou er iemand in de zaal onrustig kunnen worden, omdat hij het raar vind om een lineair model voor de $p_j(x)$ aan te nemen. Immers die $p_j(x)$ moeten tussen nul en één liggen, en een lineaire funktie trekt zich daar niets van aan. Maar stel nu eens dat $\underline{x}$ verdeeld is op een eindig interval, en dat op dat eindige interval alle $p_j(x)$ tussen nul en één liggen. Dat kan wel. En als $\underline{x}$ nu eens verdeeld is op een heel klein interval, dan is het niet alleen heel goed mogelijk dat de $p_j(x)$ tussen nul en één liggen, maar bovendien is het zo dat in dat kleine iedere differentieerbare $p_j(x)$ heel goed lineair benadert kan worden. Als $\underline{x}$ met een zeer kleine variantie σ^2 rond μ verdeeld is dan kunnen we bewijzen dat $C(\underline{h}_j, \underline{h}_\ell) \approx \sigma^2 p_j'(\mu)p_\ell'(\mu)$, wat betekent dat

het Spearman model bij benadering opgaat. Hetzelfde gebeurt als de items allemaal heel gemakkelijk of heel moeilijk zijn, de individuen zitten dan in de staarten van de $p_j(x)$, en die staarten zijn aardig lineair.

We hebben gezien dat bij het Guttman model de opeenvolgende eigenvektoren van de korrelatiematriks steeds meer toppen kregen. Hoe zit dit bij het Spearman model. Stel de diagonaalelementen van Δ zijn gerangschikt zodat $\delta_1^2 \geq \ldots \geq \delta_m^2$. Dan geldt voor de eigenwaarden $\lambda_1 \geq \ldots \geq \lambda_m$ van $R = \mu\mu' + \Delta$ dat $\mu'\mu + \delta_1 \leq \lambda_1$, voor $s=2,\ldots,m-1$ geldt $\delta_{s-1}^2 \leq \lambda_s \leq \delta_s^2$, en tenslotte $\lambda_m \leq \delta_m^2$. Deze resultaten worden op zeer vele plaatsen bewezen, een handige recente referentie is Bunch, Nielsen, en Sorensen (1978). Als t_s de eigenvektor is behorend bij λ_s, we nemen aan dat t_s genormeerd is op $t_s'\mu = 1$, dan

$$t_{js} = \frac{\mu_j}{\lambda_s - \delta_j^2}.$$

Als we t_{js} plotten tegen j krijgen we interessante plots. Voor de eerste eigenvektor is de plot monotoon en niet-negatief als $\mu \geq 0$. Voor de overige eigenvektoren is er slechts één tekenverandering, de plot gaat van hoog positief naar zeer laag negatief. De plaats waar de sprong plaats vind wordt steeds een plaats opgeschoven, in de stukken waar niet gesprongen wordt is de funktie monotoon stijgend. Het patroon is zeer karakteristiek, en goed te onderscheiden van het patroon van de Guttman schaal. Natuurlijk is de Spearman hierarchie weer niet bedoeld als een realistisch deskriptief model, In de praktijk blijkt dat de meeste gegevens tussen Spearman en Guttman inzitten. Guttman zegt wat er gebeurt als het perfekt gaat, Spearman zegt wat er gebeurt als het slecht gaat, dat wil zeggen als er vrijwel geen diskriminatie mogelijk is.

2.2.5 Het model van Rasch

Bij het model van Rasch is $\underline{x}$ verdeeld op de niet-negatieve getallen, en is $p_j(x)$ van de vorm

$$p_j(x) = \frac{x}{x + \theta_j}.$$

Wat het model van Rasch zo elegant maakt, en het weer min of meer tot norm verheft, is het volgende resultaat. Eenvoudige integratie laat zien dat

$$\pi_{j\ell} = \frac{\pi_j\theta_j - \pi_\ell\theta_\ell}{\theta_j - \theta_\ell}.$$

Uit deze formule volgt in de eerste plaats dat $\pi_{j\ell}/\pi_{j\underline{\ell}} = \theta_j/\theta_\ell$,

en op basis van deze laatste formule kunnen we de itemparameters schatten zonder dat we iets aannemen over de verdeling van x. Rasch (1966a, b) noemt dit specifieke objektiviteit, en laat zien dat het karakteristiek is voor zijn model. Enig doorrekenen geeft ook een relatief eenvoudige formule voor de phi koefficient.

$$\phi_{j\ell} = \frac{1}{\theta_j - \theta_\ell}\{\theta_j(\frac{\upsilon_j}{\upsilon_\ell}) - \theta_\ell(\frac{\upsilon_\ell}{\upsilon_j})\}.$$

Aan deze formule is te zien dat als de items zeer ver uit elkaar liggen het Rasch model zich gedraagt als het Guttman model, terwijl we hebben gezien dat allemaal makkelijke of allemaal moeilijke items tot het Spearman model leiden.

2.2.6 Enige voorbeelden

We hebben twee voorbeelden met ieder 9 variabelen gegenereerd volgens zowel het Guttman als het Rasch model. We nemen daarbij aan dat $F(x) = 1 - \exp(-x)$. In voorbeeld 1 kiezen we θ volgens .01(.01).09, in het tweede voorbeeld nemen we .05(.20)1.85. Om de π-waarden volgens Rasch te berekenen hebben we de exponentiele integraal nodig. We gebruiken formule 5.1.28 en ekspansie 5.1.11 uit Abramowitz en Segun (1968, pag 229-230). In tabel 2.1.a en b staan de waarden voor π. In 2.2.a en b staan de korrelatie matriksen voor het voorbeeld met de gemakkelijke vragen. In 2.2.c en 2.2.d staan de korrelatiematriksen voor de meer gespreide vragen. In tabel 2.3 staan de eigenwaarden van de vier korrelatie matriksen, en in figuur 2.2.a tm d staan de voor variantie gekorrigeerde eigenvektoren geplot. In figuur 2.3 tenslotte staan ter vergelijking de negen eigenvektoren van een Spearman matriks met μ gelijk aan .1(.1).9.

2.2.7 Niet-monotone één-dimensionale modellen

De modellen met stijgende $p_j(x)$ zijn geinspireerd op de testleer. Het idee wat eraan ten grondslag ligt is dat hoe beter iemand is, hoe groter de kans is dat hij het goede antwoord geeft. Maar stel nu eens dat x de politieke links-rechts dimensie is, en $p_j(x)$ is de kans dat iemand met positie x stemt voor "Nederland uit de NAVO". Een monotone $p_j(x)$ lijkt geen vreemde aanname. Maar neem nu eens een voorstel over bijvoorbeeld abortus. We kunnen verwachten dat het cofessionele midden ertegen is, en dat zowel links als rechts ervoor zijn. Door de formulering om te keren is het midden er oor en zijn links en rechts er tegen, dus is $p_j(x)$ ééntoppig. Het prototype van dit soort modellen de Coombs schaal, waarbij $p_j(x) = 1$ als $\theta_j < x \leq \theta_j + \delta$ en $p_j(x) = 0$ in andere gevallen.

θ	Rasch	Guttm
.01	.9592	.9900
.02	.9316	.9802
.03	.9085	.9704
.04	.8884	.9608
.05	.8703	.9512
.06	.8538	.9417
.07	.8385	.9324
.08	.8243	.9231
.09	.8111	.9139

tabel 2.1a: π-waarden gemakkelijke vragen

θ	Rasch	Guttm
.25	.6648	.7788
.45	.5587	.6376
.65	.4846	.5220
.85	.4352	.4274
1.05	.3943	.3499
1.25	.3612	.2865
1.45	.3338	.2346
1.65	.3106	.1920
1.85	.2966	.1572

tabel 2.1b: π-waarden gespreide vragen

.7071
.5755 .8138
.4976 .7036 .8647
.4437 .6275 .7711 .8918
.4039 .5712 .7019 .8118 .9103
.3733 .5278 .6486 .7502 .8412 .9241
.3482 .4924 .6051 .6998 .7848 .8621 .9329
.3274 .4630 .5690 .6581 .7379 .8106 .8772 .9403

tabel 2.2a: korrelaties Guttman gemakkelijk.

.2084
.2054 .2190
.2030 .2210 .2312
.1999 .2200 .2305 .2337
.1968 .2186 .2299 .2345 .2379
.1936 .2165 .2283 .2334 .2366 .2368
.1907 .2146 .2270 .2328 .2365 .2381 .2408
.1883 .2130 .2261 .2327 .2372 .2499 .2436 .2478

tabel 2.2b: korrelaties Rasch gemakkelijk.

.7069
.5569 .7878
.4604 .6513 .8267
.3910 .5531 .7020 .8492
.3377 .4777 .6064 .7335 .8637
.2951 .4174 .5298 .6408 .7546 .8737
.2598 .3675 .4665 .5642 .6645 .7693 .8805
.2302 .3256 .4133 .4999 .5887 .6816 .7801 .8860

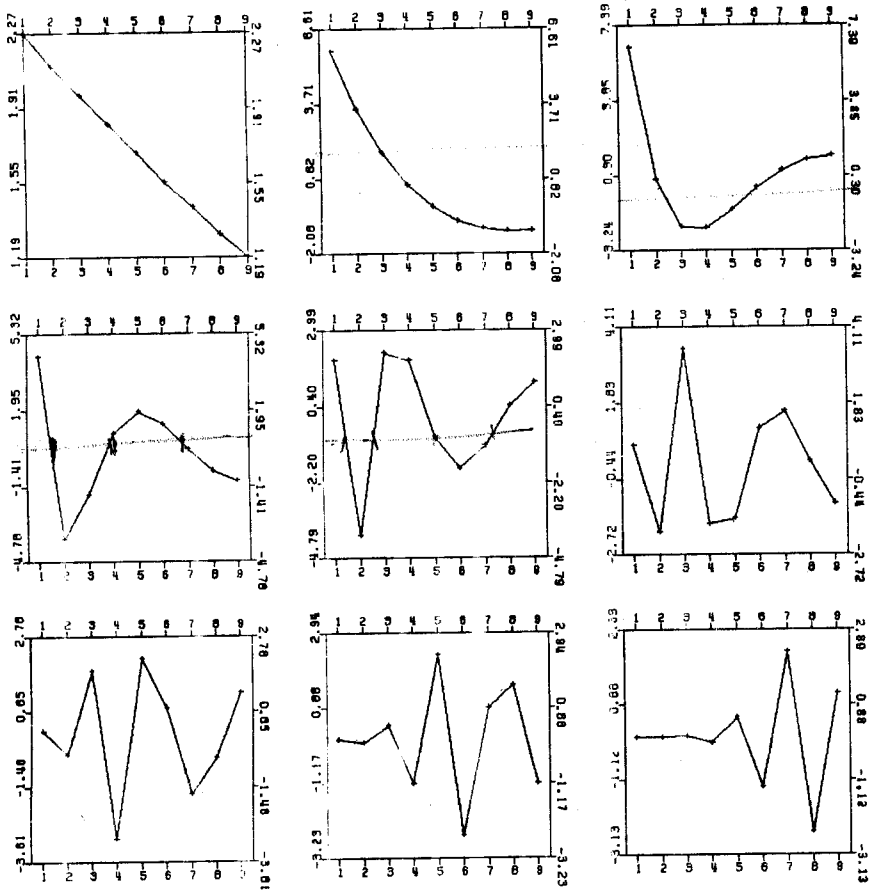
tabel 2.2c: korrelaties Guttman gespreid.

.2332
.2233 .2224
.2145 .2158 .2126
.2065 .2090 .2064 .2017
.1992 .2025 .2005 .1963 .1917
.1929 .1969 .1955 .1920 .1882 .1853
.1871 .1916 .1907 .1876 .1842 .1812 .1774
.1943 .2057 .2125 .2195 .2304 .2502 .2913 .4210

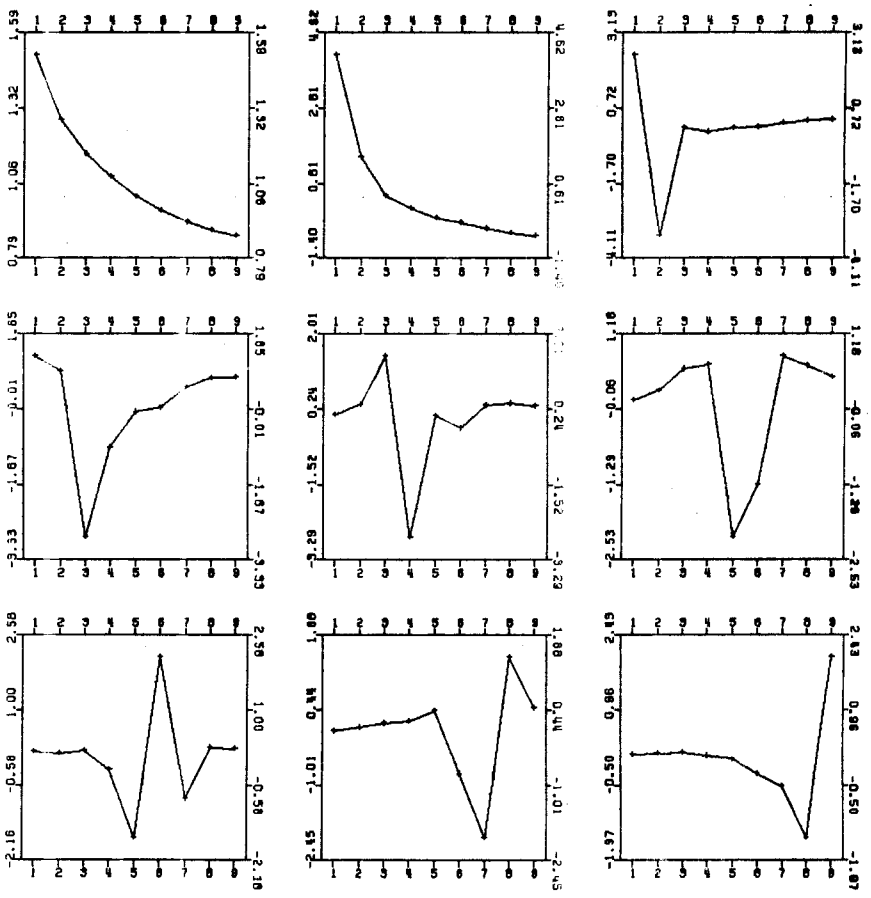
tabel 2.2d: korrelaties Rasch gespreid.

Gut Gem	6.5276	1.2544	.5058	.2680	.1647	.1091	.0762	.0548	.039
Ras Gem	2.7908	.8430	.7846	.7760	.7672	.7653	.7622	.7594	.751
Gut Ges	5.8406	1.5264	.6374	.3461	.2201	.1542	.1149	.0892	.071
Ras Ges	2.6992	.9763	.8202	.8117	.8023	.7913	.7802	.7665	.552

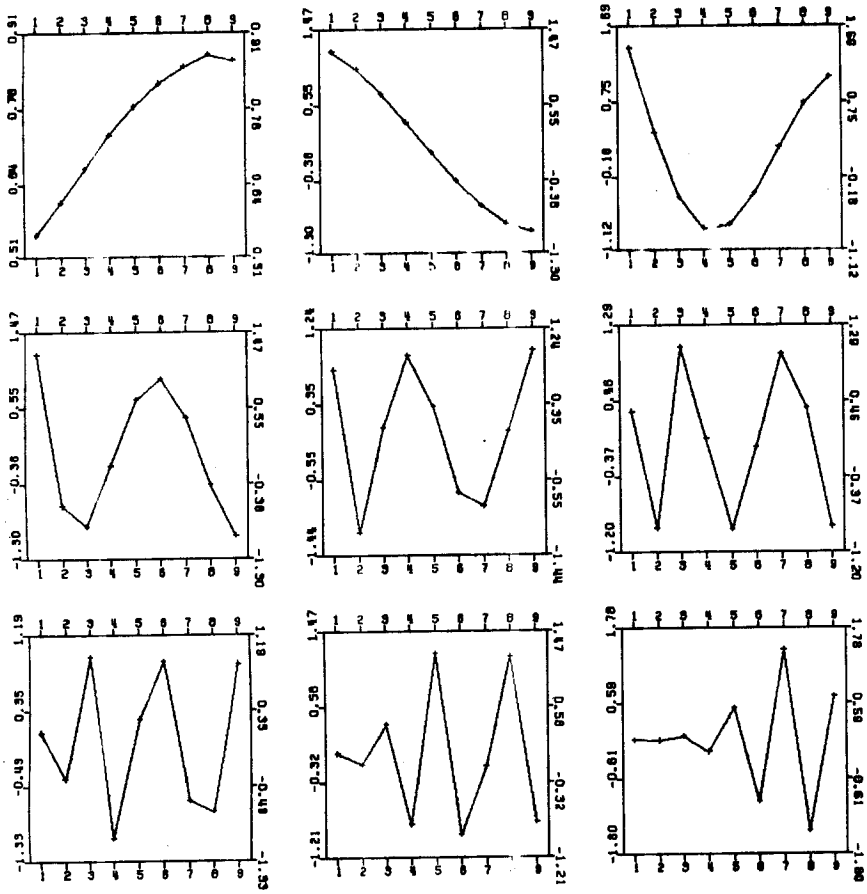
tabel 2.3: eigenwaarden vier korrelatiematriksen.



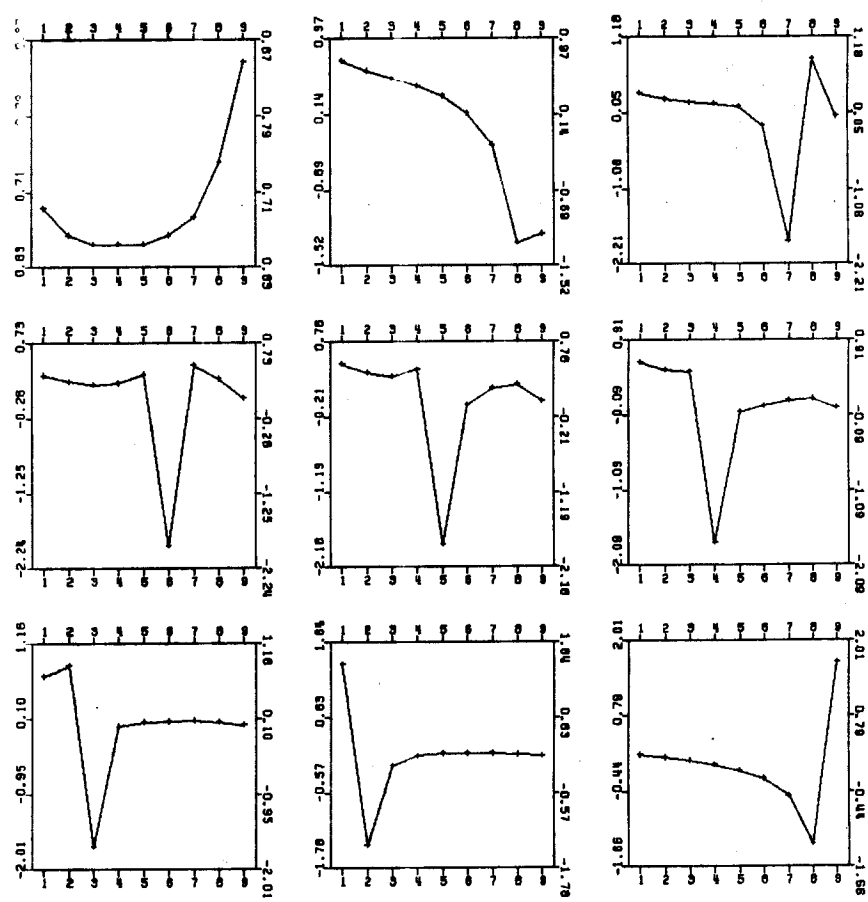
figuur 2.2a:eigenvektoren Guttman gemakkelijk



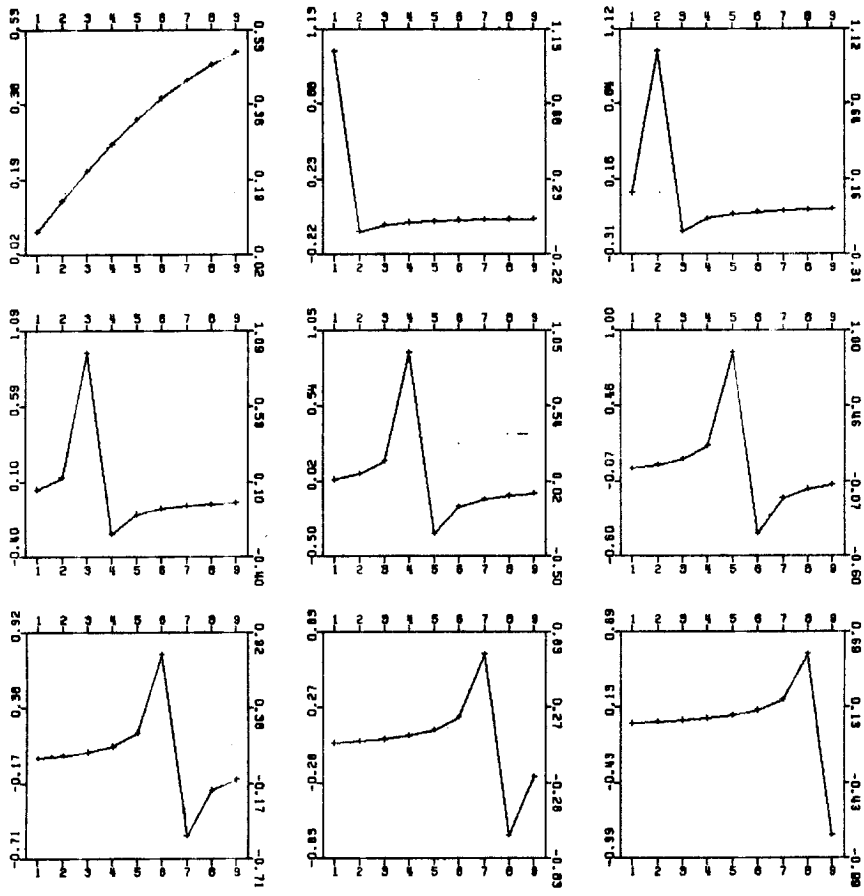
figuur 2.2b:eigenvektoren Rasch gemakkelijk



figuur 2.2cæigenvektoren Guttman gespreid



figuur 2.2d: eigenvektoren Rasch gespreid.



figuur 2.3: eigenvektoren Spearman matriks

Er zijn ook 'single-peaked' versies van het Rasch model. We gaan echter in dit hoofdstuk niet meer op deze uitbreidingen in, evenmin als op Guttman en Rasch modellen voor meerdere antwoordkategorieen, en evenmin als op meer-dimensionale modellen als de circumplex, radex of het conjunktief/disjunktief model. Verschillende van deze onderwerpen komen nog wel ter sprake als we verderop de relatie tussen HOMALS en diverse vormen van meerdimensionele schaalmethoden bespreken.

2.2.8 Gedichotomiseerde multinormaalverdelingen.

Er is nog tijd voor één toegift. Ook binaire variabelen, maar nu gegenereerd door een stel multinormaal verdeelde variabelen in tweeën te hakken. Dus we nemen aan dat $\underline{e}_1,\ldots,\underline{e}_m$ standaard normaal zijn, met korrelaties $\rho_{j\ell}$, en we definieren $\underline{h}_j = 1$ als $\underline{e}_j \geq \theta_j$, en $\underline{h}_j = 0$ als $\underline{e}_j < \theta_j$. Dit model werd in verband met PCA en FA het eerst bekeken door Lawley (1944), hoewel hij natuurlijk ruim gebruik maakte van de klassieke tetrachorische theorie van Pearson. In ieder geval kunnen we weer de Hermite Chebyshev polynomen ψ_s gebruiken, om te laten zien dat er een representatie bestaat van de vorm

$$R(\underline{h}_j,\underline{h}_\ell) = \sum_{s=1}^{\infty} \rho_{j\ell}^s \tau_{js}\tau_{\ell s}.$$

Hierbij is

$$\tau_{js} = (s!)^{-\frac{1}{2}}\,\phi(\theta_j)\psi_{s-1}(\theta_j)\{\Phi(\theta_j)(1-\Phi(\theta_j))\}^{-\frac{1}{2}},$$

en voor de gelegenheid is ϕ de dichtheid en Φ de verdeling van de standaardnormaalverdeling. Merk op dat de term tussen akkolades gelijk is aan $V(\underline{h}_j)$. Voor $s = 1$ vinden we

$$\tau_{j1} = \phi(\theta_j)V^{-\frac{1}{2}}(\underline{h}_j),$$

en voor $s = 2$

$$\tau_{j2} = \tfrac{1}{2}\sqrt{2}\;\theta_j\phi(\theta_j)V^{-\frac{1}{2}}(\underline{h}_j).$$

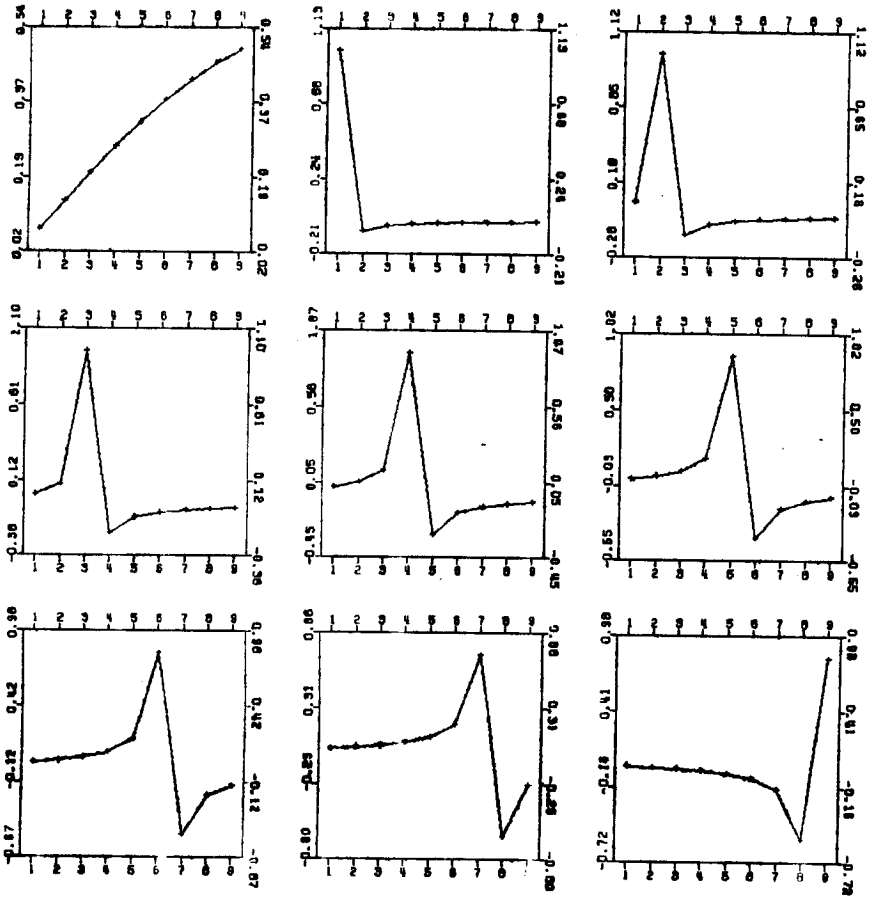
Veronderstel nu, met Lawley, dat $\rho_{j\ell} = \alpha_j\alpha_\ell$, en dat de α_j zo klein zijn dat we hun derde machten kunnen verwaarlozen. Dan geldt dus

$$R(\underline{h}_j,\underline{h}_\ell) \approx \alpha_j\alpha_\ell\tau_{j1}\tau_{\ell 1} + \alpha_j^2\alpha_\ell^2\tau_{j2}\tau_{\ell 2}.$$

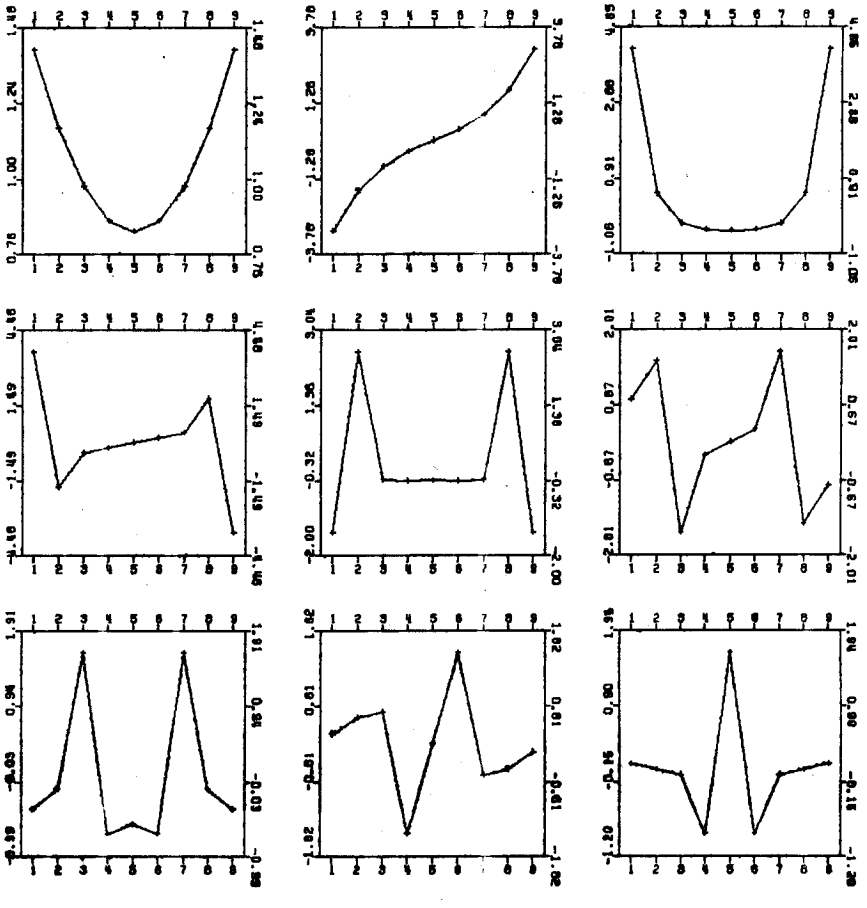
Omdat $\tau_{j2} > 0$ voor een moeilijke test, en $\tau_{j2} < 0$ voor een makkelijke test, zullen we uit PCA (of FA) dus een tweede faktor krijgen die afhangt van de moeilijkheid van de test. Een plot van de eerste twee faktoren zal, evenals bij de kontinue multinormaalverdeling en bij de Guttman schaal, een kwadratische struktuur opleveren. Een hoefijzer dus. Een voorbeeld vinden we in figuur 2.4.

In 2.4.a staat een voorbeeld met $\theta_j = 0$ voor alle j en met $a_j = .1(.1).9$. In 2.4.b staat een voorbeeld met $a_j = \frac{1}{2}\ 2$ voor alle j, en met $\theta_j = -2(\frac{1}{2})2$. In tabel 2.4.a en 2.4.b staan de korrelatiematriksen en hun eigenwaarden. We geven deze voorbeelden hier overigens zowel om illustratieve als om archivarische redenen. Het is altijd makkelijk dit soort voorbeelden bij de hand te hebben.

Alternatieve technieken voor dit soort gegevens zijn over het algemeen gebaseerd op het schatten van tetrachorische korrelaties. Hoe je die berekent staat recentelijk weer eens uiteengezet in Divgi (1979). Schattingen van de parameters van het één-faktor model worden besproken door Lawley (1944), maar moderner door Bock en Lieberman (1970). Voor het meer-faktor model verwijzen we naar Christofferson (1975, 1978) en Muthén (1978). Als de multinormaal verdeling in meerdere klassen ingedeeld is hebben we polychorische korrelatie nodig. Een interessant historisch overzicht, en een nieuwe schattingmethode, staan in Lancaster en Hamdan (1964). Het zoveelste bewijs dat een van de belangrijkste uitvoerprodukten van Zweden tegenwoordig partiele afgeleiden van likelihood funkties zijn wordt geleverd door Olsen (1979).



figuur 2.4.a:alle items even moeilijk, factorladingen verschillend



figuur 2.4.b:
faktorladingen gelijk,
moeilijkheidsgraden
verschillend.

.0127								
.0191	.0382							
.0255	.0510	.0766						
.0318	.0638	.0959	.1282					
.0382	.0766	.1152	.1543	.1940				
.0446	.0894	.1347	.1807	.2276	.2759			
.0510	.1023	.1543	.2074	.2620	.3187	.3784		
.0574	.1152	.1740	.2344	.2972	.3632	.4339	.5117	
2.5559	.9930	.9718	.9363	.8861	.8200	.7353	.6262	.4755

tabel 2.4.a: alle items even moeilijk, faktorladingen verschillend. korrelaties en eigenwaarden.

.1795								
.1799	.2362							
.1563	.2274	.2840						
.1264	.1962	.2642	.3136					
.0928	.1507	.2209	.2761	.3136				
.0600	.1074	.1645	.2209	.2642	.2840			
.0401	.0688	.1074	.1507	.1962	.2274	.2362		
.0231	.0401	.0600	.0928	.1264	.1563	.1799	.1795	
2.4572	1.2112	.8983	.8196	.7870	.7440	.7166	.6927	.6734

tabel 2.4.b: alle faktorladingen hetzelfde, moeilijkheidsgraad verschillend. korrelaties en eigenwaarden.

2.3 Bivariate verdelingen

Als m = 2 dan zijn er een groot aantal resultaten af te leiden, die voor m > 2 niet langer opgaan. De redenen waarom dit zo is zijn gekompliceerd, en spelen door de gehele algebra heen een grote rol.

2.3.1 Maten voor afhankelijkheid

In het nog steeds buitengewoon interessante artikel "On the theory of contingency and its relation to association and normal correlation definieert Pearson (1904) de gemiddelde kwadratische kontingentie $\phi^2 = \chi^2/N$, waarbij χ^2 de door hem eerder ingevoerde chi kwadraat maat voor de afwijking van een kruistabel van onafhankelijkheid is. Hij laat zien dat in een normaalverdeling met oneindig veel cellen geldt dat de korrelatieparameter ρ samenhangt met ϕ^2 volgens $\rho^2 = \phi^2/(1 + \phi^2)$. Pearson beschouwt deze laatste formule als een manier om ρ te berekenen als de binormaalverdeling in een groot aantal klassen gegroepeerd is. En het opmerkelijke is natuurlijk dat we nu ρ kunnen berekenen zonder dat we de schaalwaarden behorende bij de kategorieen kennen, dat wil zeggen we hebben een niet-lineaire manier om ρ te schatten, en we kunnen nu dus ook ρ schatten als de variabelen oogkleur of beroepssoort zijn. En dit laatste kwam Pearson vanwege zijn bekende kontinue metafysische bias prima uit (vergelijk onze opmerkingen over de Pearson-Yule kontroverse in hoofdstuk 1). In univariate situaties was Pearson de pionier van de niet-normale verdelingen, maar in bivariate situaties zien we hem zelden zonder gendarmes-muts op.

In een ander paper is Pearson (1913) geinteresseerd in de invloed van skoring op de gevonden korrelatie. Hij denkt daarbij vooral aan het samen nemen en verwisselen van kategorieen, omdat hij de theorie wil gebruiken voor nominale variabelen als haarkleur en oogkleur. En hierachter zit de overtuiging dat de korrelatiekoefficient in een bepaalde situatie een fysische konstante is, die de situatie uitputtend beschrijft. "I soon became convinced that owing to some important theoretical law hitherto overlooked, the order of the groups by which we classify our attributes is a matter of no importance when we are determining correlation." (l.c., 1904, pag 444). Pearson toont informeel aan dat lineaire regressie een voldoende voorwaarde is voor een maksimum van de korrelatie over hergroepering en herordening van de groepen. In moderne notatie wordt dit resultaat nogal triviaal: stel $\underline{x}$ en $\underline{y}$ zijn gestandariseerd (verwachting nul, variantie één), stel $E(\underline{y}|\underline{x}) = \rho\underline{x}$, dan geldt voor iedere η waarvoor $V(\eta(\underline{x}))$ eindig is dat $R(\eta(\underline{x}),\underline{y}) \leq \rho = R(\underline{x},\underline{y})$.

Een hele tijd later, in 1941, verscheen er een artikel van Gebelein waarin een nieuwe maat van afhankelijkheid gedefinieerd wordt. De korrelatiekoefficient, de korrelatieverhoudingen, en de gemiddelde kwadratische konvergentie hebben allemaal bepaalde nadelen, en Gebelein stelde als alternatief het maksimum van $E(\phi(\underline{x}),\psi(\underline{y}))$ voor, waarbij we eisen dat $E(\phi(\underline{x})) = E(\psi(\underline{y})) = 0$ en $V(\phi(\underline{x})) = V(\psi(\underline{y})) = 1$. Noem dit maksimum, als het bestaat, de <u>maksimale korrelatie</u> van $\underline{x}$ en $\underline{y}$, en schrijf het als $\kappa(\underline{x},\underline{y})$. Gebelein beperkt zich tot twee bekende speciale gevallen: of de bivariate verdeling is diskreet, of de bivariate verdeling heeft een kontinue dichtheidsfunktie. In beide gevallen berekenen we $\kappa(\underline{x},\underline{y})$ door de grootste eigenwaarde van een lineaire operator te berekenen. Gebelein laat zien dat de korrelatiekoefficient het bizondere geval is waarin we eisen dat ϕ en ψ beide lineair zijn (er valt dan niets te maksimaliseren). De korrelatieverhoudingen krijgen we wanneer we eisen dat of ϕ of ψ lineair is. Gebelein toont bovendien aan dat $\kappa(\underline{x},\underline{y}) = 0$ als en alleen als $\underline{x}$ en $\underline{y}$ onafhankelijk zijn, en dat funktionele afhankelijkheid impliceert dat $\kappa(\underline{x},\underline{y}) = 1$. Het is duidelijk dat $\kappa(\underline{x},\underline{y})$ minstens zo groot is als de korrelatie verhoudingen, die weer minstens zo groot zijn als de gewone korrelatie. Gebelein toont ook aan dat de gemiddelde kwadratische kontingentie minstens even groot is als $\kappa(\underline{x},\underline{y})$. In 1958 wordt de maksimale korrelatie herontdekt door Sarmanov (1958a,b). Hij voegt niets toe aan de theorie van Gebelein, behalve de stelling dat als beide regressies lineair zijn, dat dan de maksimale korrelatie gelijk is aan de gewone korrelatie. Dit is een tweezijdige variant van het resultaat van Pearson uit 1913, maar helaas is de stelling onjuist. Een tegenvoorbeeld staat in Sarmanov en Bratceva (1967, pag 478).

Gebelein's werk werd aanzienlijk gegeneraliseerd door Renyi (1959), en door een aantal van zijn leerlingen als Csaki en Fisher. Aan welke eisen willen we eigenlijk dat een behoorlijke afhankelijkheidsmaat voldoet ? Renyi noemt het volgende lijstje waaraan $\delta(\underline{x},\underline{y})$ best zou mogen voldoen.

A: $\delta(\underline{x},\underline{y})$ is gedefinieerd voor alle niet-konstante $\underline{x}$ en $\underline{y}$.

B: $\delta(\underline{x},\underline{y}) = \delta(\underline{y},\underline{x})$.

C: $0 \leq \delta(\underline{x},\underline{y}) \leq 1$.

D: $\delta(\underline{x},\underline{y}) = 0$ als en alleen als $\underline{x}$ en $\underline{y}$ onafhankelijk zijn.

E: Als $\underline{x} = \psi(\underline{y})$ of $\underline{y} = \phi(\underline{x})$ dan $\delta(\underline{x},\underline{y}) = 1$.

F: Als ϕ en ψ één-éénduidig zijn, dan $\delta(\phi(\underline{x}),\psi(\underline{y})) = \delta(\underline{x},\underline{y})$.

G: Als $(\underline{x},\underline{y})$ binormaal is, met korrelatie ρ, dan $\delta(\underline{x},\underline{y}) = |\rho|$.

Stel nu dat we vier verschillende associatiematen bekijken op deze zeven kriteria. De eerste maat is de absolute waarde van de korrelatiekoefficient, de tweede maat is het maksimum van de twee korrelatieverhoudingen, de derde maat is $\phi^2/1 + \phi^2$ waarbij ϕ^2 de gemiddelde kwadratische kontingentie is, en de vierde maat is de maksimale korrelatiekoefficient (gedefinieerd als sup, niet als max). We hebben dan de volgende resultaten:

	A	B	C	D	E	F	G
maat 1	0	1	1	0	0	0	1
maat 2	0	1	1	0	1	0	1
maat 3	0	1	1	1	0	1	1
maat 4	1	1	1	1	1	1	1

Men kan zich afvragen in hoeverre de maksimale korrelatie gegeneraliseerd kan worden naar meer dan twee variabelen, en in het bizonder hoe het precies zit met het onderscheid tussen ongekorrelleerdheid en onafhankelijkheid. Dit wordt uit de doeken gedaan door Lancaster (1959, 1960a, 1960b). Hij toont aan dat $\underline{x}_1,\ldots,\underline{x}_m$ onafhankelijk zijn als en alleen als $E(\phi_1(\underline{x}_1)\ldots\phi_m(\underline{x}_m)) = 0$ voor alle $\phi_j(\underline{x}_j)$ met $E(\phi_j(\underline{x}_j)) = 0$ en $V(\phi_j(\underline{x}_j)) = 1$. Natuurlijk kan men dus in het multivariate geval ook het supremum $\kappa(\underline{x}_1,\ldots,\underline{x}_m)$ definieren, en inderdaad heeft deze maksimale korrelatie ook de voor de hand liggende generalisaties van eigenschappen A t/m G. Merk op dat wij in HOMALS de maat sup $\lambda_+(\frac{1}{m} R)$, met $r_{j\ell} = E(\phi_j(\underline{x}_j),\phi_\ell(\underline{x}_\ell))$, gebruiken. Deze maat is nul als en alleen als de $\underline{x}_j$ paarsgewijs onafhankelijk zijn, de maat voldoet dus alleen aan ABC, EFG.

In het multivariate geval is het echter wel heel geforceerd om de samenhang in één getal samen te vatten. Een ander kriterium voor onafhankelijkheid is daarom nuttiger. Stel $\phi_{js}(\underline{x}_j)$ zijn gecentreerd en ortonormaal ($j=1,\ldots,m$; $s=1,\ldots\infty$). Definieer de gegeneraliseerde korrelaties $\rho_{t_1\ldots t_m} = E(\phi_{1t_1}(\underline{x}_1)\ldots\phi_{mt_m}(\underline{x}_m))$. Lancaster toont aan dat $\underline{x}_1,\ldots,\underline{x}_m$ onafhankelijk zijn als en alleen als $\rho_{t_1\ldots t_m} = 0$, oftewel als en alleen als $\sum\rho^2_{t_1\ldots t_m} = 0$. Maar $\sum\rho^2_{t_1\ldots t_m} = \phi^2(\underline{x}_1,\ldots,\underline{x}_m)$, de gemiddelde kwadratische kontingentie.

2.3.2 Skores

In 1935 publiceerde Hirschfeld een artikel waarin hij zich afvroeg of het altijd mogelijk was skores voor de rijen en kolommen van een kruistabel te vinden die beide regressies lineair maken. Als G de tabel is, van grootte $n \times m$, met $m \leq n$, en D en E zijn de diagonale matriksen met marginalen, dan willen we x en y vinden zodanig dat

$u'Dx = u'Ey = 0$ en $x'Dx = y'Ey = 1$ en zodanig dat $D^{-1}Gy = \rho x$ en $E^{-1}G'x = \lambda y$. Dit impliceert direkt dat $\rho = \lambda$. De laatste twee vergelijkingen hebben de triviale oplossing $x = y = u$, waarvoor $\rho = \lambda = 1$. Singuliere waarden dekompositie van $D^{-\frac{1}{2}}GE^{-\frac{1}{2}}$ vertelt ons dat er minstens $m - 1$ ortogonale oplossingen voor x en y bestaan, behorend bij $m - 1$ singuliere waarden, die allen tussen nul en één liggen. De grootste singuliere waarde, of preciezer de grootste niet-triviale singuliere waarde, is de maksimale korrelatiekoefficient uit de vorige paragraaf. Hirschfeld ttont tenslotte ook aan dat de kwadratensom van de singuliere waarden gelijk is aan de gemiddelde kwadratische kontingentie. De details worden overigens in het volgende hoofdstuk besproken. Gebelein (1941) leidde dezelfde resultaten af, maar liet tevens zien dat ze nog steeds opgingen in het geval waarin er een kontinue bivariate dichtheid bestaat.

Fisher (1940) herontdekt de techniek, hij noemt Hirschfeld tenminste niet, en hij is tevens de eerste die de techniek ook toepast. Of liever gezegd toe laat passen door Maung (1941a, 1941b). In het werk van Fisher en Maung spelen de drie verschillende wegen naar het maksimum al duidelijk een rol, ze leiden tot de interpretaties van maksimale kanonische korrelatie tussen indikator matriksen, van maksimale produktmoment korrelatie, en van maksimale diskriminatie tussen de kategorieen. Dit komt vrijwel tegelijkertijd vrijwel eksakt hetzelfde terug by Guttman (1941). Maung laat overigens ook zien dat voor een kontinue binormale verdeling de grootste singuliere waarde gelijk is aan $|\rho|$. Vergelijk hiervoor ook Lancaster (1957). Verdere toepassingen van Fisher's ideeen kan men vinden in Yates (1948), Johnson (1950), Williams (1952), en Bock (1960).

In zijn verhaal over de maksimale korrelatie laat Renyi zien dat er in het algemene geval niet altijd funkties $\phi(\underline{x})$ en $\psi(\underline{y})$ bestaan zodat $\kappa(\underline{x},\underline{y}) = R(\phi(\underline{x}),\psi(\underline{y}))$, dus het maximum wordt niet altijd aangenomen. We zoeken, in het algemene geval, naar oplossingen van de vergelijkingen $E(\phi(\underline{x})|\underline{y}) = \kappa\psi(\underline{y})$ en $E(\psi(\underline{y})|\underline{x}) = \kappa\phi(\underline{x})$. Of, wat hetzelfde is, van de vergelijkingen $E(E(\phi(\underline{x})|\underline{y})|\underline{x}) = \kappa^2\phi(\underline{x})$ en $E(E(\psi(\underline{y})|\underline{x})|\underline{y}) = \kappa^2\psi(\underline{y})$. Deze laatste twee vergelijkingen zijn eigenwaardenproblemen voor begrensde, symmetrische, positief semi-definiete operatoren. De eigenvektor behorend bij de grootste eigenwaarde bestaat als de operator kompakt is, en hiervoor is het voldoende dat de operator een kwadratisch integreerbare kern heeft. Deze voorwaarden zijn echter niet noodzakelijk.

2.3.3 Expansies

In Maung (1941a,b) vinden we de volgende representatie van een bivariate verdeling. We gebruiken de notatie uit de vorige paragraaf.

$$g_{ij} = d_i e_j (1 + \sum_{s=1}^{m-1} \lambda_s x_{is} y_{js}).$$

Maung schrijft dit resultaat toe aan Fisher, hoewel we iets dergelijks dus al eerder vinden bij Hirschfeld. In feite is de representatie een tamelijk direkte toepassing van de singuliere waarde dekompositie, die rond 1930 herontdekt was door Eckart en Young. Op dezelfde manier was Gebelein's representatie van kontinue bivariate dichtheden een eenvoudige toepassing van de theorie van Hilbert en Schmidt, waarin ook singuliere waarden en singuliere funkties een rol spelen.

Het klassieke voorbeeld van zo'n kontinue representatie was de identiteit van Mehler, die we hier schrijven als

$$p_\rho(x,y) = p(x)p(y)(1 + \sum_{s=1}^{\infty} \rho^s \psi_s(x)\psi_s(y)).$$

Hierbij is $p_\rho(x,y)$ de bivariate dichtheid van de normaalverdeling met korrelatieparameter ρ , $p(x)$ is de univariate dichtheid, en $\psi_s(x)$ zijn de Hermite-Chebyshev polynomen. De Fisher representatie van diskrete bivariate verdelingen en de formule van Mehler waren de belangrijkste inspiratiebronnen voor het klassieke artikel van Lancaster (1958). Daarin bespreekt Lancaster kanonische analyse van bivariate verdelingen, waarin representaties gezocht worden van de vorm van de Fisher of Mehler formule. In plaats van de machten ρ^s in de Mehler formule komen de algemene kanonische korrelaties, in plaats van de Hermite-Chebyshev polynomen komen de algemene kanonische variabelen. We gaan hier niet al te diep op de theorie in maar we noemen wat referenties, voorbeelden, en problemen.

Een eerste stap die we doen is dat we onze ekspansies steeds schrijven in de vorm

$$\mathbb{E}(\phi(\underline{x})\psi(\underline{y})) = \sum_{s=0}^{\infty} \lambda_s \, \mathbb{E}(\phi(\underline{x})g_s(\underline{x}))\mathbb{E}(\psi(\underline{y})f_s(\underline{y})).$$

Dit noemen we dus een kanonische analyse: vindt ortonormale funkties g_s op $L^2(X)$, de ruimte van alle funkties op X met eindige variantie, en vindt ortonormale funkties h_s op $L^2(Y)$, en getallen λ_s, zodat voor iedere ϕ in $L^2(X)$ en voor iedere

ψ in $L^2(Y)$ de boven genoemde ekspansie geldt. We gebruiken verwachte waarden om redenen uiteengezet in Whittle (1970, voorwoord), en bovendien omdat het een wat algemenere benadering mogelijk maakt. Zo hoeft het helemaal niet zo te zijn dat er een dichtheidsfunktie bestaat, en bovendien neemt de gebruikelijke theorie over het algemeen aan dat $\sum\lambda_s^2 < \infty$. (Cambanis en Liu (1971), Chesson (1976)). We gebruiken de konventie dat zowel g_0 en h_0 gelijk zijn aan één. Hierdoor kunnen we ook schrijven

$$E(\phi(\underline{x})\psi(\underline{y})) = E(\phi(\underline{x}))E(\psi(\underline{y}))(1 + \sum_{s=1}^{\infty} \lambda_s E(\phi(\underline{x})g_s(\underline{x}))E(\psi(\underline{y})h_s(\underline{y}))).$$

Zowel de Fisher als de Mehler ekspansie zijn van deze vorm. Merk op dat als ϕ en ψ indikator funkties zijn van Borel verzamelingen de verwachte waarden gelijk worden aan waarschijnlijkheden. We vinden dan dus een ekspansie voor de waarschijnlijkheidsmaat, voor de gebruikelijke keuze van de Borel verzamelingen vinden we een ekspansie van de bivariate verdelingsfunktie. Een interessante toepassing is de volgende (Jensen, 1971). Stel X = Y en $g_s = h_s$ voor alle s. Dan is het duidelijk dat $\lambda_s \geq 0$ voor alle s impliceert dat $E(\phi(\underline{x})\phi(\underline{y})) \geq E(\phi(\underline{x}))E(\phi(\underline{y}))$. In het bizonder geldt dan voor iedere Borel verzameling A dat $\text{prob}(\underline{x} \in A \ \& \ \underline{y} \in A) \geq \text{prob}(\underline{x} \in A)\text{prob}(\underline{y} \in A)$. Dit wordt wel positieve afhankelijkheid genoemd. Merk op dat we niet noodzakelijk aangenomen hebben dat X en Y stukken van de reele lijn zijn, ze kunnen ook deelverzamelingen zijn van $\mathbb{R}^s$ (Venter, 1966). In Hannan (1961) en Chesson (1976) vinden we zelfs voorbeelden van oneindig dimensionale ruimten, waarvoor $L^2(X)$ en $L^2(Y)$ niet noodzakelijk separabel zijn, zodat we de som in de ekspansie moeten vervangen door een integraal (kontinue kanonische analyse).

De eerste vraag die ons interesseert is: wanneer zijn de kanonische variabelen orthogonale polynomen ? Bij de meeste voorbeelden is dit het geval (Barrett & Lampard, 1955; McFadden, 1966; McGraw & Wagner, 1968; Lee, 1971), maar een stel voorbeelden zijn nog geen algemene theorie. Brown (1958) geeft de eenvoudige noodzakelijke en voldoende voorwaarde dat de $E(\underline{y}^s|\underline{x})$ en $E(\underline{x}^s|\underline{y})$ polynomen van graad ten hoogste s zijn, maar dat is natuurlijk niet erg diepgaand. Het zegt in feite alleen maar dat de oplossingen van de stationaire vergelijkingen uit de vorige paragraaf polynomen zijn. Een meer konstruktieve benadering is het werk van Eagleson (1964), Eagleson & Lancaster (1967), en Lancaster (1975).

Daar wordt een algemene theorie opgezet van bivariate verdelingen die als speciale gevallen de bivariate versies van de normaal-verdeling, en van de gamma, Poisson, hypergeometrische, en positieve en negatieve binomiaal verdeling heeft. In alle gevallen zijn de kanonische variabelen de orthogonale polynomen op de marginalen, en in al deze gevallen zijn dit ook klassieke ortogonale polynomen. Men zou kunnen zeggen dat Eagleson en Lancaster een systeem van bivariate verdelingen ontwikkeld hebben vergelijkbaar met Pearson's systeem van univariate verdelingen.

We kunnen de theorie van kanonische ekspansies ook toepassen op de overgangswaarschijnlijkheden van symmetrische stationaire Markov processen met een kontinue tijdsparameter. Voor een kontinue toestandenruimte werd dit gedaan door Wong en Thomas (1962), en door Sarmanov (1963). Voor een diskrete toestandenruimte, en in het bizonder voor geboorte en doodprocessen, door Eagleson (1969). Uitgaande van de differentiaalvergelijkingen die stationaire Markov processen definieren (afhankelijk van discipline of van land van oorsprong van de auteur de Fokker-Plank of de Chapman-Kolmogorov vergelijkingen genoemd) laat men zien dat, als we wat technische toevoegingen over het hoofd zien, de volgende representatie geldt. Als $t < u$, dan

$$E(\phi(\underline{x}_t)\psi(\underline{x}_u)) = E(\phi(\underline{x}_t))E(\psi(\underline{x}_u))\left[1 + \sum_{s=1}^{\infty} \exp(-\lambda_s(u-t))E(\phi(\underline{x}_t)g_s(\underline{x}_t))E(\psi(\underline{x}_u)g_s(\underline{x}_u)\right]$$

Hierbij is $\underline{x}_t$ dus de toestand op tijdstip t, vanwege de symmetrie en de stationariteit kunnen we aannemen dat de g_s dezelfde zijn voor alle t. Ook hier kunnen we ons de vraag stellen wanneer de kanonische variabelen ortogonale polynomen zijn. Deze vraag wordt in de aangehaalde literatuur volledig beantwoord. De kondities worden aangegeven, en in het bescheiden aantal mogelijke gevallen worden weer klassieke ortogonale polynomen als oplossingen gevonden. In een meer recent artikel van Cooper, Hoare, en Rahman (1977) worden de overgangswaarschijnlijkheden van een klasse Markov ketens geanalyseerd, de kanonische variabelen zijn hier diskrete ortogonale polynomen, door diverse grootheden naar grenswaarden te laten konvergeren vinden we ook weer processen met kontinue toestandenruimte en met klassieke ortogonale polynomen als kanonische variabelen.

Een tweede interessante vraag is de volgende. Gegeven dat een

kanonische analyse met ortogonale polynomen als kanonische variabelen mogelijk is, elke waarden kunnen dan de kanonische korrelaties aannemen ? In Sarmanov en Bratoeva (1967) werd aangetoond dat als de kanonische variabelen Hermite-Chebyshev polynomen zijn, dat dan de λ_s de momenten van de een of andere waarschijnlijkheidsverdeling op het interval (-1,+1) moeten zijn. Dit resultaat kan gebruikt worden om diverse bivariate verdelingen te konstrueren met univariaat normale marginalen. Het eenvoudigste voorbeeld krijgen we door vermengen van een binormaalverdeling met korrelatie $\rho > 0$ en een binormaalverdeling met korrelatie $-\rho$. De eerste komponent krijgt menggewicht p, de tweede menggewicht 1 - p. De kanonische variabelen zijn de Hermite-Chebyshev polynomen, de kanonische korrelaties zijn $\lambda_s = \rho^s\{p + (-1)^s(1 - p)\}$. Als $\rho > 1 - 2p$, dan geldt $\lambda_2 > \lambda_1$, en de kanonische variabele behorende bij de grootste kanonische korrelatie is de polynoom van de tweede graad. Niettemin geldt $E(\underline{x}|\underline{y}) = (1 - 2p)\rho\underline{y}$ en $E(\underline{y}|\underline{x}) = (1 - 2p)\rho\underline{x}$, dus de regressies zijn lineair. Als $p = \frac{1}{2}$ dan geldt $E(\underline{x}|\underline{y}) = E(\underline{y}|\underline{x}) = 0$, en alle λ_s met s oneven zijn gelijk aan nul, Het werk van Sarmanov en Bratoeva werd aangevuld en uitgebreid door Eagleson (1969), Griffiths (1969, 1970), Tyan en Thomas (1975).

Zoals we eerder gezegd hebben is het voor het bestaan van dit soort ekspansies niet nodig dat $\underline{x}$ en $\underline{y}$ ééndimensionale stochastische veranderlijken zijn. In Venter (1966) wordt onderzocht wat er gebeurt als $\underline{x}$ en $\underline{y}$ vektor variabelen zijn, en met name dan $\underline{x}$ en $\underline{y}$ gezamenlijk multinormaal. Min of meer dezelfde resultaten worden ook besproken in Naouri (1970), en in Pousse en Dauxois (1976). Voor deze generalizatie hebben we een multivariate generalizatie van de Hermite-Chebyshev polynomen nodig, die wordt bijvoorbeeld besproken in Appell en Kampe de Feriet (1926) of in Erdélyi (1953, deel II). De multivariate Hermite-Chebyshev polynomen zijn orthogonaal met betrekking tot een multinormaal verdeling met algemene dispersiematriks. Stel dat de gezamenlijke multinormaalverdeling van $\underline{x}$ en $\underline{y}$ kanonische korrelaties in de klassieke zin $\rho_1,\ldots,\rho_t$ heeft, dus de ρ_s zijn de singuliere waarden van $\Sigma_{xx}^{-\frac{1}{2}}\Sigma_{xy}\Sigma_{yy}^{-\frac{1}{2}}$. Dan zijn de kanonische korrelaties voor ons ekspansie probleem van de vorm $\rho_1^{k_1}\cdots\rho_t^{k_t}$, waarbij de eksponenten k_s alle maal van 1 naar ∞ lopen. De grootste kanonische korrelatie is dus gewoon weer ρ_1, en deze wordt aangenomen als alle transformaties lineair zijn. Dit laatste resultaat is volgens Sarmanov en Zacharov

(1960) het eerst ontdekt door Kolmogorov: het laat dus zien dat in het geval van de multinormaalverdeling de maksimale korrelatie tussen twee groepen gelijk is aan de eerste kanonische korrelatie, en dit geldt niet alleen als we lineaire kombinaties bekijken, maar ook als we iedere variabele apart niet-lineair mogen transformeren alvorens lineair te kombineren, en zelfs als we zowel niet-lineair mogen transformeren als kombineren. Een bizonder geval is natuurlijk multipele korrelatie, omdat er daar maar één klassieke kanonische korrelatie is, de multipele korrelatie koefficient. De kanonische korrelaties in de ekspansie zijn dus de machten van de multipele korrelatiekoefficient, en de grootste multipele korrelatie niet-lineair is de gewone lineaire multipele korrelatie.

2.3.4. HOMALS als eerste stap

Uit de voorafgaande bladzijden is het duidelijk dat in een aantal gevallen de eerste dimensie (het eerste stel transformaties), behorende bij de grootste eigenwaarde, van bizondere betekenis is. Bij de multinormaalverdeling bijvoorbeeld geeft de eerste set transformaties de oorspronkelijke variabelen terug. In plaats van verdere dimensies te berekenen is het daarom dikwijls nuttig te volstaan met die ene dimensie, en de gevonden transformaties te gebruiken op de variabelen te kwantificeren of te metrizeren. We kunnen met deze gekwantificeerde variabelen dan gewoon verder rekenen alsof we ze gemeten hebben. We kunnen bijvoorbeeld de standaard lineaire technieken als multipele regressie of principale komponenten analyse nu zonder verdere komplikaties toe passen. Uit de voorbeelden in dit hoofdstuk, en met name uit de Monte Carlo studie in hoofdstuk 7, blijkt dat een dergelijke aanpak aantrekkelijke resultaten op kan leveren.

2.4. Toepassingen van één-dimensionale HOMALS.

2.4.1. Inleiding.

Het wordt hoog tijd een paar toepassingen van één-dimensionale HOMALS te bekijken. We hebben daartoe vier typen onderzoek gekozen die vrij representatief zijn voor hoe niet-experimentele sociale wetenschappers veelal te werk gaan. Een aantal van de voorbeelden komt in latere hoofdstukken in meer dimensies opnieuw aan bod. We hebben geenszins de pretentie de betrokken toepassingen metodologies uitputtend te analyseren. De bedoeling is: mogelijkheden aanduiden, kijken wat de gebruikelijke aannamen waard zijn, visualisering van resultaten.

De eerste toepassing betreft een item-analyse van een multiple choice tentamen. Kenmerkend is, dat elke variabele één favoriete kategorie heeft. Dan bekijken we een aantal traditionele manieren van attitude-schaling, waarbij de kategorieën meestal op z'n minst als geordend opgevat worden. In hoeverre dit klopt gaan we na bij een onderzoek naar de houding tegenover abortus en een aantal aanverwante kwesties betreffende leven en dood. Vervolgens komt een omvangrijk survey onderzoek aan bod, naar de achtergronden van schoolcarrière en beroepskeuze van nederlandse kinderen (van Jaar tot Jaar). Hier is kenmerkend, dat we een groot aantal variabelen van verschillend type hebben. Het hoofdstuk wordt besloten met een voorbeeld van een veldstudie, ondernomen om enig licht te werpen op wat mensen beweegt, bloed te geven (Bloedbank). In dit soort toepassingen is het vaak de bedoeling, een nieuwe typering van individuen te vinden.

Omdat het Abortus en van Jaar tot Jaar onderzoek beide nogal omvangrijk zijn, en ook in latere hoofdstukken terug komen, is een uitgebreide beschrijving van individuen, variabelen en kategorieën in de Appendix opgenomen.

2.4.2. Item-analyse: multiple-choice tentamen.

Het konstrueren van een multiple-choice tentamen is geen eenvoudige zaak. Naast het feit dat het moeilijk kan zijn om één-één-duidig interpreteerbare vragen te stellen en om inzicht te meten i.p.v. pure feitenkennis, blijkt het vaak niet gemakkelijk

4 kategorieën							2 kategorieën (goed/fout)				
nr	marginale frekwenties					disk.m.	nr	marginale frekwenties			disk.m.
	M	1	2	3	4			M	1-goed	2-fout	
+ 1	0	16	24	10	140=	.103	+ 1	0	140	50	.105
+ 2	0	11	147=	25	7	.285	+ 2	0	147	43	.264
3	0	17	34	32	107=	.067	3	0	107	83	.088
+ 4	0	22	9	17	142=	.068	+ 4	0	142	48	.041
5	0	102=	0	81	7	.001	5	0	102	88	.004
6	0	1	29	68	92=	.284	6	0	92	98	.155
7	0	4	44	52	90=	.264	7	0	90	100	.182
8	0	49	11	100=	30	.182	8	0	100	90	.214
9	0	1	153=	31	5	.062	9	0	153	37	.102
+ 10	0	109=	26	2	53	.136	+ 10	0	109	81	.113
11	0	26	121=	39	4	.260	11	0	121	69	.195
+ 12	0	11	36	5	138=	.201	+ 12	0	138	52	.183
13	0	90	4	92=	4	.081	13	0	92	98	.018
14	0	19	9	36	126=	.026	14	0	126	64	.021
15	0	40	14	28	108=	.134	15	0	108	82	.128
+ 16	0	5	5	144=	36	.302	+ 16	0	144	46	.211
17	0	106=	48	27	9	.074	17	0	106	84	.110
18	0	5	17	15	153=	.262	18	0	153	37	.144
+ 19	0	66	1	102=	21	.058	+ 19	0	102	88	.000
20	1	3	35	138=	13	.252	20	0	138	52	.228
+ 21	0	12	47	103=	28	.093	+ 21	0	103	87	.102
22	0	22	38	10	120=	.102	22	0	120	70	.105
23	0	8	11	148=	23	.150	23	0	148	42	.142
24	0	44	126=	6	14	.154	24	0	126	64	.116
25	0	154=	11	10	15	.380	25	0	154	36	.213
+ 26	1	103=	29	35	22	.058	+ 26	0	103	87	.040
27	0	13	135=	37	5	.131	27	0	135	55	.102
28	0	117=	28	39	6	.163	28	0	117	73	.198
29	0	5	167=	9	9	.377	29	0	167	23	.290
+ 30	0	27	149=	11	3	.309	+ 30	0	149	41	.293
e.w.						.1673	e.w.				.1369

tabel 2.5. Multiple-choice tentamen: marginale frekwenties, diskriminatiematen en eigenwaarde.

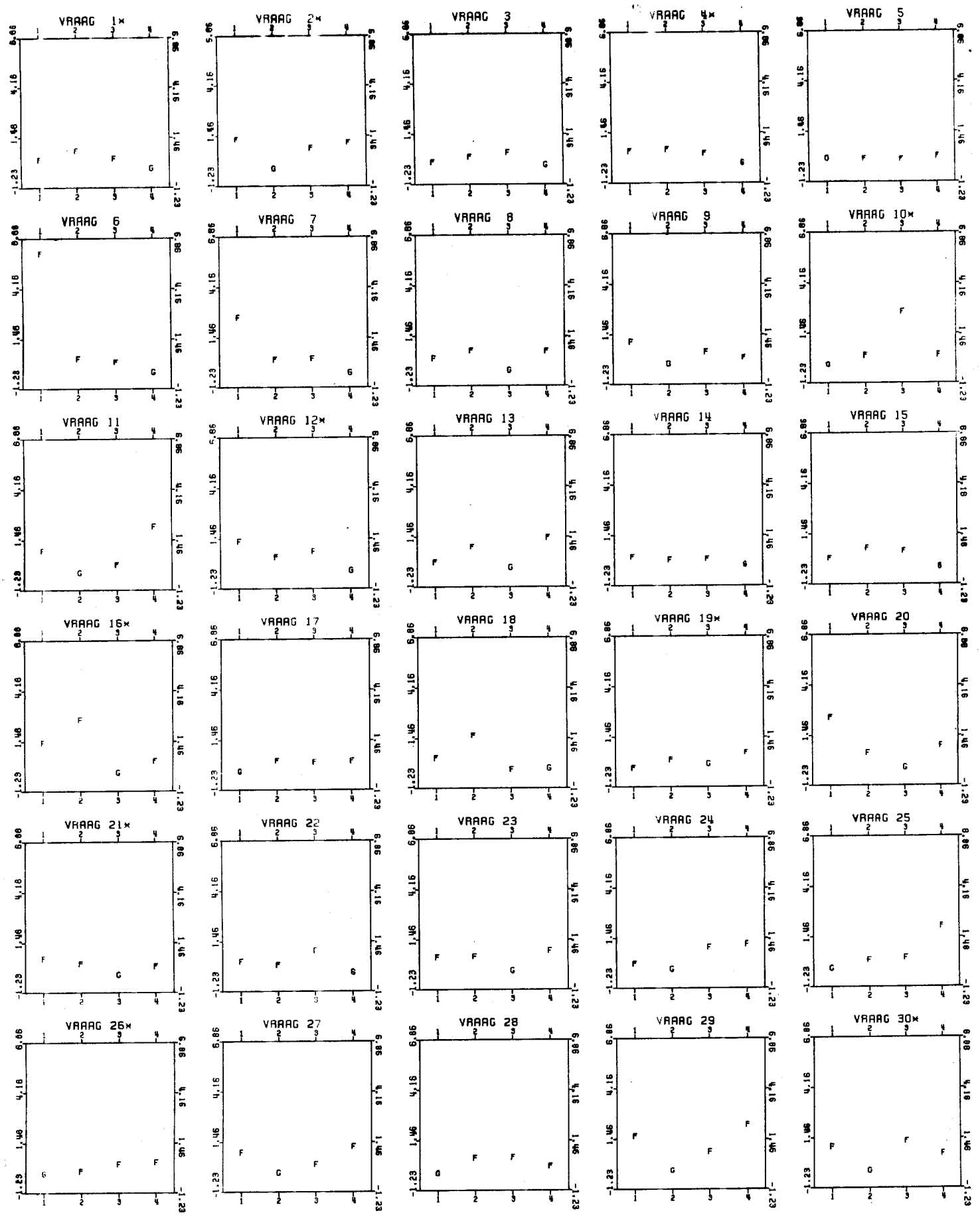
foute alternatieven te verzinnen. Dat wil zeggen, alternatieven die inspelen op een verwachte denkfout van de tentaminandus, terwijl ze enerzijds duidelijk aantoonbaar fout zijn en anderzijds niet op het eerste gezicht al onzinnig lijken. Bij bijv. statistiek tentamens is dit nog wel te doen: als men verwacht dat mensen fouten maken door met de variantie te rekenen i.p.v. met de standaarddeviatie, kan men de opgave doorrekenen met de variantie en zo een fout alternatief konstrueren. Hetzelfde geldt bijv. voor tweezijdig toetsen i.p.v. éénzijdig. De fout is zo direkt

interpreteerbaar en gelijksoortige fouten van verschillende vragen kunnen met elkaar vergeleken worden.

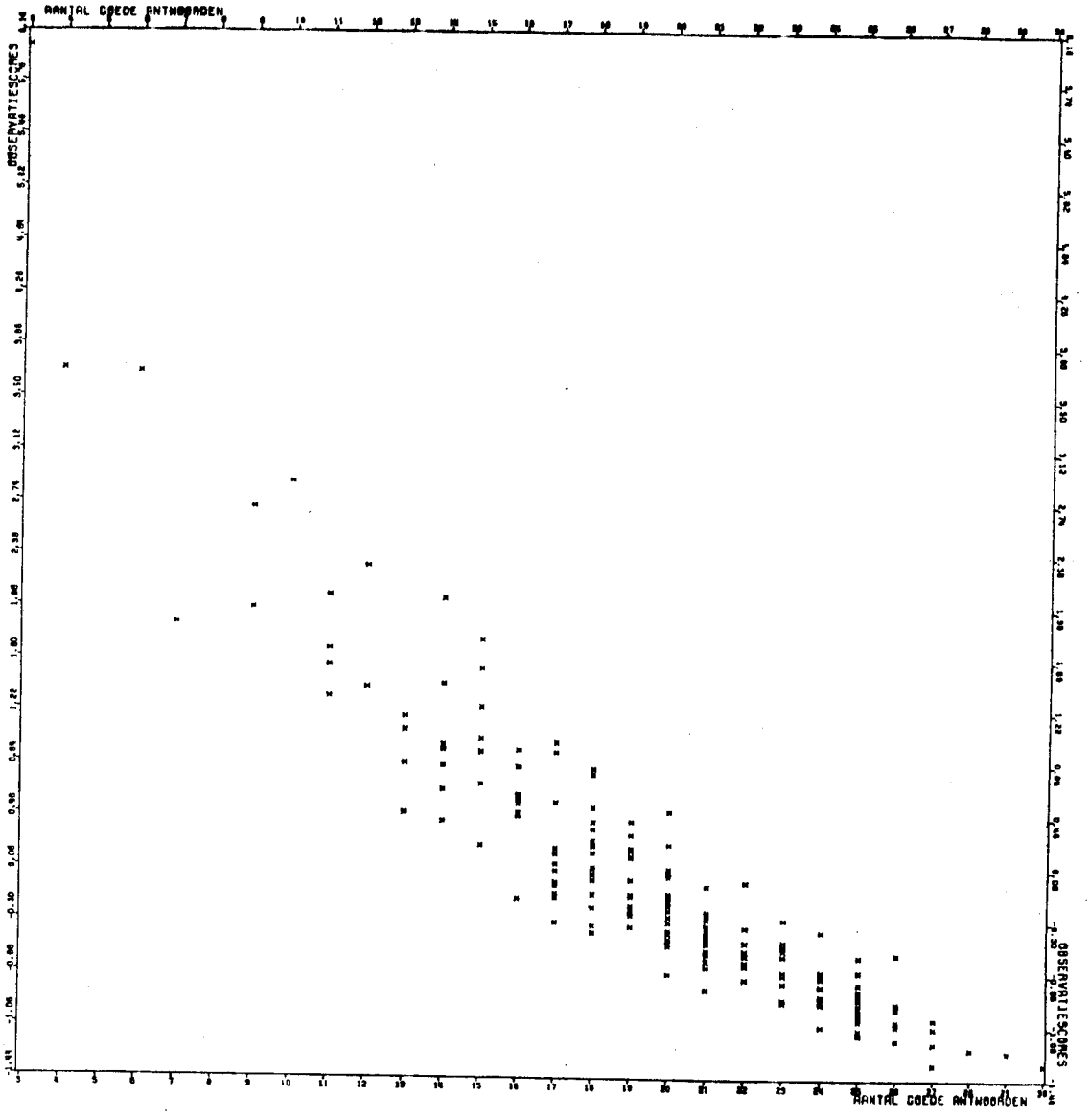
Bij het hier gebruikte tentamen is een en ander minder eenvoudig. Het gaat om het tentamen 'Inleiding in de Psychologie', dat in januari 1978 door 190 leidse eerstejaars studenten gemaakt is. Bij ieder van de 30 vragen kon de student kiezen uit vier alternatieven, waarvan er zoals gebruikelijk slechts één goed was. In tabel 2.5 is een overzicht gegeven van de vragen. Het goede antwoord is steeds aangeduid met een =. Rechts in de tabel staan de vragen en de HOMALS diskriminatiematen weergegeven voor het geval dat de variabelen gedichotomiseerd worden naar goed/fout. Een + voor het nummer van de vraag geeft aan, dat deze deel uitmaakt van de zgn. 'deeltoets': een stel vragen die in een eerder tentamen ook al gebruikt zijn. Men gaat er van uit dat dit vragen zijn met een hoge item-totaal korrelatie en een p-waarde (kans op goed) van ongeveer .70. Door het herhaaldelijk opnemen van de zelfde deeltoets kan men de tentamenresultaten over verschillende jaren met elkaar vergelijken en steeds dezelfde norm hanteren.

De gebruikelijke procedure bij het bekijken van een tentamen is, dat men alle foute alternatieven van een vraag op een hoop gooit en dan de item-totaal korrelaties en de p-waarden bepaalt. Vragen met een hoge p-waarde zijn gemakkelijk, met een lage moeilijk. Een lage item-totaal korrelatie geeft aan, dat het om een 'slechte' vraag gaat: fout beantwoord door mensen met veel goede antwoorden en/of goed beantwoord door mensen met veel foute antwoorden.

HOMALS biedt de mogelijkheid om de foute alternatieven apart te bekijken. In figuur 2.5 zijn de door HOMALS berekende optimale kategorie-transformaties uitgezet tegen het nummer van het alternatief. In het algemeen hebben de goede alternatieven (G) de laagste waarden gekregen. Deze alternatieven liggen dan ook het dichtst bij elkaar op de door HOMALS gekonstrueerde schaal: 'goed' heeft een duidelijke betekenis, in de zin dat er een gebied wordt afgebakend waarin bijna alle goede alternatieven liggen en dus ook de mensen met veel goede antwoorden. De schaal loopt dus als het ware van een goed resultaat naar een slecht. Als we dan ook de gevonden individu-skores vergelijken met het aantal goede



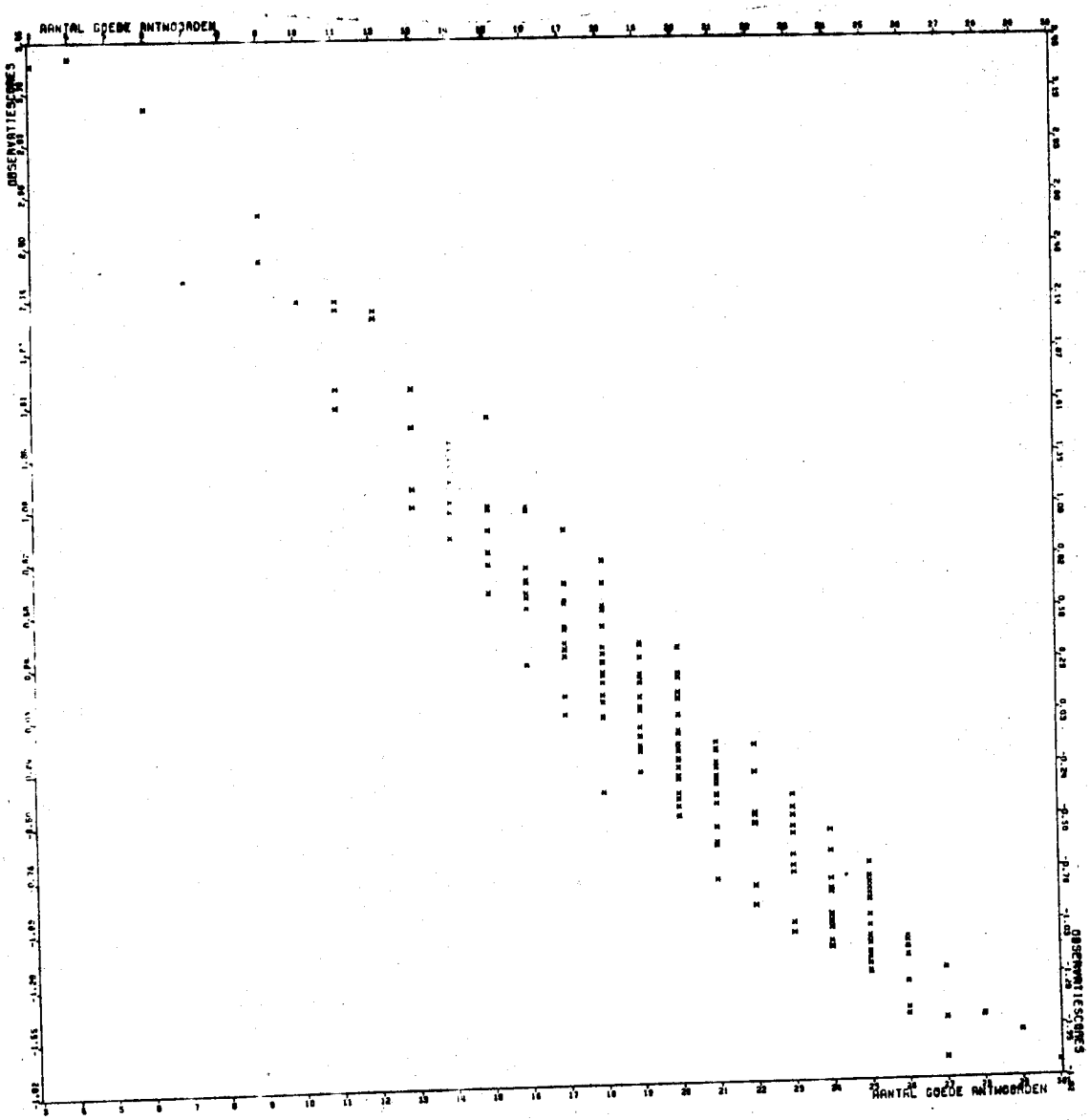
figuur 2.5. Kategorie-transformaties voor MC-tentamen (vragen met een * maken deel uit van de deeltoets; goede alternatieven: G).



figuur 2.6. Individu-skores na HOMALS (4 kat. MC) uitgezet tegen het aantal goede antwoorden.

antwoorden, blijken deze twee sterk samen te hangen (zie figuur 2.6). Optimaal wegen heeft hier dus geen dramaties effekt.

Het is echter niet zo, dat alle vragen 'goed' op de goede plek (nl. beneden de foute alternatieven) hebben. Deze 'rare' vragen hebben ook een lage diskriminatiemaat (vgl. vraag 5 en vraag 19). De diskriminatiematen zijn gelijk aan het kwadraat van de item-totaal korrelaties van de geinduceerde skores en hiervoor geldt dus hetzelfde als wat eerder over de oorspronkelijke skores werd gezegd: een lage diskriminatiemaat betekent dat een vraag slecht bij de andere past en blijkbaar iets heel anders meet dan de op



figuur 2.7. Individu-skores na HOMALS (2 kat. MC) uitgezet tegen het aantal goede antwoorden.

de overige items berustende goed-fout schaal.

Ook bij HOMALS op twee kategorieën (goed versus fout) blijken vraag 5 en vraag 19 slecht te diskrimineren. In figuur 2.7 zijn de individu-skores van deze HOMALS analyse weer uitgezet tegen het aantal goede antwoorden. De korrelatie van de individu-skores met het aantal goede antwoorden is hier overigens nog hoger dan in het 4-kategorieën geval. De homogeniteit is nu echter lager (.137 versus .167). Dit komt, omdat de 2-kategorieën oplossing minder vrijheidsgraden heeft: personen met een fout antwoord op een bepaalde vraag moeten op elkaar liggen, terwijl ze zich bij

vier kategorieën over drie fouten konden verspreiden.

Zoals gezegd kunnen we op grond van de 4-kategorie HOMALS oplossing de fouten apart bekijken. Als een 'foute' kategorie dicht bij de goede ligt, is hij waarschijnlijk vaak gekozen door mensen die veel andere vragen goed hebben. Dit geldt des te sterker, als het 'foute' alternatief nog lager ligt dan het goede. Hoe verder een F ligt van G, hoe fouter dat alternatief als het ware is. Als alle foute alternatieven op gelijke hoogte liggen (zie bijv. vraag 17), zijn alle fouten ook even fout.

Doordat we nu zo de betekenis of het belang van verschillende fouten kunnen achterhalen, is het mogelijk om dieper in te gaan op de relatie tussen de individu-skores en het aantal goede antwoorden. Om te slagen moet men 20 vragen goed hebben beantwoord; waar komt dan de spreiding van de skores van de individuen met 20 goed vandaan? En sterker nog: hoe komt het dat de persoon met 27 goed een lagere individu-skore heeft dan degenen met 28, 29 en zelfs 30 goed en dus volgens HOMALS de 'beste' zou moeten zijn?

Het antwoord hierop is heel eenvoudig: de betrokken student had drie fouten, te weten op vraag vijf 3 i.p.v. 1, op vraag dertien 1 i.p.v. 3 en op vraag negentien 1 i.p.v. 3. Aan de plot van de kategorie-transformaties (figuur 2.5) kunnen we zien, dat de foute kategorie van vijf en negentien waar hij in geskoord heeft een lagere getransformeerde skore hebben dan de goede kategorie. Hij of zij heeft dus fouten gemaakt in vragen die slecht diskrimineren (vgl. ook de diskriminatiematen in tabel 2.5), en bovendien die foute alternatieven gekozen die door veel 'goede' mensen ook gekozen zijn; de goede kategorieën van de betrokken items zijn juist relatief vaak door 'slechte' mensen gekozen.

Dit alles geldt ook, hoewel in mindere mate, voor de resultaten van de 2-kategorieën oplossing. Hier wordt geen onderscheid gemaakt tussen 'goede' en 'slechte' fouten. Blijft, dat twee van de drie fouten van de betrokken persoon gemaakt zijn in items die slecht diskrimineren. Bij vraag vijf heeft zelfs nu nog het foute alternatief een kwantifikatie die dichter bij 'goed' van de andere items ligt dan bij 'goed' van hetzelfde item.

Tenslotte een kort overzicht van de aard van de kategorieën bij

Vraag

2) parasympatisch deel autonoom zenuwstelsel :
1. dominant als opgewonden 2. dominant als ontspannen

4) behavioristische ontwikkelingstheorie
1. Piaget 2. Erikson 3. Kohlberg 4. geen van drieën

6) impulsroute naar occipitaal kwab vanaf linker (gestippeld) en rechter (ononderbroken lijn) helft retina
1.  2.  3.  4. 

10) primaire reinforcer in konditioneringsexperiment
1. voedsel 2. lampje 3. toon + lampje 4. toon

13) volgorde moeilijkheid konsepten : objekts-, ruimtelijke en vormbegrippen
1. OGR 2. ROG 3. ORG 4. RGO

17) identificeren drie situaties met drie arousal grafieken (A, B en C)
1. A1 B2 C3 3. A3 B1 C2 3. A2 B3 C1 4. A3 B2 C1

20) primary process thinking hoort bij ... , secondary process thinking bij ...
1. ego id 2. ego superego 3. id ego 4. id superego

24) wel (+) of niet (-) aanwezig zijn van drie kenmerken (A , B en C) in omschrijving psychopaat van Szasz
1. $A^{+}B^{-}C^{+}$ 2. $A^{-}B^{-}C^{+}$ 3. $A^{+}B^{+}C^{-}$ 4. $A^{-}B^{+}C^{-}$

28) twee groepen (A;B) : welke ervaart in een bepaalde situatie dissonantie volgens Festinger;welke pleegt dispositionele attributies volgens Bem. Welke groepen ..
1. A;A 2. A;B 3. B;A 4. B;B

tabel 2.6. Beknopte omschrijving van een aantal antwoordkategorieën.

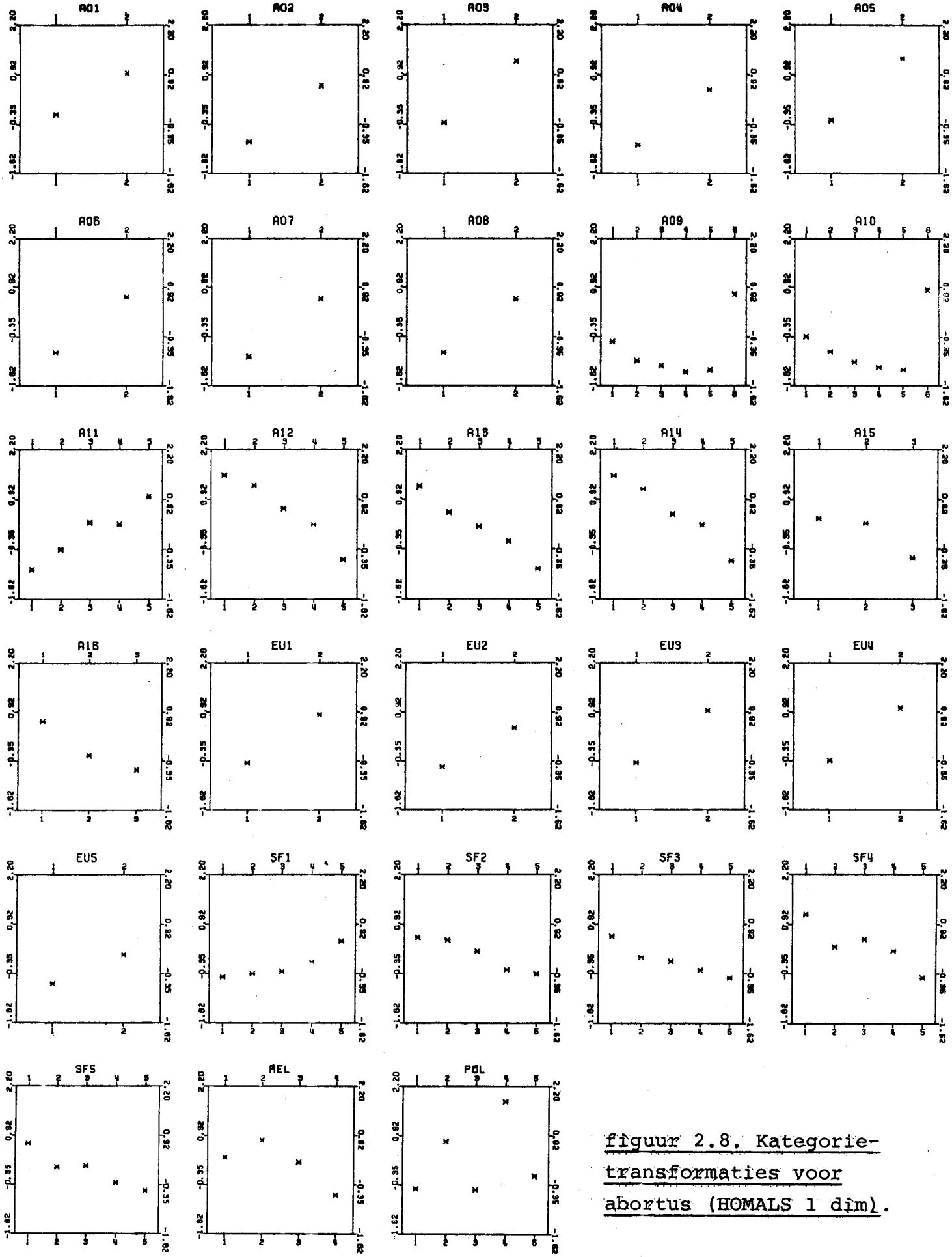
enkele vragen. In sommige gevallen zijn de kategorie-kwantificaties goed te interpreteren, omdat bijv. de meest extreme waarden binnen één vraag betrekking hebben op het goede antwoord en het inhoudelijk tegengestelde alternatief, terwijl de er tussenin liggende kategorieën een gedeeltelijk goed alternatief aangeven, of een alternatief zonder één-één-duidige relatie met het juiste (zie tabel 2.6).

2.4.3. Attitude schaling; abortus.

Naar aanleiding van de gegevens van het abortus onderzoek (zie Appendix B.2) zullen we hier de oplossingen uit één-dimensionale HOMALS-runs vergelijken met de resultaten van enkele attitude schalings methoden. Hierbij moet men er rekening mee houden dat, bij iedere schaalmethode, men uitgaande van hetzelfde model de schaalmethode kan zien als kriterium en als techniek. Bij de methode als kriterium denkt men van te voren dat de variabelen/items een bepaalde onderliggende schaal vertegenwoordigen. Men toetst dan als het ware of de data aan deze veronderstellingen beantwoorden. Bij de methode als techniek probeert men uit de items een ééndimensionale schaal te konstrueren. Hierbij zal men ook items weglaten, die niet in deze schaal passen.

Verschillende schaalmethoden gaan uit van verschillende soorten items. In figuur 2.8 zijn de getransformeerde kategorieskores van de één-dimensionale HOMALS-oplossing van de abortus variabelen weergegeven die hier en in volgende hoofdstukken geanalyseerd worden. Men ziet, dat er 9 zgn 'Likert' items (A10 t/m A14 en SF1 t/m SF5) bij zijn : items met 5 kategorieën (zeer mee oneens .... zeer mee eens). Verder zijn er 13 binaire 'Guttman' items (A01 t/m A08 en EU1 t/m EU5) : items met twee kategorieën zoals eens-oneens , goed-fout , geoorloofd-niet geoorloofd , of meer algemeen wel-niet in het bezit van een bepaalde eigenschap (overigens kan men ook uitgaan van items met meer kategorieën, waarbij men dan één of meer kategorieën definieert als het positief alternatief (bv alle kategorieën die een bepaalde gradatie van instemming weergeven) en de overige kategorieën als het negatief alternatief).
Verder kan men POL (politieke partij) opvatten als een item dat in relatie tot abortus een niet-monotone ééndimensionale schaal, een zogenaamde Coombs schaal zal weergeven (zie hoofdstuk 2.2.7.). Links (PvdA ed : kat 1) en rechts (VVD ed : kat 3) nemen nl volgens de HOMALS oplossing een positievere houding aan dan het midden (CDA : kat 2). Hier wordt verder niet op de Coombs schaal ingegaan.

De HOMALS oplossing voor verschillende groepen gelijksoortige variabelen (per groep zijn afzonderlijke analyses gedaan) kan men vergelijken met de resultaten van de standaard attitude schalingsmethoden. Overigens is het van belang dat men in een HOMALS analyse zowel 'Guttman' als 'Likert' als 'Coombs' als wat voor soort andere items (bv allerlei zeer nominale achtergrondsvariabelen, zoals beroep) tegelijk op kan nemen. HOMALS kan verder ook gezien worden als een manier om schalen te konstrueren (zie boven) en bovendien biedt het



figuur 2.8. Kategorie-transformaties voor abortus (HOMALS 1 dim).

de gelegenheid om relaties tussen variabelen en tussen kategorieën aan het daglicht te brengen, die bij genoemde schaalmethoden niet te achterhalen zijn.
Hier wordt eerst uiteengezet hoe men bij de standaard attitude schalingsmethoden die hier aan de orde komen te werk gaat. Daarna worden de resultaten van deze methoden vergeleken met de HOMALS data voor de 'abortus'-items. Al eerder in dit hoofdstuk is ingegaan op de Likert-schaal (2.1.1) en op de Guttman-schaal (2.2.3). Deze schaal methoden hebben gemeen dat ze op zoek zijn naar één onderliggende dimensie : items die naar een andere dimensie lijken te verwijzen worden uit de schaal verwijderd. De Mokken-schaal , een uitbreiding van de Guttman-schaal, probeert , als er meerdere dimensies zijn, verschillende schalen te konstrueren die ieder één-dimensionaal zijn. Voor de theoretische formulering van deze schaal (2.2.2) en de beide andere wordt verwezen naar genoemde hoofdstukken , hier wordt meer op de praktische kant en op de interpretatie van de resultaten ingegaan.

Likert schaal : methode van gesommeerde skores

Paradoxaal genoeg is het eerste wat we over deze schaal moeten opmerken, dat het eigenlijk geen schaal is. De achtergrond moet gezocht worden in de test-theorie : kriteria als *interne konsistentie* en maten als *item-totaal korrelatie* komen daar vandaan. Kenmerkend voor de Likert schaal is ook een bepaald type *rating scale* , maar deze schaal zou men ook volgens de methode van Guttman kunnen analyseren (door de kategorieën op te delen in positieve en negatieve alternatieven). Tenslotte wordt de Likert schaal ook wel gekarakteriseerd als *poor man's factor analysis* : een poging om de intrinsieke struktuur van de data weer te geven toen faktoranalyse door het ontbreken van computers nog geen haalbare techniek was.
Waarom hier dan toch aandacht besteed aan de Likert-schaal ? In de praktijk wordt er nog steeds mee gewerkt. Men spreekt over Likert-items en men konstrueert Likert-schalen. De moeite waard dus om de resultaten van de Likert-schaal te vergelijken met HOMALS.

Als men een Likert schaal wil konstrueren begint men met het verzamelen van een groot aantal items, dwz uitspraken die een bepaalde mening over een onderwerp en/of een bepaalde gedragsintentie mbt dat onderwerp weergeven. De onderzoeker bepaalt dan voor ieder item eerst of het een positieve dan wel negatieve attitude/houding over het onderwerp aangeeft. Als de betekenis van een item voor een onderwerp niet duidelijk te interpreteren is of op een neutrale attitude lijkt te wijzen, wordt het item uit de set verwijderd.

De gehandhaafde items worden dan voorgelegd aan een grote steekproef waarbij ieder op een item kan reageren met één van de volgende kategorieën :

zeer mee eens - mee eens - noch mee eens, noch mee oneens - mee oneens - zeer mee oneens

Antwoorden op positieve items worden 1 t/m 5 gekodeerd, op negatieve items 5 t/m 1 . De voorlopige attitude skore van een persoon is dan de som van zijn skores op alle items.

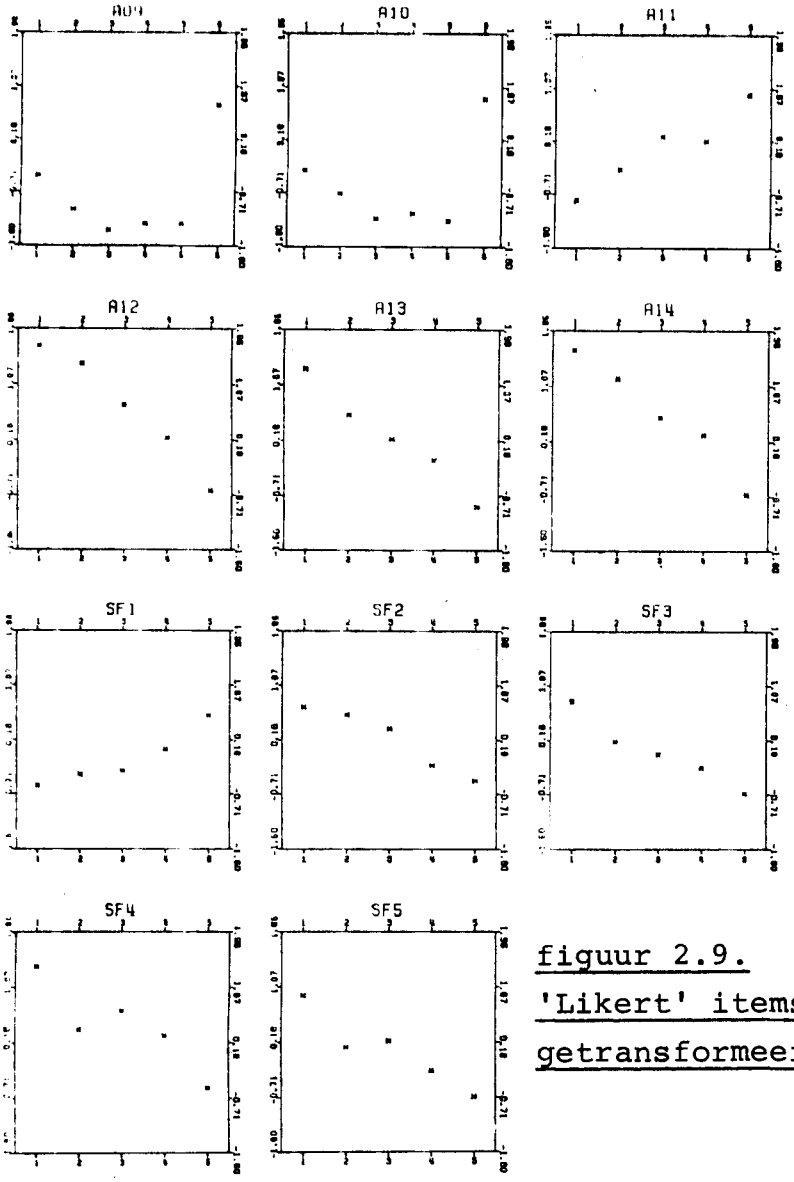
Vervolgens wordt gekeken welke items een monotone één-dimensionale schaal vormen. Aanvankelijk bepaalde men dit aan de hand van het *criterion of internal consistency* : men gaat voor ieder positief geformuleerd item na, of de 10% mensen met de hoogste attitudeskores op dit item ook duidelijk hoger geskoord hebben dan de 10% mensen met de laagste attitudeskores. Voor negatief geformuleerde items geldt het tegenovergestelde.

Toen de uitvinding van moderne rekenapparatuur niet alleen faktor-analyse, maar ook itemanalyse door het berekenen van item-totaal korrelaties mogelijk maakte, ging men ipv het criterion of consistency de item-totaal korrelaties gebruiken om afwijkende (dwz niet passend in de schaal) items op te sporen. Volgens Likert leveren het criterion of consistency en item analyse trouwens vergelijkbare resultaten op (Likert , 1974 : p.238).

De uiteindelijke Likert-schaal bestaat dan uit de $\pm$ 20 meest diskriminerende items ofwel de 20 items met de hoogste item-totaal korrelatie. Deze schaal kan nu aan andere personen voorgelegd worden. De toekenning van de skores aan de items gebeurt op dezelfde manier als boven beschreven. De som van de skores op de items bepaalt de 'definitieve' (itt voorlopige) skore op de attitudeschaal.

In de praktijk spreekt men ook wel van een Likert item en een Likert schaal als men bovenstaande procedure niet helemaal volgt, maar uitgaat van een beperkt aantal 'Likert-items' waarvan men bij voorbaat uitgaat dat ze een schaal vormen. Men kan dan op grond van de resultaten nagaan of de items een schaal vormen en tegelijkertijd de attitude skores berekenen. Ook bij het 'abortus' voorbeeld zijn er slechts een beperkt aantal 'Likert' items opgenomen, die niet van te voren bij een andere groep op schaalbaarheid zijn gekontroleerd. Omdat de interpretatie van item-totaal korrelaties ed hetzelfde blijft, lijkt dit hier bij vergelijking van de schaal met HOMALS niet zo'n bezwaar.

De relatie tussen de schaal en de afzonderlijke items impliceert, dat men - naarmate men een hogere attitude skore heeft - het meer eens is met positief geformuleerde items en minder met negatieve. De skore op



figuur 2.9.
'Likert' items
getransformeerd

één item zegt daarbij niets over de skore op een ander item.
De individuen hebben zo een positie op een één-dimensionale attitude-schaal. Door de schalingsprocedure zijn items die naar een andere dimensie lijken te verwijzen uit de set verwijderd.

In figuur 2.9 vinden we de optimale kategorie-transformaties na HOMALS voor de Likert items. Op de volgende pagina vinden we in tabel 2.7.a de korrelaties en item-totaal korrelaties van de oorspronkelijke skores. Twee dingen moeten hierbij opgemerkt worden : item A11 en SF1 zijn vergeleken met de andere items negatief geformuleerd. Deze items zijn omgepoold en de korrelaties met andere items zijn voor deze omgepoolde skores berekend. Aan de plot van de optimaal getransformeerde kategorieskores ziet men, dat HOMALS deze items zelf ahw ompoolt : bij A11 en SF1 wordt kategorie 1 tot een lage waarde getransformeert en krijgen de volgende kategorieën een steeds hogere waarde (muv een kleine afwijking in de volgorde bij A11 : kat 3 en 4). Bij A12 t/m A14 en SF2 t/m SF5 verlopen deze transformaties precies omgekeerd (met enkele kleine verschuivingen bij SF4 en SF5).
Ten tweede zijn in de HOMALS oplossing en in de Likert-schaal twee items opgenomen, die geen Likert items zijn , nl A09 en A10. Deze items beschrijven een bepaalde situatie en de antwoordkategorieën geven aan tot wanneer men vindt dat abortus in die situatie geoorloofd is : tot 3 (kat 1) , 4 (2) , 5 (3) , 6 (4) maanden , na 6 maanden (5) en 'niet geoorloofd' (6) (zie Appendix). Deze items bevatten ieder eigenlijk 2 vragen : mag het of niet en als het mag, tot wanneer dan. Men zou deze items ook kunnen opvatten als reakties op de uitspraak

tabel 2.7.
a. Oorspronkelijke skores (met wijzigingen)

	A09	A10	A11	A12	A13	A14	SF1	SF2	SF3	SF4	SF5	I-T
A09		.55	.39	.35	.41	.37	.27	.24	.25	.26	.20	.61
A10			.46	.37	.43	.41	.23	.19	.22	.27	.20	.64
A11				.50	.57	.49	.24	.14	.18	.26	.21	.65
A12					.71	.68	.26	.27	.28	.36	.30	.71
A13						.70	.28	.27	.28	.39	.34	.77
A14							.23	.26	.25	.38	.38	.72
SF1								.18	.30	.25	.30	.52
SF2									.37	.39	.39	.52
SF3										.37	.54	.58
SF4											.50	.63
SF5												.63
											hom	.4109

b. Skores na HOMALS transformaties

	A09	A10	A11	A12	A13	A14	SF1	SF2	SF3	SF4	SF5	I-T
A09		.65	.52	.49	.56	.51	.32	.31	.27	.35	.29	.73
A10			.58	.52	.59	.57	.27	.23	.25	.36	.29	.75
A11				.55	.58	.51	.26	.15	.19	.28	.24	.69
A12					.73	.68	.29	.28	.28	.40	.32	.78
A13						.71	.30	.29	.30	.41	.36	.83
A14							.27	.27	.25	.40	.40	.79
SF1								.15	.30	.28	.29	.47
SF2									.36	.43	.40	.49
SF3										.38	.54	.52
SF4											.51	.63
SF5												.60
											hom	.4541

'in deze situatie is abortus niet geoorloofd' waarbij kat. 6 als geen mee dan opgevat zou kunnen worden en 1 t/m 5 in toenemende mate oneens zijn met deze uitspraak aangeven.
Deze interpretatie is ook in overeenstemming met de HOMALS transformaties. Kategorie 6 van A09 en A10 heeft, net als de abortusverwerpende kategorieen van de andere variabelen een hoge optimale skore, terwijl 1 t/m 5 (behoudens kleine afwijkingen) min of meer aflopen wat betreft optimale skores.

Deze twee niet-Likert items zijn dus wel in de schaal opgenomen. Hiertoe zijn ze echter eerst geherkodeerd : kat. 6 werd 1 , 1 werd 2 , enz.

Na deze ompolingen en herkoderen kan men tabel 2.7.a vergelijken met tabel 2.7.b . Hier staan de korrelaties en item totaal korrelaties berekend op grond van de getransformeerde skores als resultaat van een HOMALS analyse. De diskriminatiematen uit HOMALS zijn gelijk aan het kwadraat van de item-totaal korrelaties. De som van de diskriminatiematen, gedeeld door het aantal variabelen, is gelijk aan de eigenwaarde van de oplossing, de maat voor de homogeniteit. Uitgaande van de item-totaal korrelaties van de 'oorspronkelijke' skores, is voor deze oplossing ook de homogeniteit berekend. Deze staat in de tabel vermeld.

Als men de Likert-schaal met de HOMALS oplossing vergelijkt, kan men het volgende opmerken:

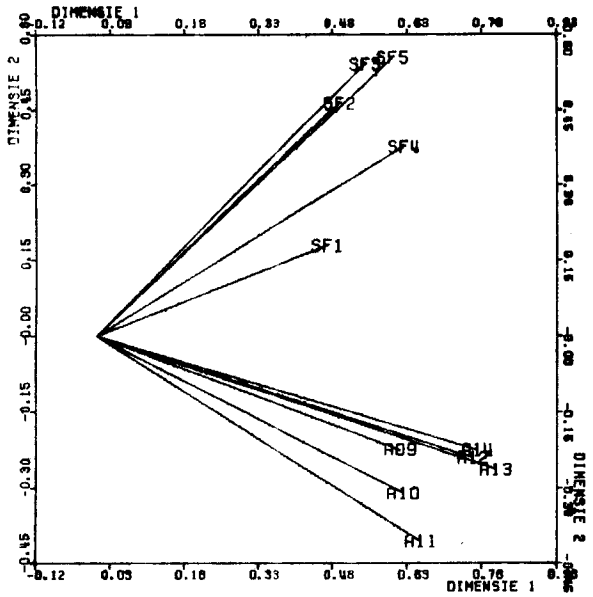
- er is duidelijk sprake van een belangrijke eerste dimensie: de abortus items laden hier iets zwaarder op dan de items over sexuele vrijheid. Dit effekt vindt men versterkt in de HOMALS oplossing, waarvan de homogeniteit iets hoger is. Door HOMALS worden de abortussitems

ahw zwaarder gewogen dan de SF items, hetgeen men ook ziet aan de plotjes van de optimale kategorie transformaties : de abortus items hebben iha een grotere spreiding op de vertikale as dan de SF items, hetgeen wijst op het verschil in belangrijkheid.
Als de items uitsluitend één onderliggende schaal zouden vertegenwoordigen, dwz als er slechts één belangrijke dimensie zou zijn, zou de item-totaal korrelatie voor alle items na optimale transformatie hoger moeten worden. Het feit dat dit niet zo is (voor de SF items worden ze lager) wijst erop, dat er nog een tweede dimensie is, waarop de SF items beter uit de verf moeten komen.

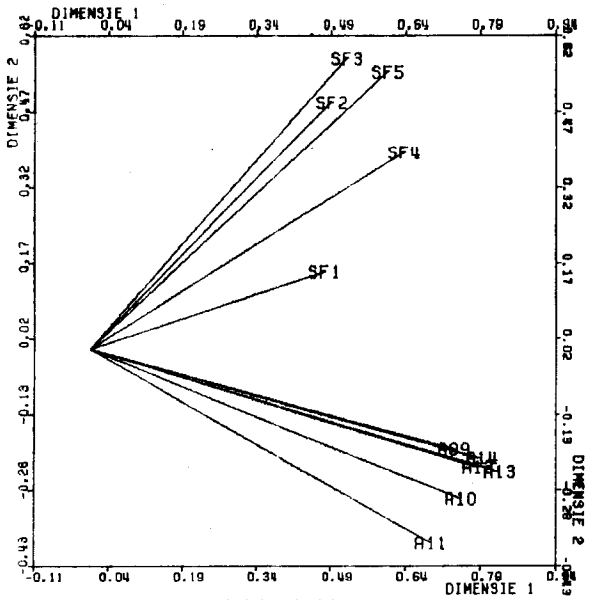
- zowel op grond van de item-totaal korrelaties van de oorspronkelijke als van de getransformeerde skores, kan men konkluderen dat sommige items meer diskrimineren dan andere (dwz een hogere item-totaal korrelatie hebben) . In een Likert schaal tellen echter alle items even zwaar mee , itt bij HOMALS en bv ook bij principale komponent analyse, waar faktorladingen voor items het verschil in belang voor de gevonden eigenwaarde weergeven. In de laatste gevallen tellen belangrijke items zwaarder mee bij het bepalen van de positie op de (attitude) schaal.
- de Likert methode gaat ervan uit, dat de variabelen op interval nivo gemeten zijn. De HOMALS oplossing kan men interpreteren als een toets voor deze aanname : daar wordt immers totaal niets verondersteld over meetnivo en vat men de items aanvankelijk als nominaal op. De transformaties geven de indruk, dat men (na de genoemde herkoderingen) de items (behoudens enkele kleine violaties) als ordinaal opgevat kunnen worden. Voor sommige variabelen zijn de afstanden tussen de kategorieën onderling ongeveer even groot (A13) na transformatie, bij andere variabelen (A09, A10, SF5) is dit duidelijk niet het geval : dan schijnt het verschil tussen bv *mee eens* en *noch mee eens,noch mee oneens* minimaal te zijn. HOMALS hecht aan het verschil tussen deze twee kategorieën ahw geen waarde, terwijl in de Likert methode het onderscheid tussen deze twee kategorieën voor alle variabelen even zwaar meetelt.

De voordelen van het werken met HOMALS in vergelijking tot de Likert schaal kunnen als volgt samen gevat worden. Men hoeft niet eerst een item-set te konstrueren, die tot een één-dimensionale schaal leidt : hoe één-dimensionaal de schaal is blijkt vanzelf uit de oplossing. Verder is men niet gebonden aan 'Likert' items : items met allerlei soorten kategoriseringen kunnen opgenomen worden, net als achtergrondsvariabelen. Men hoeft niet meer om te polen en kwasi-Likert items zoals hier eerst te herkoderen, omdat HOMALS automatisch zo 'herkodeert' dat

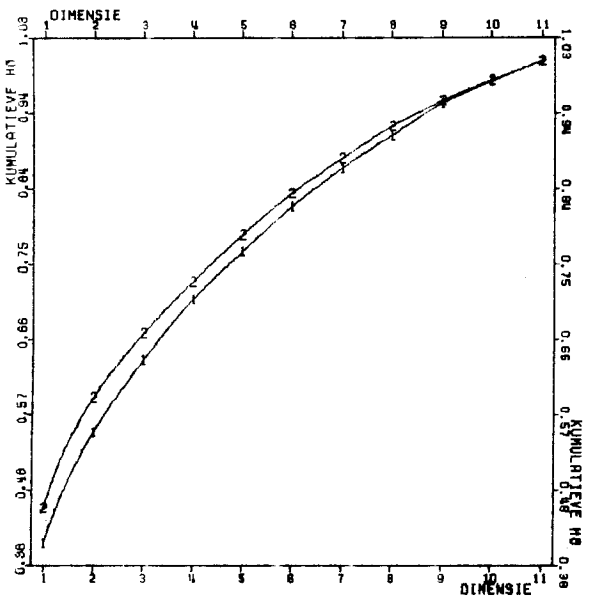
A. PCA OORSPRONKELIJKE SKORES



B. PCA GETRANSFORMEERDE SKORES



C. KUMULATIEVE HOMOGENITEIT
OORSPRONKELIJK(1) EN GETRANSFORMEERD(2)



de optimale onderliggende schaal wordt gevonden. Uit de oplossing blijkt dan vanzelf welke items belangrijk zijn in de 'attitude' schaal. Men weegt de items voor het diskriminatievermogen. Kategorie transformaties leiden tot informatie over lineariteit van de items en men hoeft geen aannames te maken over schaalnivo. Hoewel in dit geval beide oplossingen niet heel erg veel verschillen, biedt HOMALS zoveel meer informatie en zijn de mogelijkheden ervan zoveel groter, dat er niet zo erg veel reden is om een Likert schaal te gebruiken.

Tenslotte is, om de twee methoden te vergelijken en om de al vermoede 2^{e} dimensie te achterhalen op beide korrelatie matrices uit tabel 2. 7. PCA gedaan. De oplossingen voor de eerste twee dimensies staan hiernaast. Veel verschil is er niet, maar dat bleek al uit de korrelatiematrices. Na HOMALS transformatie worden de abortus items naar elkaar toegetrokken. Ook dat was al bekend.
Wat wel nieuw is, is dat de tweede dimensie de abortus items en de sexuele vrijheid items zeer duidelijk onderscheidt. De 2^{e} dimensie heeft ondanks z'n veel kleinere homogeniteit dan de eerste (.1351 vs .4114 bij a. en .1349 vs .4541) dus best iets nieuws te vertellen.

figuur 2.10.

Guttman schaal : scalogram analyse

Zoals gezegd zijn de theoretische aspekten van deze schaalmethode al eerder in dit hoofdstuk aan de orde gekomen (2.2.3). We zullen daarom hier slechts een korte beschrijving geven en vooral ingaan op de betekenis van de resultaten en op de vergelijking met HOMALS analyse van dezelfde data en met de Mokken schaal (zie 2.2.2).
De Guttman schaal is een kumulatieve schaal, waarbij men ervan uitgaat, dat de items van elkaar verschillen omdat de een méér van iets heeft dan de ander : moeilijker is, groter is, afwijzender is, enz. Deze toename in 'moeilijkheid' (ofwel steeds minder skores in het positieve alternatief) wordt geacht een één-dimensionale schaal te representeren : als men bij een 'moeilijk' item in de positieve kategorie skoort, wordt men geacht dit voor alle makkelijkere items ook te doen, als men het eens is met een extreem negatieve uitspraak, zal men op alle iets minder extreem negatieve uitspreken zeker ook ja zeggen. Hierbij kan het positieve alternatief meerdere kategorieën bevatten ; als men 'Likert' items zou nemen, zou men *zeer mee eens* en *mee eens* als het positieve alternatief kunnen definiëren en *noch mee eens*,*noch mee oneens* , *mee oneens* en *zeer mee oneens* samen als het negatieve alternatief. Ook kan men één item met meerdere kategorieën opdelen in verschillende binaire alternatieven. Dit verandert verder niets aan het uitgangspunt van de Guttman schaal.
Uitgaande van de kumulativiteit van de Guttman schaal geldt ten eerste, dat als persoon A een hogere skore heeft dan persoon B, hij -bij een perfekte Guttman schaal- op alle items, waarop persoon B bevestigend heeft geantwoord, ook ja heeft gezegd, plus nog op minstens één moeilijker item. Als men bovendien de schaalpositie van een individu kent (dwz het aantal positieve alternatieven) kan men hieruit z'n skore wat betreft alle items voorspellen.

Bij een perfekte Guttman schaal heeft men dus te maken met binaire items en kunnen slechts een beperkt aantal antwoord patronen voorkomen (zie de mogelijkheden voor 5 items hiernaast). Van de items neemt men verder slechts aan dat ze geordend zijn.

antwoord patroon	'moeilijkheid' item laag 1 2 3 4 5 hoog	schaal-positie
A	0 0 0 0 0	0
B	1 0 0 0 0	1
C	1 1 0 0 0	2
D	1 1 1 0 0	3
E	1 1 1 1 0	4
F	1 1 1 1 1	5

tabel 2. 8 : perfekte Guttman schaal.

ITEM	A01		A05		A03		A06		A08		A02		A07		A04		TOTAL
RESP	E	O	E	O	E	O	E	O	E	O	E	O	E	O	E	O	
	ERR		ERR		ERR		ERR		ERR		ERR		ERR		ERR		
n 8	0	17	0	17	0	17	0	17	0	17	0	17	0	17	0	17	17
i	ERR																
e 7	40	4	3	41	1	43	0	44	0	44	0	44	0	44	0	44	44
t			ERR														
6	38	6	36	8	14	30	0	44	0	44	0	44	0	44	0	44	44
m					ERR												
e 5	97	4	91	10	98	3	4	97	6	95	4	97	1	100	2	99	101
e							ERR										
4	56	4	57	3	57	3	17	43	21	39	20	40	8	52	4	56	60
e									ERR								
e 3	47	1	47	1	45	3	31	17	23	25	21	27	16	32	10	38	48
n											ERR						
s 2	45	5	49	1	46	4	36	14	38	12	30	20	34	16	22	28	50
													ERR				
1	33	3	36	0	34	2	32	4	23	13	29	7	35	1	30	6	36
															ERR		
0	143	0	143	0	143	0	143	0	143	0	143	0	143	0	143	0	143
SUMS	499	44	462	81	438	105	263	280	254	289	247	296	237	306	211	332	543
PCTS	92	8	85	15	81	19	48	52	47	53	45	55	44	56	39	61	
ERR.	0	27	3	23	15	15	4	78	27	50	45	27	59	1	68	0	442
B.C.	.3556		.7143		.7083		.9049		.8438		.8668		.9568		.9284		

COEFFICIENT OF REPRODUCIBILITY = .8983
MINIMUM MARGINAL REPRODUCIBILITY = .6680
PERCENT IMPROVEMENT = .2302
COEFFICIENT OF SCALABILITY = .6935

(E - eens)
(O - oneens)

tabel 2.9. SPSS

In veel gevallen zal er geen sprake zijn van een perfekte kumulatieve schaal. Afwijkingen kunnen veroorzaakt worden door meet-fouten, ofwel random error, ofwel door het feit dat er geen één-dimensionale onderliggende schaal bestaat.

Men gebruikt dan de schaalmethode, in dit geval Guttman's scalogram analyse, als kriterium, om te kijken hoever de gevonden schaal afwijkt van een perfekte Guttman schaal.

In tabel 2. 9 ziet men de resultaten van een Guttman scalogram analyse op de abortusvariabelen A01 t/m A08. In tabel 2. 10 staan de resultaten van een HOMALS analyse uitgaande van dezelfde data.

De Guttman schaal tabel kan men als volgt interpreteren. De items zijn geordend wb moeilijkheidsgraad. Een fout treedt op als men de skore op een item moet veranderen om niet in strijd te zijn met een perfekte Guttman schaal. Hoe meer fouten er zijn, hoe minder reden er is om een onderliggende ééndimensionale kumulatieve schaal te veronderstellen. Guttman (1944) suggereerde de *REP*-coëfficient, ofwel de *coefficient of reproducibility* als index voor de mate waarin een empirisch gevonden schaal een perfekte Guttman schaal benadert:

$$REP = 1 - \frac{\text{aantal fouten}}{\text{totale aantal antwoorden}},$$

opt. transf. kat. skores

VAR	EENS	ONEENS	DISKR. MAAT
A01	-0.07	0.82	0.058
A02	-0.88	0.72	0.623
A03	-0.29	1.22	0.342
A04	-1.05	0.65	0.673
A05	-0.24	1.33	0.308
A06	-0.84	0.78	0.657
A07	-0.98	0.74	0.718
A08	-0.84	0.72	0.598
	eigenwaarde		0.497

tabel 2. 10. HOMALS

Voor de waarde van de REP wordt door Guttman .85 aangegeven als minimum voor het veronderstellen van een één-dimensionale schaal. We zien, dat de abortus-items aan dit kriterium beantwoorden.
Een van de bezwaren van de REP-coefficient is, dat wanneer de rangorde van de items bepaald wordt met behulp van dezelfde steekproef als waarop we onze schaal-analyse willen uitvoeren, het maximale aantal fouten kleiner zal zijn dan het totale aantal antwoorden, zodat de minimale REP , REP_{min} of MMR (coefficient of minimum marginal reproducibility) groter dan nul is en zelfs vaak groter dan .50 (Mokken, 1971 p.50-54). MMR wordt gedefinieerd als :

$$MMR = \frac{\text{totale aantal antwoorden in modale kategorieën}}{\text{totale aantal antwoorden}}$$

Edwards (1957) stelde, dat alleen als REP veel groter is dan MMR en R groter of gelijk aan .85, men kan aannemen, dat de items een één-dimensionale kumulatieve schaal vormen. In het 'abortus'-geval is REP .8983 en MMR .6680 , maar hoeveel is veel groter?

Nog steeds levert de Guttman schaal echter problemen op. Niemöller (1976) noemt de volgende:

- Gaan we uit van de nulhypothese, dat een verzameling van k items bij gegeven steekproef marginalen door de subjekten uit de steekproef random beantwoord wordt, dan kan die onder de nul-hypothese verwachte REP zeer hoog zijn, vooral indien het aantal items klein is. Zo hoog zelfs, dat de items zich als schaal kwalificeren.
- De MMR als regel niet precies benaderd kan worden en dus benaderd moet worden. Zo kan voor het SPSS-programma GUTTMAN SCALE (dat ook in het abortus voorbeeld gebruikt is), waar deze benadering gebruikt wordt, gemakkelijk worden aangetoond dat de MMR uit het voorbeeld in de SPSS manual in feite kleiner is dan door de computer berekende waarde (Niemöller, 1976 , p. 20).
- Gezien het tweede probleem blijft ook de coefficient of scalability problematisch. In verband hiermee worden zgn S_2 coëfficienten gesuggereerd, die rekening houden met het verwachte aantal fouten onder de nulhypothese :

$$S_2 = 1 - \frac{E}{E_0} \quad ,$$

waarbij E het totaal aantal fouten is en E_0 het verwachte aantal fouten onder de nul-hypothese van random reponse.
Een van deze coefficienten van het S_2 type is Jane Loevingers *coefficient of homogenity* die bij het model van Mokken terugkomt, want:

Na alle kritiek op het deterministische model wordt het tijd daar iets positiefs tegenover te stellen. We kunnen dat niet beter doen dan met een algemene behandeling van het probabilistiese model van Mokken (Niemöller, 1976 : p. 21).

Hoewel we hier inmiddels met andere alternatieven hebben kennis gemaakt en we het model van Mokken hier slechts oppervlakkig zullen behandelen, kunnen we hier toch ook de resultaten van de Mokken schaalanalyse op de abortusitems A01 t/m A08 presenteren.

tabel 2.11.
Mokkenschaal abortusitems

Var	moeilijkheid	H(I)	DELTA(I) *
A01	.9183	.3928	8.7100
A02	.4452	.6450	29.9607
A03	.8000	.6943	21.7567
A04	.3774	.7434	31.4437
A05	.8436	.7203	21.0286
A06	.4783	.6824	30.9557
A07	.4243	.7152	32.6700
A08	.4504	.6312	29.3177
schaal		.6730	53.2012
RHO(1) = .8869		RHO(2) = .8809	
ALPHA = .50	LOWERBOUND H = .3000		

De Mokken-schaal is hier gebruikt als schaal-konstruktie methode : uitgaande van de 8 items werd gekeken welke samen een één-dimensionale schaal vormden. Alle items konden in deze schaal opgenomen worden.
Hoewel de Mokken schaal al in 2.2.2 besproken is, hier nog kort enkele kenmerken :
In tegenstelling tot de Guttman-schaal is de Mokken-schaal probabilistisch, dwz dat naarmate de schaalpositie van een individu hoger is, wordt voor ieder item de kans dat hij het positieve alternatief skoort groter. De items zijn daarbij dubbel monotoon, ofwel *holomorf* . Dwz dat, zoals gezegd, bij toenemende schaalpositie da kans om het positieve alternatief te skoren groter wordt en bovendien dat, als bij één bepaalde schaalpositie de kans op het positive alternatief van item A groter is dan van item B, dit voor alle schaalposities geldt : de tracelines snijden elkaar niet.
De moeilijkheid van een item (zie tabel 2.11) geeft aan, hoe vaak het positieve alternatief geskoord is.
De homogeniteits-coefficient H(I) voor alle items en H voor de totale schaal geven respektievelijk aan hoe goed de items in de schaal passen en hoe homogeen de uiteindelijke schaal is. Meer in 't bijzonder is H(I) de som van de kovarianties van item i met de andere items, gedeeld door de som van de maximale kovarianties die i gegeven de marginalen met de anderen zou kunnen hebben.
H is de som van alle kovarianties gedeeld door de som van alle maximale kovarianties. DELTA* is een statistiek die gebruikt wordt om te kijken of H signifikant afwijkt van nul en is asymptoties standaard normaal verdeeld.

ALPHA geeft het betrouwbaarheidsnivo aan De hier gevonden waarden leiden ons ertoe om te zeggen dat de nul-hypotese dat de schaalkoefficient H op random response is gebaseerd, verworpen kan worden op 10^{-100} nivo (zie voor de preciese betekenis en motivatie voor deze berekeningen Niemöller, 1976: 2.3.4.1).

De betrouwbaarheidscoefficiënt van de verzameling items kan op twee manieren worden berekend (zie Niemöller , 1976 : 2.3.4.3). Beide manieren leiden hier tot een hoge coefficiënt (RHO).

De in de tabel vermelde LOWERBOUND tenslotte zegt iets over kriteria van schaalkonstruktie : tijdens het zoekproces wordt een item niet aan een schaal toegevoegd, ten eerste als de H(I) van dat item beneden die ondergrens ligt en ten tweede als het toevoegen van dat item de H voor de totale set beneden die grens zou doen dalen.

Als we nu de HOMALS oplossing, de Guttman-schaal en de Mokken-schaal van dezelfde 8 abortus variabelen vergelijken, kunnen we het volgende opmerken :

- als we de items ordenen wb schaalbaarheid, krijgen we vogens de diskriminatiematen van HOMALS , de biseriële korrelaties van Guttman en de schaalbaarheidscoefficiënten H(I) van Mokken :

HOMALS -	A07	A04	A06	A02	A08	A03	A05	A01
Guttman-	A07	A04	A06	A02	A08	A05	A03	A01
Mokken -	A04	A05	A07	A03	A06	A08	A02	A01

  De Guttman volgorde komt vrijwel overeen met die van HOMALS. Mokken komt tot in totaal andere indeling. In alle drie de gevallen past A01 duidelijk het slechts in de schaal.

- HOMALS en Mokken kunnen beide gezien worden als een alternatief voor de Guttman schaal, waarop veel van de daarop geuite kritiek niet van toepassing is. De Mokken-schaal heeft het bezwaar, dat als de items verschillende dimensies vertegenwoordigen, er verschillende schalen worden gekonstrueerd, waarvan de onderlinge relatie niet geheel duidelijk is. Als een item bij HOMALS slecht bij de andere past, wordt het in de totale schaal ook minder meegewogen. Men merkt aan de diskriminatiemaat of een item slecht in de schaal past en als dat zo is, heeft hij ook weinig aan de schaal bijgedragen.
- HOMALS heeft bovendien, zoal ook al bij de bespreking van de Likert schaal werd uiteengezet, het voordeel dat er geen veronderstellingen gedaan hoeven te worden over de onderliggende schaal; men hoeft dus

niet uit te gaan van een 'kumulatieve' schaal. Bovendien kan men ook allerlei soorten items, zoals bv 'Likert' items of achtergrondvariabelen in de set opnemen. Als dan sommige items een kumulatieve schaal blijken te vormen, blijkt dit vanzelf uit de kategorie transformaties. Deze geven ook informatie over de relatie tussen de items die meer behelst dan alleen de volgorde : de afstanden tussen de positieve alternatieven van alle items (en de negatieve) zijn dan precies te bepalen.

Door de flexibiliteit en de hoeveelheid informatie lijkt het niet teveel gezegd, dat HOMALS ook bij attitude schaling een heel prettige techniek is om een onderliggende één-dimensionale schaal van een set items te bepalen.

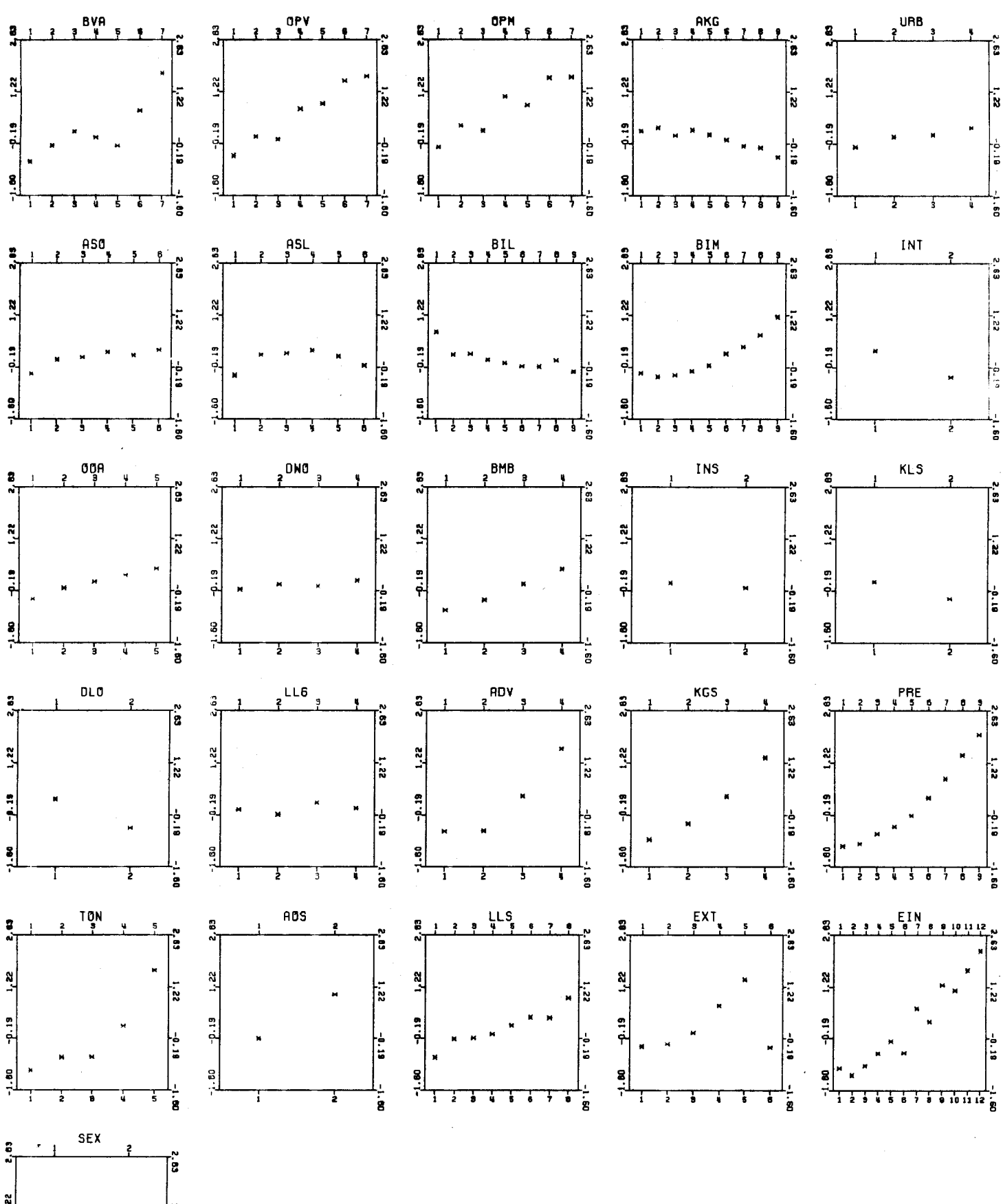
2.4.4. Survey onderzoek: van Jaar tot Jaar.

De gegevens uit het Van Jaar tot Jaar onderzoek (zie voor een uitgebreide beschrijving Appendix B.1)zijn op verschillende manieren met HOMALS geanalyseerd. Allereerst is gekeken naar de resultaten voor de totale groep wb de optimale transformaties van de kategorieën van de 26 variabelen. Een overzicht van deze transformaties, uitgezet tegen de oorspronkelijke kategorieskores, vindt u op de volgende pagina in figuur 2. 11. Deze weergave maakt het mogelijk de transformaties in detail te bekijken. Allereerst kan het belang van een variabele (de korrelatie met de 'schaal') afgelezen worden aan de spreiding in vertikale richting. Grote spreiding betekenent een hoge waarde van de diskriminatiemaat (zie tabel 2. 14 op pag.124. variabelen met een lage diskriminatiemaat zijn niet erg goed in de rest 'in te passen'. Belangrijke variabelen zijn hier duidelijk EIN , TON , PRE , en ADV . De onderliggende schaal kan duidelijk geïnterpreteerd worden als 'schoolsukses'.

Het geslacht van de leerling heeft daar blijkbaar niet erg veel mee te maken.

Mbt de transformaties kunnen hier verschillende eigenschappen van het programma en verschillende kenmerken van de data duidelijk gemaakt worden. In de eerste plaats korrigeert HOMALS dikwijls voor scheefheid van de marginalen ; we vinden dan monotone transformaties die konvex of konkaaf zijn, afhankelijk van de aard van de scheefheid. Een voorbeeld hiervan is ASO. In de tweede plaats korrigeert HOMALS voor te zware staarten. U-vormige verdelingen komen in de 'Van Jaar tot Jaar' gegevens niet voor, maar URB bv heeft een bijna rechthoekige frekwentieverdeling, wat tot gevolg heeft dat de transformatie de beide uiteinden relatief ver weg zet. Als deze verschijnselen na inspektie van de plotjes niet erg opvallend of zelfs niet erg zichtbaar lijken, is het van belang eraan te denken, dat het in deze twee voorbeelden ging om relatief onbelangrijke variabelen, waarvan de kategorieën dicht bij elkaar liggen. Als de kleine plotjes niet op dezelfde schaal waren gemaakt, maar iedere variabele gebruik kon maken van de volledige hoogte van z'n plotje, zouden bovenstaande opmerkingen geloofwaardiger zijn, maar dan heb je weer het probleem dat het belang van de verschillende variabelen niet naar voren komt, omdat de verschillen in spreiding geen rol meer spelen. Het risiko hiervan is ook dat men kleine verschillen binnen onbelangrijke variabelen teveel gewicht gaat geven, omdat ze op het oog net zo belangrijk zijn als grote verschillen tussen kategorieën van belangrijke variabelen. Uiteindelijk leek deze weergave toch de minste bezwaren te hebben.

Maar we zijn afgedwaald. De algemene vorm van de transformatie kan dus dikwijls begrepen worden uit de marginalen, tezamen genomen met het idee dat HOMALS naar multinormaliteit tracht te transformeren. De meest regelmatige transformatie van numerieke variabelen vinden we dus bij afgeronde stanine skores (hier BIL , BIM en PRE).



figuur 2.11. Optimaal getransformeerde kategorie-skores van Jaar tot Jaar, totale groep (n=1345).

Interessanter is welke transformaties we vinden voor typisch nominale variabelen zoals BVA , OPV , OPM , ADV , TON en EIN (hoewel bv Dronkers , 1978 - zie de beschrijving van de variabelen - er,mede op grond van andere literatuur, niet voor schroomt een variabele als EIN als ordinaal te beschouwen) . De transformaties zijn over het algemeen monotoon met de kategorienummers en wijken op het eerste gezicht weinig van lineair af. Dit bevestigt globaal de aanname van lineariteit waarop eerdere analyses, zoals Dronkers (1978) en Dronkers en Jungbluth (1979) (zie de onderzoeksbeschrijving), gebaseerd waren . Er zijn echter interessante afwijkingen van lineariteit te vinden. Bij BVA bv blijkt, dat de zelfstandige middenstand (4) en boeren en tuinders (5) lager op de 'schoolsukses-schaal'- liggen dan aanvankelijk werd gedacht. Ook bij opleidingsnivo van de vader en moeder zijn de afwijkingen van lineariteit interpreteerbaar. Bij EIN is er wel heel duidelijk iets aan de hand: het 'verwachte' ordinale karakter van deze variabele wordt met konsistente afwijkingen gerepresenteerd : waar men aanvankelijk dacht dat zonder diploma van een hogere opleiding meer was dan met diploma van een lagere opleiding vertrekken (bv lbo-met (6) versus ulo/havo/mavo/mms/vhmo - zonder (7)) . De HOMALS oplossing lijkt aan te geven dat schoolsukses zeker in relatie staat met het behalen van een diploma , dat bij 'sukses' de 'hoogte' van de schoolsoort niet zo'n grote rol speelt als verwacht, maar het wel of niet behalen van een diploma juist zwaarder meetelt.

Aardig is, dat bij EXT de 6^{e} kategorie (geen extracurriculaire aktiviteiten) op de goede plaats (beneden 1 aktiviteit) terecht komt. Bijna alle afwijkingen van lineariteit kunnen zo uit de betekenis van de betrokken kategorieën verklaard worden.

Aparte HOMALS analyses uitgaande van verschillende subgroepen in de data kunnen inzicht geven in verschillen in belang van variabelen voor de subgroepen. Bij vergelijking van de diskriminatiematen uit de HOMALS analyses voor jongens en meisjes apart (zie tabel 2.12) ziet men bv dat het merendeel van de achtergronds-, gezins-variabelen voor meisjes belangrijker is. Ook is bv voor meisjes de opleiding van de vader (OPV) belangrijker dan voor jongens en het verschil tussen OPV en OPM is bij meisjes ook duidelijk groter. AOS is voor jongens belangrijker ; dit komt waarschijnlijk door verschillende kenmerken van scholen waar jongens en meisjes na het L.O. naar toe gaan (huishoudscholen hebben bv een andere struktuur dan LBO opleidingen). Ook bij EXT en LLS spelen kenmerken van sexe-specifieke scholen een rol.

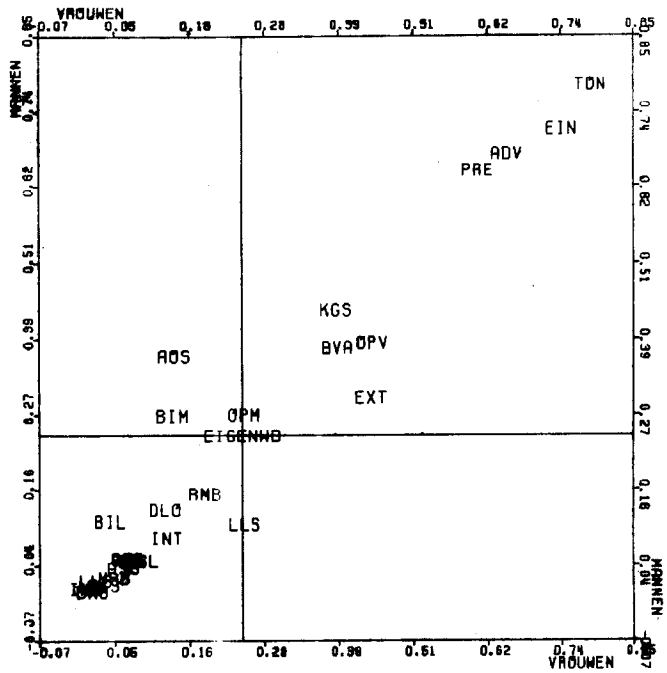
Men zou de uitkomsten voor de aparte analyses van jongens en meisjes kunnen weergeven als in tabel 2.12 , met dan de optimale transformaties voor beide sexen tegen elkaar uitgezet. Hier is gekozen voor een kompaktere weergave van de resultaten : in een figuur (2.12) zijn de diskriminatiematen van de variabelen tegen elkaar uitgezet. Het assenstelsel in de plot geeft per sexe de gemiddelde diskriminatiemaat, ofwel de homogeniteit van de oplossing weer. Variabelen, die voor beide groepen even belangrijk zijn, zouden op een rechte lijn, die met een hoek van 45^{o} door het snijpunt van de homogeniteiten loopt, moeten liggen. Afwijkingen van

deze lijn (zoals bij AOS, LLS, EXT , BIL , BIM en KGS) geven aan dat een variabele niet voor beide groepen even belangrijk is. Het gaat hier verder toch echter om relatief kleine verschillen. De echt belangrijke variabelen spelen bij beide groepen een grote rol, terwijl het ook niet voorkomt, dat één variabele bij de ene groep belangrijk is en bij de andere groep totaal niet. Ook de homogeniteit van de oplossing wijst niet in de richting van verschil tussen jongens en meisjes.

tabel 2.12
diskriminatiematen HOMALS voor vrouwen, mannen en totaal

	VROUW	MAN	TOTAAL
BVA	.388	.376	.384
OPV	.442	.384	.411
OPM	.244	.274	.259
AKG	.065	.055	.055
URB	.044	.026	.034
ASO	.058	.041	.049
ASL	.088	.052	.067
BIL	.038	.112	.059
BIM	.134	.273	.186
INT	.125	.087	.109
OOA	.063	.055	.053
DWO	.011	.004	.008
BMB	.184	.153	.165
INS	.000	.009	.003
KLS	.028	.014	.021
DLO	.123	.129	.125
LL6	.016	.021	.015
ADV	.652	.670	.651
KGS	.386	.434	.412
PRE	.606	.645	.627
TON	.783	.776	.774
AOS	.137	.364	.213
LLS	.245	.107	.145
EXT	.440	.301	.359
EIN	.737	.709	.711
SEX	----	----	.001
hom.	.2414	.2428	.2269

figuur 2.12
diskriminatiematen HOMALS voor vrouwen (hor.) en mannen (vert.)

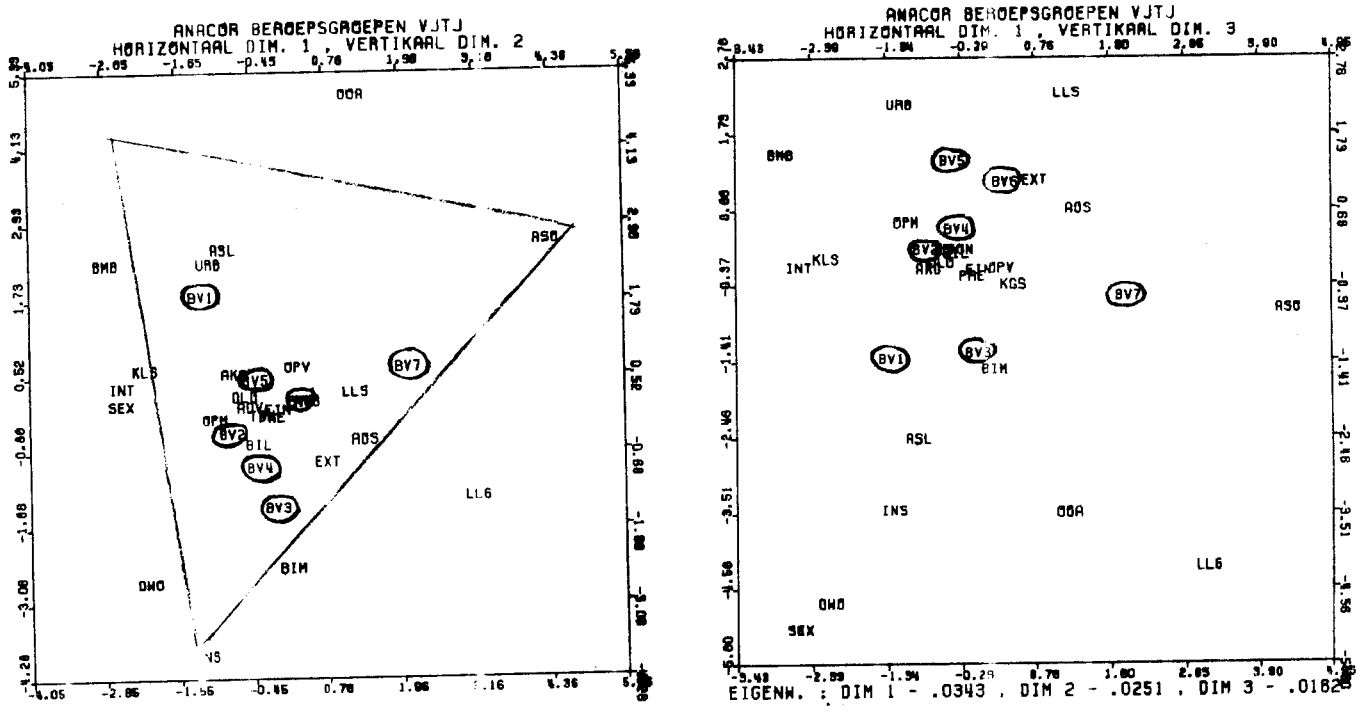


Ook per beroepsgroep zijn afzonderlijke HOMALS analyses gedaan. De resultaten vindt u in tabel 2.13 . Opvallend is de toename in belang van de variabelen die kenmerken van het secundair onderwijs weergeven naarmate het beroep van de vader 'hoger' wordt. Duidelijke verschillen tussen beroepsgroepen komen verder naar voren bij OPV (.064 vs .202) , ASO (.011 vs .295) , INT (.004 vs .130) en BMB (.007 vs .163) .

Een afbeelding van deze tabel vindt u in figuur 2.13 . Deze afbeelding is tot stand gekomen door het gebruik van de diskriminatiematen als invoer van ANACOR, de naam van zowel een programma als een techniek, waarmee men rijen en kolommen van een willekeurige matrix met niet-negatieve elementen in een euclidische ruimte kan weergeven. Deze techniek zal later in hoofdstuk 3 nog uitvoerig aan de orde komen. In figuur 2.13 zijn ook de beroepsgroepen weergegeven Inspektie van de eerste

	landarbeid + ongesch. hand	geschoold hand	uitvoerend hoofd	zelfst. middenstand	boeren + tuinders	midden-kader	akad. vrije ber. + leidinggevenden
OPV	.174	.064	.150	.185	.188	.165	.202
OPM	.098	.105	.127	.080	.151	.116	.056
AKG	.078	.142	.045	.071	.083	.036	.088
URB	.033	.028	.000	.030	.068	.006	.028
ASO	.064	.011	.043	.030	.095	.087	.295
ASL	.137	.027	.072	.020	.058	.069	.063
BIL	.077	.074	.127	.106	.133	.081	.086
BIM	.090	.171	.339	.288	.095	.116	.194
INT	.130	.056	.091	.113	.129	.056	.004
OOA	.110	.027	.009	.006	.015	.040	.110
DWO	.035	.031	.078	.018	.011	.006	.004
BMB	.163	.099	.015	.093	.156	.100	.007
INS	.008	.013	.033	.016	.006	.001	.005
KLS	.028	.028	.001	.035	.000	.023	.006
DLO	.173	.090	.171	.162	.167	.191	.117
LL6	.019	.027	.088	.025	.006	.010	.108
ADV	.613	.683	.646	.633	.711	.684	.580
KGS	.308	.293	.389	.292	.317	.417	.447
PRE	.576	.635	.723	.632	.684	.611	.691
TON	.666	.758	.783	.768	.824	.799	.727
AOS	.093	.084	.185	.253	.149	.295	.278
LLS	.062	.165	.127	.189	.328	.299	.309
EXT	.118	.340	.392	.375	.412	.479	.445
EIN	.617	.668	.779	.616	.766	.713	.762
SEX	026	.011	.000	.001	.002	.001	.009
e.w.	.1799	.1852	.2165	.2015	.2222	.2160	.2269

tabel 2.13. Diskriminatie maten per beroepsgroep.



figuur 2.13 ANACOR op diskriminatiematen uit HOMALS per beroepsgroep

drie dimensies leert, dat de belangrijkste variabelen, die voor iedere beroepsgroep een hoge diskriminatiemaat hebben (EIN , TON , PRE , ADV) in het centrum van de plot liggen. Variabelen, waarvan de diskriminatiemaat per groep sterk varieert, liggen excentrisch. De groepen liggen weer meer in het centrum, waarbij iedere groep in de richting ligt van de variabele die hem van de ander groepen onderscheidt (vgl BV7-ASO , BV3-INS , BV7 en BV3-LL6 , BV1 en BV7-OOA).

De plotjes geven zo op een inzichtelijke manier de informatie uit een redelijk grote tabel neer.

Op grond van de HOMALS-analyses per sexe en per beroepsgroep zijn twee (voor iedere sexe) en zeven (voor iedere beroepsgroep) korrelatiematrices berekent uitgaande van de optimaal getransformeerse kategorieskores ipv de oorspronkelijke skores. Deze korrelatiematrices zijn dan weer gebruikt als invoer voor andere programma's als INDSCAL en TUCKALS. Het zou hier te ver voeren om op de resultaten van deze analyses in te gaan ; het is echter wel belangrijk zich te realiseren, dat transformatie leidt tot nieuwe metrische variabelen , terwijl men begonnen is met alle variabelen als nominaal te beschouwen. Allerlei nieuwe analyses zijn dan toepasbaar.

MARGINALE FREKWENTIES EN DISKRIMINATIEMATEN 'VAN JAAR TOT JAAR'

		KATEGORIEëN													
NAAM	M	1	2	3	4	5	6	7	8	9	10	11	12	D.M.1	D.M.2
BVA	35	408	366	171	235	197	331	102						.383	.137
OPV	17	761	358	325	146	87	56	95						.411	.143
OPM	11	1229	103	260	130	47	35	30						.259	.137
AKG	1	103	300	346	304	220	169	151	91	160				.055	.145
URB	0	464	457	516	408									.034	.144
ASO	75	422	386	281	258	267	156							.049	.053
ASL	0	454	373	307	278	258	175							.067	.148
BIL	35	118	163	249	299	298	307	198	87	91				.059	.121
BIM	35	96	156	195	323	316	294	177	164	89				.186	.124
INT	71	1162	612											.109	.124
OOA	75	260	477	521	376	136								.053	.051
DWO	75	218	349	390	813									.008	.054
BMB	75	255	263	479	773									.165	.053
INS	82	1482	281											.003	.050
KLS	0	1634	211											.021	.145
DLO	2	1295	548											.125	.141
LL6	4	118	478	709	536									.015	.141
ADV	43	230	756	522	294									.651	.182
KGS	21	215	732	659	218									.412	.142
PRE	0	67	128	262	299	334	319	237	113	86				.627	.174
TON	0	57	157	690	605	336								.774	.302
AOS	60	1440	345											.218	.114
LLS	190	69	250	413	302	159	113	146	203					.145	.100
EXT	60	394	393	318	282	185	213							.359	.111
EIN	7	62	24	99	153	100	629	61	357	25	161	115	52	.711	.292
SEX	0	924	921											.001	.143
													HOM.	.2269	.1334

tabel 2.14. Marginale frekwenties en diskriminatiematen vJtJ.

De hier gerapporteerde konklusies staan geheel buiten verantwoordelijkheid van de onderzoekers.

2.4.5. Veldstudie: Bloedbank.

We verrichten hier een sekundaire analyse op een gedeelte van de gegevens die door Stammeijer en Staallekker (1977) verzameld zijn t.b.v. de Stichting Rode Kruis Bloedbank. Het onderzoek richtte zich op de volgende vragen:

- Welke faktoren leiden tot donorschap van bloed?
- Welke faktoren leiden tot voortzetting van het donorschap?
- Welke klachten en/of bedenkingen hebben de donors?
- Welke faktoren leiden tot het afbreken van het donorschap?

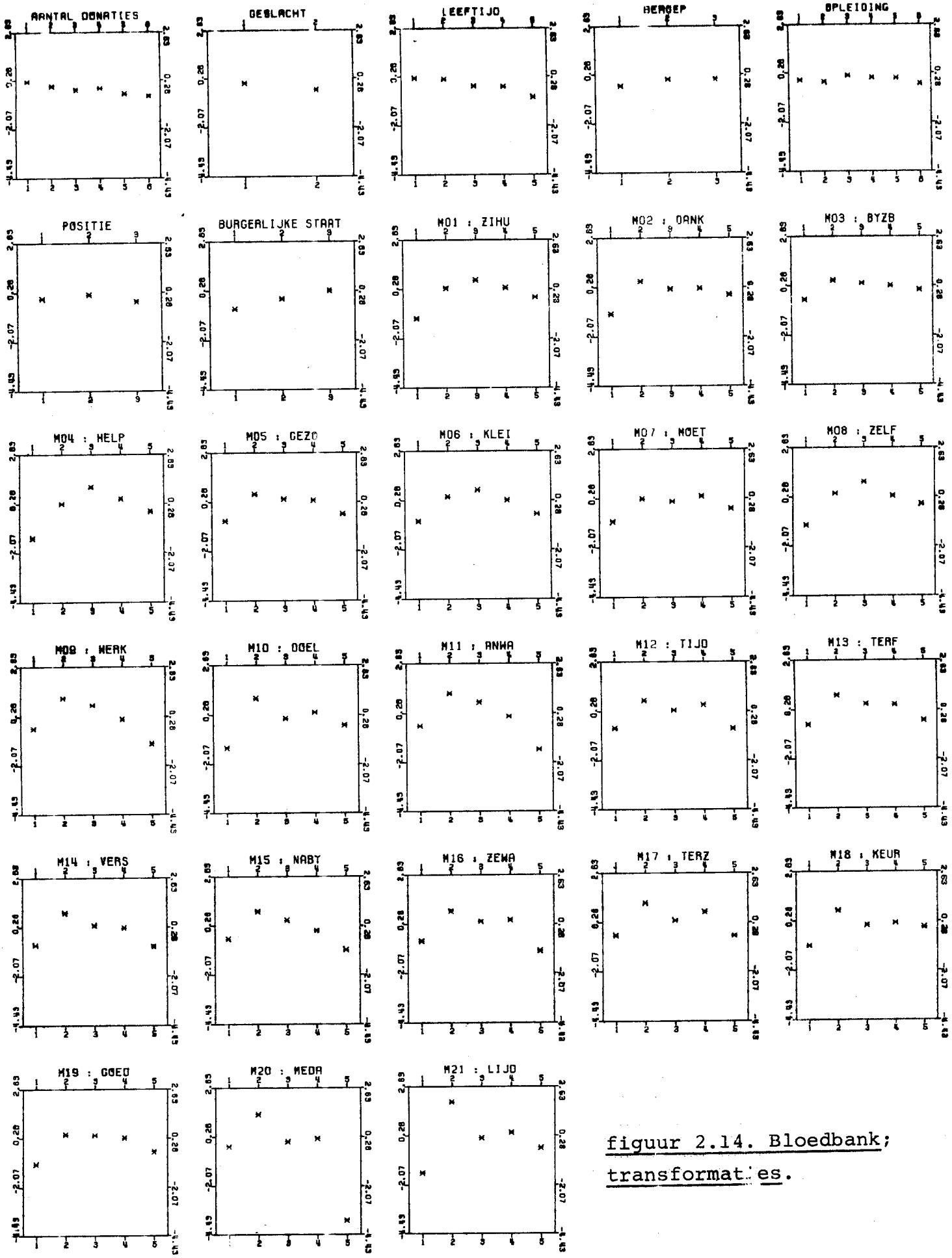
Men richtte zich hierbij op drie groepen in de Haagse regio, die ieder per post een vergelijkbare vragenlijst toegestuurd kregen, te weten: aanstaande donores, mensen die een aanmeldingsformulier hebben ingestuurd, na een oproep op de eerste keuring zijn verschenen en nog nooit eerder bloed hebben gegeven; donores, vrijwel alle voor de Haagse Bloedbank beschikbare donores; ex-donores, gewezen donores, die niet om één van de volgende redenen uit het donorbestand zijn afgevoerd: overlijden, bereiken van de 65-jarige leeftijd, mediese afkeuring of duidelijke mediese klachten en verhuizing. De vragenlijsten bevatten - aan de situatie van de drie groepen aangepaste - vragen over direkte aanleidingen om donor te worden, motieven voor het donorschap, klachten en bedenkingen t.a.v. de procedure en de gevolgen van bloedafname, de organisatie rondom de bloedbank, de kennis van de donor omtrent mediese konsekwenties voor het eigen lichaam en tenslotte vragen naar enkele achtergrondsgegevens.

Een dergelijke omvangrijke en ambitieuze onderzoeksopzet komt vaker voor bij zgn. beleidsondersteunend onderzoek en maakt dat men, om de grote lijn te bewaren, zelfs op onderdelen van het bestand de data grof bewerkt. In onze sekundaire analyse wordt uitsluitend aandacht besteed aan de donores en wel aan diegenen, die meer dan één maal bloed hebben gegeven. De onderzoekers zelf (pag. 41 ev.) hebben, ter verkenning van de faktoren die bij de voortzetting van het donorschap een rol spelen, relatief veel aandacht besteed aan de motieven voor het donorschap (deze staan getabelleerd in tabel 2.15, samen met een aantal achtergrondgegevens). Over deze motieven hebben ze "... een zogenaamde faktoranalyse uitgevoerd om een genuanceerder beeld te krijgen van de

	ACHTERGRONDGEGEVENS			
DONA	1) Aantal Donaties	1 : 2 - 9	3 : 20 - 29	5 : 40 - 49
		2 : 10 - 19	4 : 30 - 39	6 : 50 of meer
SEXE	2) Geslacht	1 : man	2 : vrouw	
LFTD	3) Leeftijd	1 : 19 - 29	3 : 40 - 49	5 : 60 of ouder
		2 : 30 - 39	4 : 50 - 59	
BERO	4) Beroep	1 : hoger	2 : middelbaar	3 : lager (eventueel van vader of echtgenoot)
OPLD	5) Opleiding	1 : laag	6 : hoog	
POSI	6) Positie	1 : loondienst	2 : zelfstandig	3 : geen beroepsbevolking
BURG	7) Burgerlijke Stand	1 : getrouwd gew.	2 : getrouwd	3 : ongetrouwd
	MOTIEVEN OM BLOED TE GEVEN			
	1 : speelt helemaal geen rol	*2 : speelt geen rol*	*3 : weet ik niet*	*4 : speelt een rol* / *5 : speelt een grote rol*
	Ik ben donor omdat ...			
ZIHU	1. ... ziekenhuizen erg veel bloed nodig hebben			
DANK	2. ... ik dankbaar ben dat ik zelf goed gezond ben			
BYZB	3. ... ik denk dat ik een weinig voorkomende bloedgroep heb			
HELP	4. ... ik anderen ermee kan helpen			
GEZO	5. ... het goed voor de gezondheid is			
KLEI	6. ... het eigenlijk een kleine moeite is			
MOET	7. ... ik vind dat eigenlijk iedere gezond mens het zou moeten doen			
ZELF	8. ... als ik het zelf nodig heb bij ziekte of operatie, ik óók wil dat het er is			
WERK	9. ... ik in werktijd bloed kan geven			
DOEL	10. ... het voor een goed doel is			
ANWA	11. ... anderen me vaak waarderen, als ik vertel dat ik bloed geef			
TIJD	12. ... het me weinig tijd kost			
TERF	13. ... ik een beetje terug wil geven voor het bloed dat een familielid/bekende in het ziekenhuis toegediend kreeg			
VERS	14. ... ik nu regelmatig vers bloed aanmaak			
NABY	15. ... er bij mij thuis/op school/op m'n werk véél zijn die bloed geven. Ik voel me min of meer verplicht het ook te doen			
ZEWA	16. ... ik mezelf meer waardeer als ik zoiets doe			
TERZ	17. ... ik een beetje terug wil geven voor het bloed dat ik zelf kreeg toen ik in het ziekenhuis lag			
KEUR	18. ... ik nu regelmatig gekeurd wordt			
GOED	19. ... het bloed geven me het gevoel geeft iets goeds te doen			
MEDA	20. ... ik na een bepaald aantal keren bloed geven een medaille krijg			
LIJD	21. ... ik het lijden van iemand anders ermee kan verlichten			

tabel 2.15. Omschrijving deelbestand BloedBank.

relaties tussen de diverse motieven"(pag. 41). Wij hebben de gegevens van 309 donoren met HOMALS bewerkt, en de één-dimensionale kwantifikaties in figuur 2.14 tegen de oorspronkelijke kategorie-labels uitgezet. In tabel 2.16 staan de marginale frekwenties van de variabelen en hun diskriminatiematen. Het eerste wat opvalt is dat de achtergrondvariabelen niet 'meedoen'. Zijn we hier een geheel nieuwe dimensie van de werkelijkheid op



figuur 2.14. Bloedbank; transformaties.

MARGINALE FRI KWENTIES BLOEDDONOR ONDERZOEK

Vars	/ Kat	M	1	2	3	4	5	6	Diskr.m.
Achtergrond	1	12	96	94	52	34	14	7	.034
	2	0	233	76					.021
	3	1	90	86	61	54	17		.061
	4	26	36	116	131				.011
	5	8	22	105	117	26	7	24	.021
	6	10	251	11	37				.003
	7	0	13	234	62				.037
Motieven	1	2	20	6	4	84	193		.127
	2	5	44	33	7	96	124		.192
	3	11	157	62	25	34	20		.151
	4	2	11	1	1	64	230		.129
	5	6	106	45	78	40	34		.269
	6	8	76	74	9	94	48		.239
	7	3	26	24	28	91	137		.132
	8	6	47	37	4	107	108		.246
	9	8	222	58	4	12	5		.334
	10	5	17	7	8	115	157		.198
	11	8	222	55	14	6	4		.394
	12	10	163	90	16	23	7		.365
	13	10	175	59	14	25	26		.324
	14	8	123	46	53	64	15		.329
	15	7	204	71	12	10	5		.326
	16	8	157	72	19	42	11		.415
	17	9	210	51	11	14	14		.373
	18	6	63	28	8	143	61		.243
	19	7	46	30	21	142	63		.234
	20	8	259	36	1	4	1		.292
	21	5	12	4	23	108	157		.251
								e.w.	.2054

tabel 2.16. Marginale frekwenties en diskriminatiematen Bloedbank.

het spoor? Het tweede wat opvalt is dat de transformaties niet stijgen, laat staan in enig opzicht lineair genoemd mogen worden, zodat de aanname dat het interval-schalen betreft, niet wordt bevestigd. De faktor-analyse resultaten van de onderzoekers (zie tabel 2.17) moeten dus met argwaan bekeken worden.

Wanneer we dit doen, schijnt de faktor-analyse in ieder geval te zeggen, dat als men HELP en LIJD belangrijke motieven vindt, men ook ZIHU, DOEL en DANK aankruist; ook oordelen over ANWA, WERK, ZEWA en MEDA komen geklusterd voor, maar niet noodzakelijk bij de mensen

	Faktor 1	2	3	4	5	6
Faktor-lading motief (als groter .40)	HELP .634 LIJD .580 ZIHU .513 DOEL .481 DANK .465	ANWA .864 WERK .580 ZEWA .432 MEDA .408	VERS .681 GEZO .654	KLEI .758	TERF .581 TERZ .553	GOED .568
Eigenw.	3.597	2.778	1.494	1.332	1.140	1.031
Kum. % var	17.1 %	30.4 %	37.5 %	43.8 %	49.2 %	54.2 %

tabel 2.17. Principale assen faktor-analyse met iteraties voor kommunaliteiten, gevolgd door VARIMAX rotatie over faktoren met eigenwaarde groter dan 1.

die het eerste kluster belangrijk vinden. We zouden de eerste faktor 'altruisme' en de tweede 'eigenbelang' kunnen noemen; de derde heeft iets met de gezondheidsaspekten te maken en de vijfde met afbetaling. De faktor-analyse vertelt ons een aantal dingen over de motieven, maar niet erg veel nieuws; over de donoren zijn we helemaal niet veel wijzer geworden.

We merken tenslotte nog op, dat HOMALS één-dimensionaal vooral de kategorieën 1 en 5 kontrasteert met 2,3 en 4, wat eigenlijk ook alleen maar zegt dat er waarschijnlijk een dijk van een 'response-bias' in zit. Hoe dit precies in elkaar zit, behandelen we in hoofdstuk 3.1., waar we de twee-dimensionale HOMALS oplossing bekijken.

3. HOMALS en ANACOR.

3.1. Geometriese benadering van HOMALS.

In deze sektie willen we laten zien, hoe de representatie die HOMALS geeft van individuen, variabelen en kategorien geometries in elkaar zit. Met name zullen we het hebben over wat we verstaan onder een 'ideale' representatie, hoe dan de HOMALS verliesfunktie voor de dag komt en hoe bepaalde eigenschappen van de data in een HOMALS oplossing naar voren komen.

Uitgangspunt is, dat we individuen afbeelden als punten in een euclidiese ruimte van dimensionaliteit p (we zullen hier voor de overzichtelijkheid in de voorbeelden steeds p=2, het platte vlak dus, nemen). De koordinaatwaarden van deze individu-punten op een bepaald gekozen assenstelsel staan verzameld in de matrix X, van afmetingen nxp, eerder genoemd de matrix van individu-skores. Verder beelden we ook de kategorien af als punten in dezelfde ruimte, en verzamelen de koordinaatwaarden van deze punten (hun meervoudige kwantifikaties) in de m matrixen Y_j, van afmetingen $k_j$xp. Alles wat we weten over individuen en kategorien staat gekodeerd in de indikator matrixen G_j; we interpreteren de elementen van G_j als een "horen bij"-relatie, d.w.z.

$$\begin{aligned} &\text{als } g^j_{ir} = 1 \quad \text{dan hoort individu i bij kat. r van var. j} \\ &\text{als } g^j_{ir} = 0 \quad \text{dan hoort individu i niet bij kat. r van var. j} \end{aligned} \tag{1}$$

en representeren de (empiriese) "horen bij"-relatie af als afstand (in de euclidiese ruimte). Bekijken we nu één variabele apart, dan behoort ieder individu maar tot één kategorie, en kunnen we dus zijn individu-punt zonder meer identificeren met het punt van de kategorie waar hij bij hoort:

$$\text{als } g^j_{ir} = 1 \quad \text{dan } \delta^j_{ir} = 0 \tag{2}$$

waarbij dus δ gebruikt wordt als symbool voor de euclidiese afstand. Nu is het natuurlijk zo, dat elk individu bij m kategorien hoort, en intuitief verwachten we niet, dat dan eis (2) voor alle variabelen tegelijkertijd kan opgaan als we ook de dimensionaliteit laag willen houden. Toch eisen we zonder blikken of blozen, dat voor een

ideale representatie geldt:

$$\text{als } g^{j}_{i\,r} = 1 \quad \text{dan } \delta^{j}_{i\,r} = 0 \quad \underline{\text{voor alle i,j in kleine p}}. \qquad (3)$$

Als we een X en een Y kunnen vinden waarvoor (3) geldt, spreken we van <u>perfekte homogeniteit</u>. Voorzover het niet lukt aan de (strenge) eisen (3) te voldoen lijden we <u>verlies</u>. Liever dan onze eisen af te zwakken (voorbeelden hiervan zullen we in hoofdstuk 10 tegenkomen), zullen we dit verlies kwantificeren in een <u>verliesfunktie</u> en ons ten doel stellen X en Y zodanig te vinden, dat de verliesfunktie een zo klein mogelijke waarde aanneemt.

Maar eerst bespreken we ter illustratie de zgn. GBS-data (Guttman's (1966) versie van een tabel uit het boek van Bell en Sirjamaki, 1961). De 'individuen' zijn hier <u>groepskonsepten</u>, de 'variabelen' zijn diverse <u>karakteriseringen</u> van die konsepten (zie tabel 3.1; om de terminologiese verwarring tot een minimum te beperken hebben we de termen niet vertaald).

nr.	konsepten	nr.	variabelen	kategorien
1	Crowd (CR)	1	Intensity of interaction (INTE)	a.slight
				b.low
2	Audience (AU)			c.moderate
				d.high
3	Public (PU)	2	Frequency of interaction (FREQ)	a.slight
				b.non-recurring
4	Mob (MO)			c.infrequent
				d.frequent
5	Primary group (PG)	3	Feeling of belonging (FEEL)	a.none
				b.slight
6	Secondary group (SG)			c.variable
				d.high
7	Modern Community (MC)	4	Physical Proximity (PHYS)	a.distant
				b.close
		5	Formality of relationship (FORM)	a.no relation
				b.formal
				c.informal

<u>tabel 3.1. Identifikatie van rij- en kolomobjekten van GBS-data</u>.

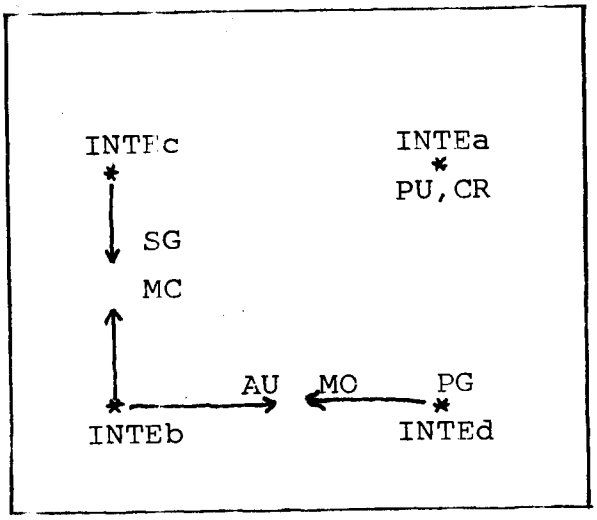
	1	2	3	4	5
CR	a	a	a	b	b
AU	b	b	b	b	b
PU	a	a	b	a	a
MO	d	b	d	b	c
PG	d	d	d	b	c
SG	c	c	c	a	b
MC	b	c	c	b	b

tabel 3.2.a
GBS data matrix

	INTE				FREQ				FEEL				PHYS		FORM		
	a	b	c	d	a	b	c	d	a	b	c	d	a	b	a	b	c
CR	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0
AU	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	1	0
PU	1	0	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0
MO	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	1
PG	0	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	1
SG	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	1	0
MC	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	1	0

tabel 3.2.b. GBS indikator super matrix.

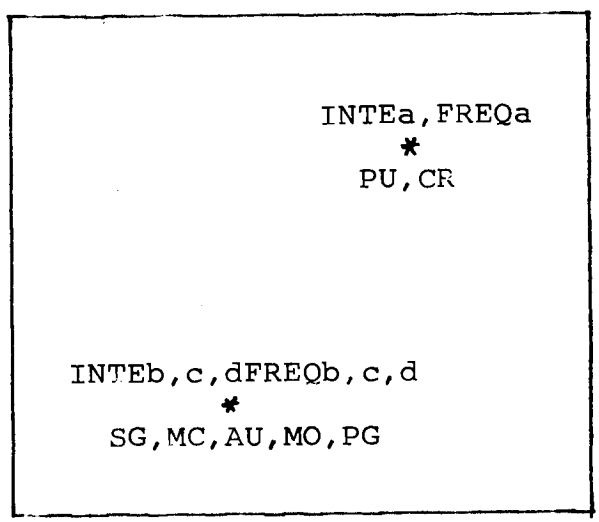
Het is vanzelfsprekend een zaak voor sociologen of sociaal-psychologen om te beoordelen in hoeverre de gebruikte variabelen en kategorien gelukkig gekozen zijn en in hoeverre de datamatrix juist is ingevuld; vanuit ons gezichtspunt kan aan de data matrix niet getornd worden, maar het is wel mogelijk om met een homogeniteitsanalyse uit te zoeken of de gebruikte karakteriseringen redundant zijn.



figuur 3.1. Willekeurige start voor GBS data.

In figuur 3.1 zien we, dat het natuurlijk geen enkel probleem is

om aan eis (2) te voldoen. De kategorien van de eerste variabele, INTEa t/m INTEd, zijn hier als vier willekeurige punten in het platte vlak weergegeven, en we leggen de punten die de konsepten moeten weergeven eenvoudigweg bovenop hun overeenkomstig kategoriepunt. Tevens zien we, dat om ook de tweede variabele in te passen, SG en MC naar elkaar toe bewogen zouden moeten worden om samen te met FREQb. AU en CR kunnen blijven liggen (allebei FREQa), terwijl PG gewoon met MO kan meeverhuizen (in z'n eentje FREQd). We krijgen nu figuur 3.2. Hierin wordt voldaan aan (3), voor de eerste twee



figuur 3.2. Eerste twee variabelen van GBS data.

variabelen en in feite met p=1. Het geval van perfekte homogeniteit is dus niet onmogelijk, hoewel meestal niet bijster informatief. Als we op deze manier doorgaan, zullen we steeds meer kategoriepunten samen moeten laten vallen om de representatie homogeen te houden, en inderdaad, FEELb zou ons dwingen om alle punten op elkaar te leggen. Er moeten dus extra eisen gesteld worden. Dit kan bijv. door af te spreken dat de konfiguratie van individu-punten gecentreerd is (u'X = 0, eigenlijk geen extra eis, afstanden zijn invariant onder translatie) en ortonormaal (X'X = I). Bovendien gaan we (3) uitwerken naar de HOMALS verliesfunktie.

Allereerst roepen we in herinnering de matrix van geinduceerde skores:

$$V_j \triangleq G_j Y_j \tag{4}$$

en merken op dat de indikator matrix hier precies dat kategorie punt uit Y_j selekteert dat hoort bij individu i; V_j geeft als het ware de koordinaatwaarden voor individu i "voor zover het ligt aan variabele j". Nu zegt (2) dus, dat voor individu i en variabele j moet gelden

$$\text{als } g^j_{ir} = 1 \text{ dan } \sqrt{\sum_{s=1}^{p} (x_{is} - v^j_{is})^2} = 0 \tag{5}$$

We laten nu verder als-dan weg, de wortel doet er duidelijk ook niet toe, en omdat (5) voor ieder individu geldt, krijgen we

$$\sum_{i=1}^{n}\sum_{s=1}^{p} (x_{is} - v^j_{is})^2 = 0 \qquad \text{voor alle } j = 1, \ldots, m \tag{6}$$

We brengen dit nu in matrix notatie en vervangen V_j weer door (4):

$$SSQ(X - G_j Y_j) = 0 \qquad \text{voor alle } j = 1, \ldots, m \tag{7}$$

In het perfekte geval zal ook de gemiddelde sum of squares gelijk zijn aan nul. In het niet-perfekte geval is dus een natuurlijke maat voor verlies:

$$\sigma(X;Y_1,\ldots,Y_m) = \frac{1}{m}\sum_{j=1}^{m} SSQ(X - G_j Y_j) \tag{8}$$

We zijn hier bij de HOMALS verliesfunktie aangeland, die we willen minimaliseren over alle Y en genormaliseerde X. Op details van het algoritme gaan we hier niet in (zie hoofdstuk 5), maar we merken alvast op, dat voor vaste X het minimum van $\sigma(X;Y_1,\ldots,Y_m)$ over Y_j bereikt wordt voor

$$\hat{Y}_j = (G_j'G_j)^{-1}G_j'X = D_j^{-1}G_j'X \triangleq U_j \tag{9}$$

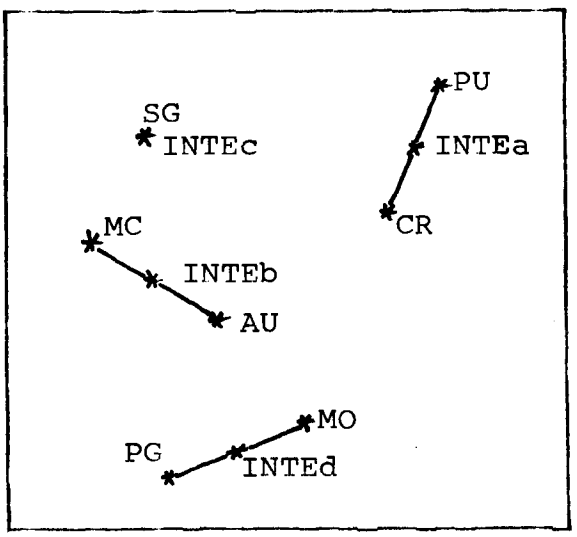
In woorden staat hier, dat als we de individu punten even op hun plaats laten en de beste positie van de kategorie punten van variabele j willen vinden, we het punt voor kategorie r moeten laten samenvallen met het <u>zwaartepunt</u> van de individu punten die horen bij r. De koordinaatwaarden van deze zwaartepunten staan verzameld in de matrix U_j, een $k_j$xp matrix van <u>geinduceerde kwantifikaties.</u>

Op overeenkomstige wijze vinden we voor vaste Y_j, dat het minimum van $\sigma(X;Y_1,\ldots,Y_m)$ over X bereikt wordt voor

$$\hat{X} = \frac{1}{m} \sum_{j=1}^{m} G_j Y_j = V. \qquad (10)$$

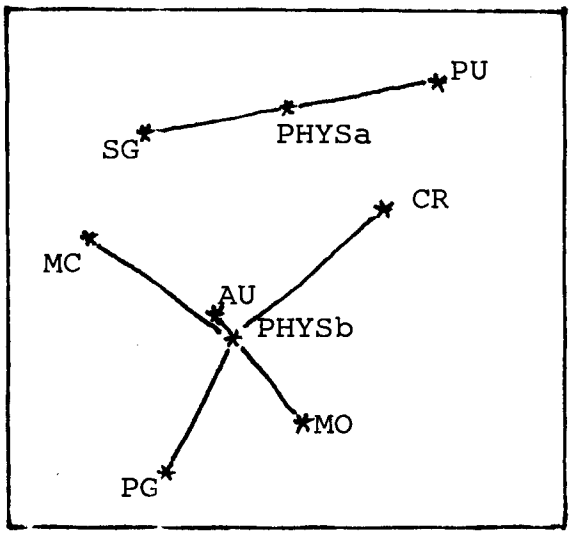
het gemiddelde van de geinduceerde skores (altans, in principe, want we moeten $\hat{X}$ ook nog laten voldoen aan de eis van ortonormaliteit). Het basisprincipe van het HOMALS algoritme is het afwisselen van stappen (9) en (10), en we kunnen HOMALS dus heel goed karakteriseren als een "method of reciprocal averages" (vgl. p. 71).

We illustreren een en ander nu voor de kategorien van variabele 1 van GBS. In figuur 3.3 hebben we eerst zeven punten voor de konsep-



figuur 3.3. Het bepalen van de kategorie punten van INTE voor vaste X.

ten getekend en daarna de vier zwaartepunten voor INTEa t/m INTEd bepaald. In figuur 3.4. hebben we dat ook gedaan voor de twee kategorien van variabele PHYS. Behalve dat deze plaatjes laten zien hoe we de kategorie punten bepalen, brengen ze ook het verlies in beeld. Immers, (8) zegt dat het verlies in de eerste plaats een gemiddelde is over variabelen en, in de tweede plaats, dat per variahet verlies gelijk is aan de gekwadrateerde lengtes van alle lijntjes die de kategoriepunten met de individu punten verbinden. Het zal dus intuitief duidelijk zijn dat we bijvoorbeeld INTEa in fi-

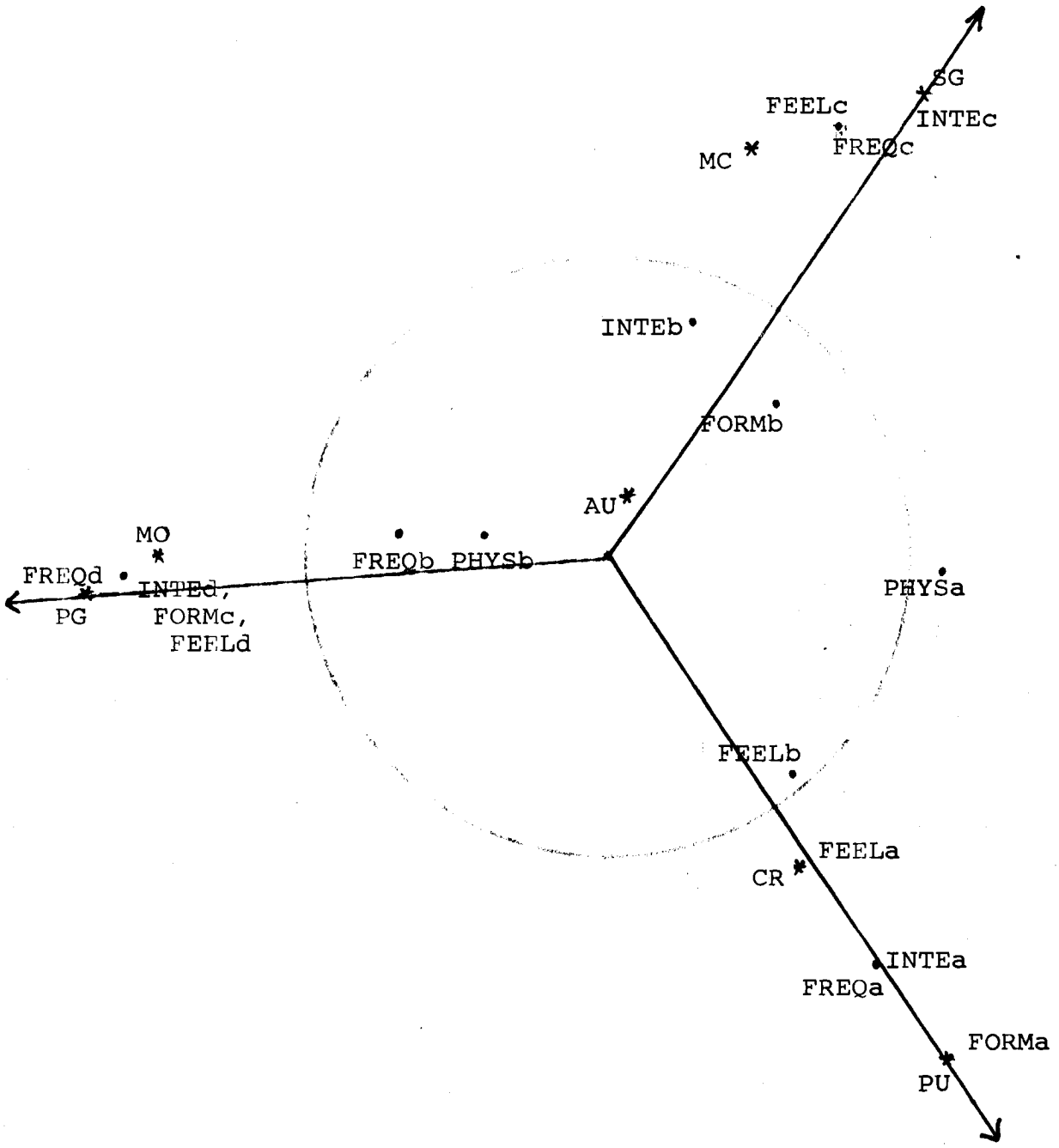


figuur 3.4. Het bepalen van de kategorie punten van PHYS voor vaste X.

guur 3.3 niet naar links of rechts kunnen verschuiven zonder het verlies te vergroten (de lengtes van de lijnen naar PU en CR zouden beide toenemen); maar we kunnen INTEa zelfs niet langs de lijn die PU en CR verbindt bewegen, omdat dan weliswaar niet de som van de lengtes, maar wel de som van de gekwadrateerde lengtes zou toenemen.

Als we figuur 3.3 en 3,4 onderling vergelijken valt het ook op, dat PHYS meer tot het verlies bijdraagt dan INTE. Dit lijkt niet eerlijk, want de variabelen worden in (8) niet gewogen. Maar we moeten bedenken dat de positie van de konsepten ook van de rest van de variabelen afhangt en bovendien hier provisories gekozen is; het is dus heel goed mogelijk dat de HOMALS oplossing een wat andere konfiguratie van konsepten geeft, waarin PHYS relatief beter past. Aan de andere kant is het in 't algemeen natuurlijk zo, dat een variabele met weinig kategorien meer verlies oplevert dan eentje met veel; een variabele met evenveel kategorien als er individuen zijn (k_j=n) levert nooit verlies op als we aan de positie van die kategorien geen extra eisen stellen (hier gaan we in hoofdstuk 5 verder op in).

We bekijken nu wat HOMALS van ons illustratief voorbeeld maakt; dit staat getekend in figuur 3.5. We hebben er wat lijnen bijgetekend om de interpretatie te vergemakkelijken. De konsepten vallen in 4



figuur 3.5. Gezamenlijke plot van konsepten en kategorien voor GBS data (HOMALS opl.), $\lambda_1 = .8047$, $\lambda_2 = .6245$, $\lambda_3 = .3972$.

groepen uiteen: in het midden Audience, gekarakteriseerd door het profiel bbbbb, waarvan de kategoriepunten alle binnen de getrokken cirkel liggen; aan de randen resp. Modern Community en Secondary group, gekarakteriseerd door c-kategorien, Mob en Primary group, gekarakteriseerd door d-kategorien, en Crowd en Public, gekarakteriseerd door a-kategorien. Een andere manier om dit te bekijken ontstaat wanneer we even de twee rechter poten alleen nemen: we krijgen dan van onder naar boven de volgorde PU, CR, AU, MC, SG, waarbij de volgorde van de kategorien per variabele overeenkomt met hun oorspronkelijke kodering: INTEa, INTEb, INTEc; FREQa, FREQb, FREQc; FEELa, FEELb, FEELc; FORMa, FORMc. De variabele PHYS valt hierbij eigenlijk uit de boot, die bepaalt dan ook samen met de hoogste kategorien van de andere variabelen de horizontale onderscheidingen.

We resumeren nu een aantal eigenschappen van HOMALS oplossingen die voor interpretatiedoeleinden van belang zijn.

1. Individuen en kategorieën worden als punten in één gemeenschappelijke euclidiese ruimte afgebeeld.
2. Ofwel de individu-punten liggen in het zwaartepunt van de kategorie-punten waar zij bij horen, ofwel de kategorie-punten liggen in het zwaartepunt van de individu-punten die er bij horen (zwaartepunt-principe). De keuze tussen beide komt overeen met de keuze tussen alternatieve normalisatiemethoden. De spreiding van de projekties op de assen van de punten die niet genormaliseerd zijn, komt overeen met de grootte van de eigenwaarden.
3. De koordinaatwaarden van individuen en kategorieën op de eerste p_1 assen van een p_2-dimensionale oplossing zijn gelijk aan die van een p_1-dimensionale oplossing (genestheid).
4. Variabelen definieren in de indikator matrix een partitie van individuen, in de oplossing komen zij overeen met een opsplitsing van individu-punten in (ev. elkaar overlappende) puntenwolken.

De volgende vuistregels kunnen ook van belang zijn:

5. Naarmate de puntenwolken die behoren bij de kategorie-punten van een variabele verder uit elkaar liggen en minder overlappen diskrimineert die variabele beter, hij is meer homogeen, en draagt meer bij tot de eigenwaarde.
6. Individuen met overeenkomstige antwoordprofielen zullen dicht

bij elkaar in de ruimte terechtkomen. Een groep individuen waarvoor dit geldt vormt een homogene groep.

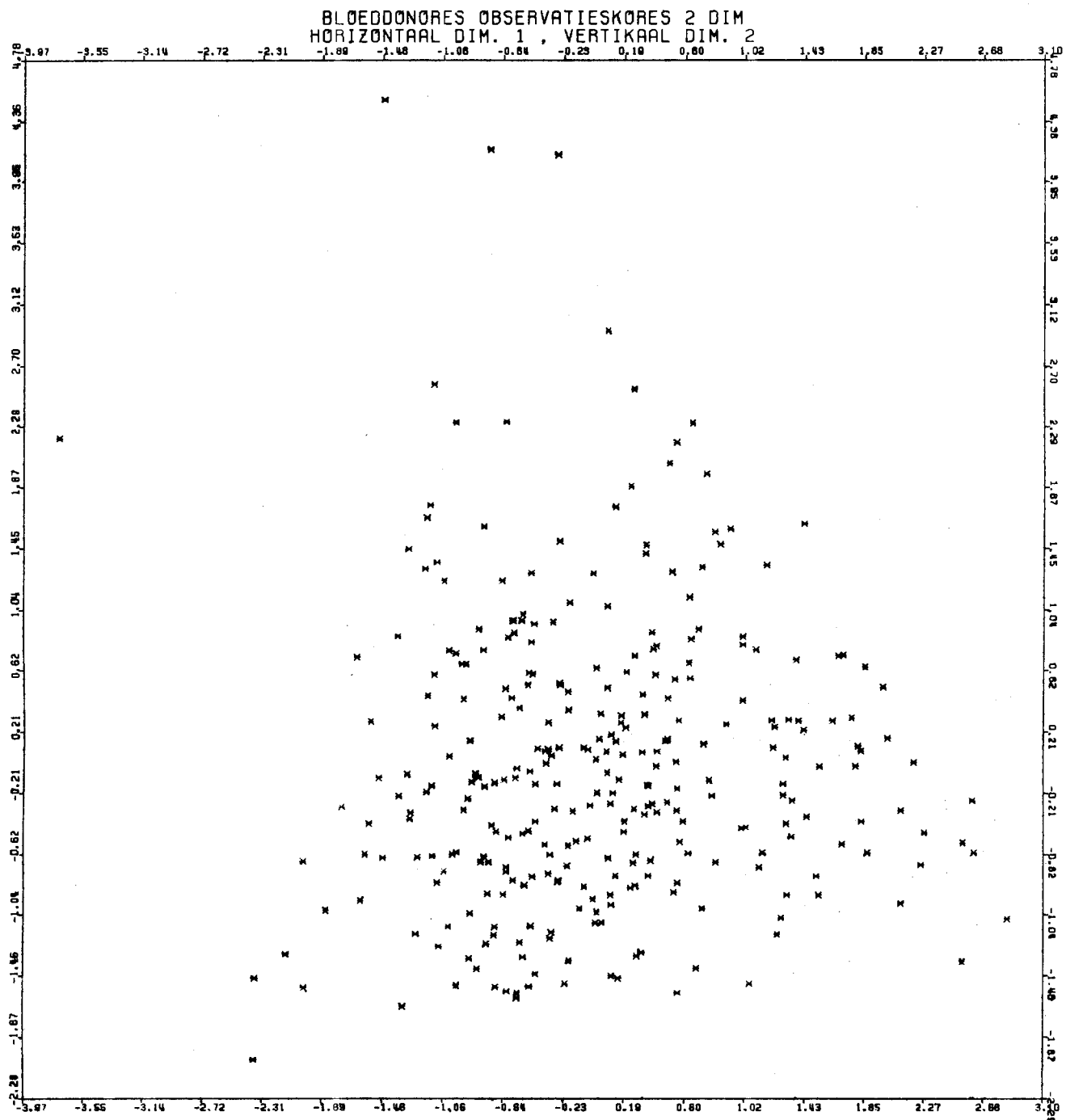
7. Individuen met in hoge mate verschillende antwoordprofielen zullen ver van elkaar in de ruimte terechtkomen. Twee of meer groepen waarvoor dit geldt, worden goed gediskrimineerd.
8. Individuen met antwoordprofielen die veel gemeen hebben met de antwoordprofielen van de meeste andere individuen zijn representatief en worden centraal in de ruimte afgebeeld.
9. Individuen met antwoordprofielen die weinig gemeen hebben met de antwoordprofielen van de meeste andere individuen zijn uniek en worden perifeer in de ruimte afgebeeld.
10. Naarmate de marginale frekwentie van een kategorie hoger is, zal hij meer centraal worden afgebeeld; naarmate deze lager is, meer perifeer.

Tot besluit van deze sektie bekijken we twee 2-dimensionale HOMALS oplossingen voor gegevens die we in het vorige hoofdstuk al tegenkwamen: die van het Bloeddonoren onderzoek bij de Haagse Bloedbank en de Abortus/Sexuele vrijheid vragenlijst.

Bloedbank.

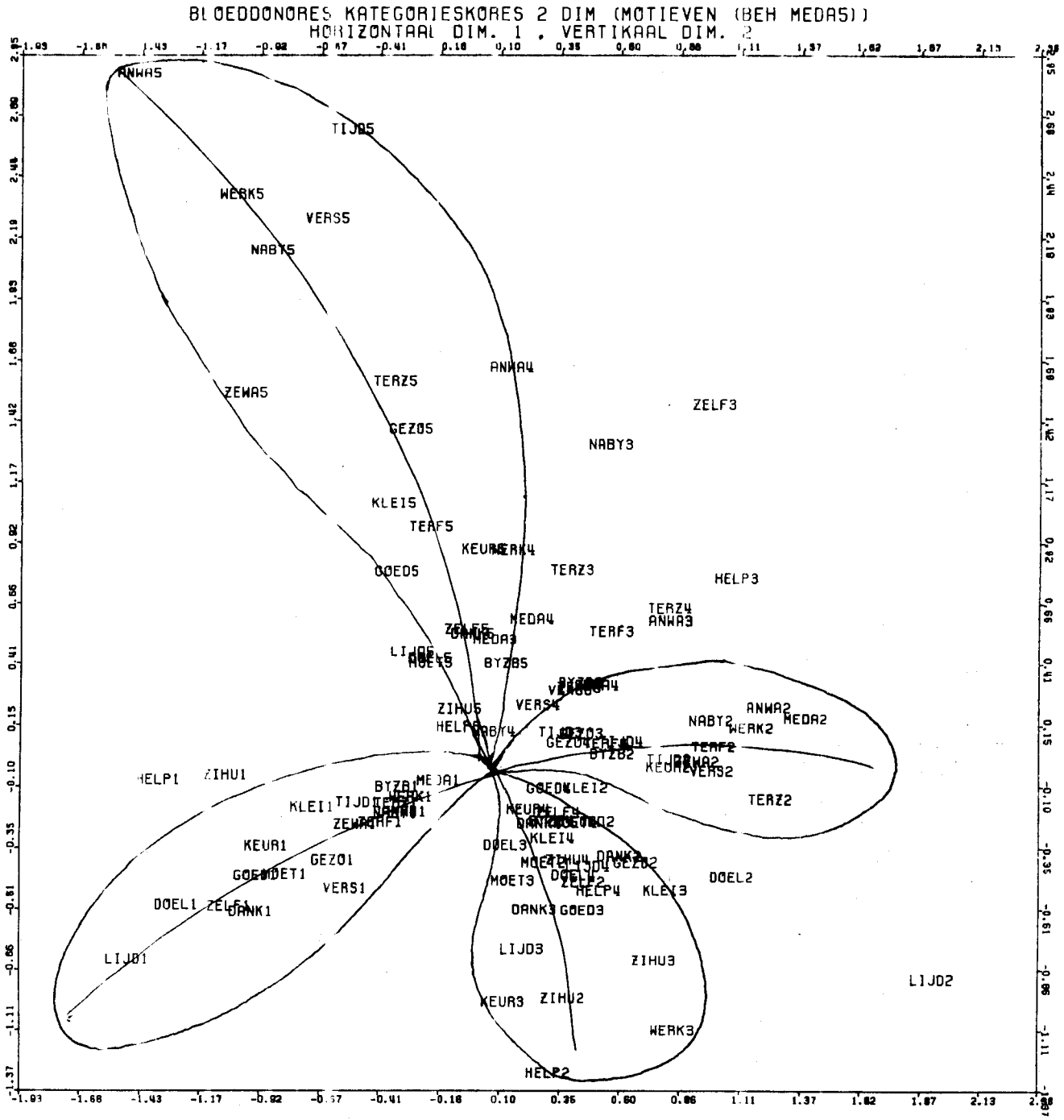
De 1-dimensionale oplossing voor de Bloedbankgegevens vertoonde niet de monotoniciteit die we van dergelijke Likert-type items verwachten. Bij veel variabelen werden de kategorien 1 en 5 samen geschaald versus 2. Nu zijn de variabelen erg scheef verdeeld, en er zal zeker een effekt van "sociale wenselijkheid" meespelen, zodat op sommige items iedereen 'speelt helemaal geen rol' zegt en op andere iedereen 'speelt een grote rol', maar dit verklaart nog niet waarom 1 en 5 per variabele zo dicht bij elkaar liggen. Het is bovendien nogal onwaarschijnlijk, dat opleiding en beroep zo'n geringe rol spelen (niet soms?).

In figuur 3.6 zijn de individu punten geplot van de 2-dimensionale oplossing. Hoewel zeker bijzonder van vorm, is deze plot op zich weinig informatief; hij laat eigenlijk voornamelijk zien, dat de gegevens niet 1-dimensionaal zijn en dat er geen duidelijke klusters zijn. De plot van de kategorie punten in figuur 3.7 is duidelijk beter te interpreteren; de kategorien vormen een klaverblad.



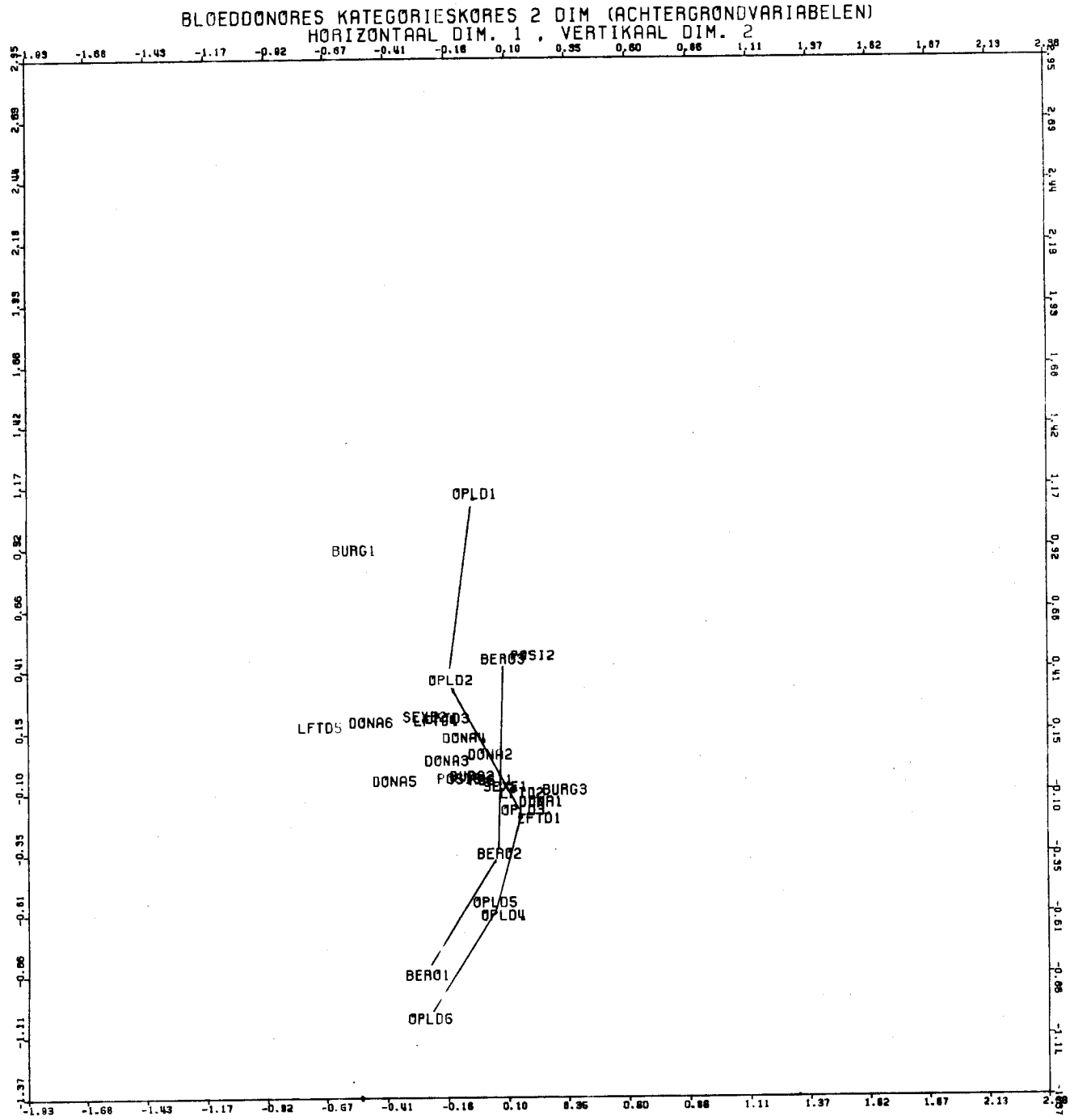
figuur 3.6. Individupunten Bloedbankgegevens.

De blaadjes van dit klavertje-4 worden gevormd door 'zuivere types'; het feit dat de individuen-plot geen overeenkomstige klusters vertoont betekent, dat er weinig individuen 'puur' zijn, de meesten zijn een mengvorm. Links onder zitten de cynici, die van de meeste motieven absoluut ontkennen dat ze een rol spelen (vooral LIJD, DOEL, ZELF, DANK, GOED en KEUR zijn onbelangrijk). Meer naar rechts komen de nobelen, voornamelijk gekenmerkt door 'weet-niet' antwoorden (3),



figuur 3.7. Kategoriepunten Bloedbankgegevens (motieven).

die echter wel LIJD, DOEL, HELP en KLEI relatief hoog in hun vaandel hebben (4). Het derde blaadje is de moeilijkste groep: veel 2-en, eigenlijk wordt alleen specifiek enig belang gehecht aan ZEWA, TERF en TIJD (4). Omdat de tweeën in deze oplossing zo gekontrasteerd worden met de enen, zijn we geneigd deze kategorie te herinterpreteren als 'kan niet ontkennen dat het een rol speelt', er van uit gaand dat deze mensen een bepaald idee hebben van wat sociaal (on)-



figuur 3.8. Kategoriepunten bloedbankgegevens (achtergrondvariabelen).

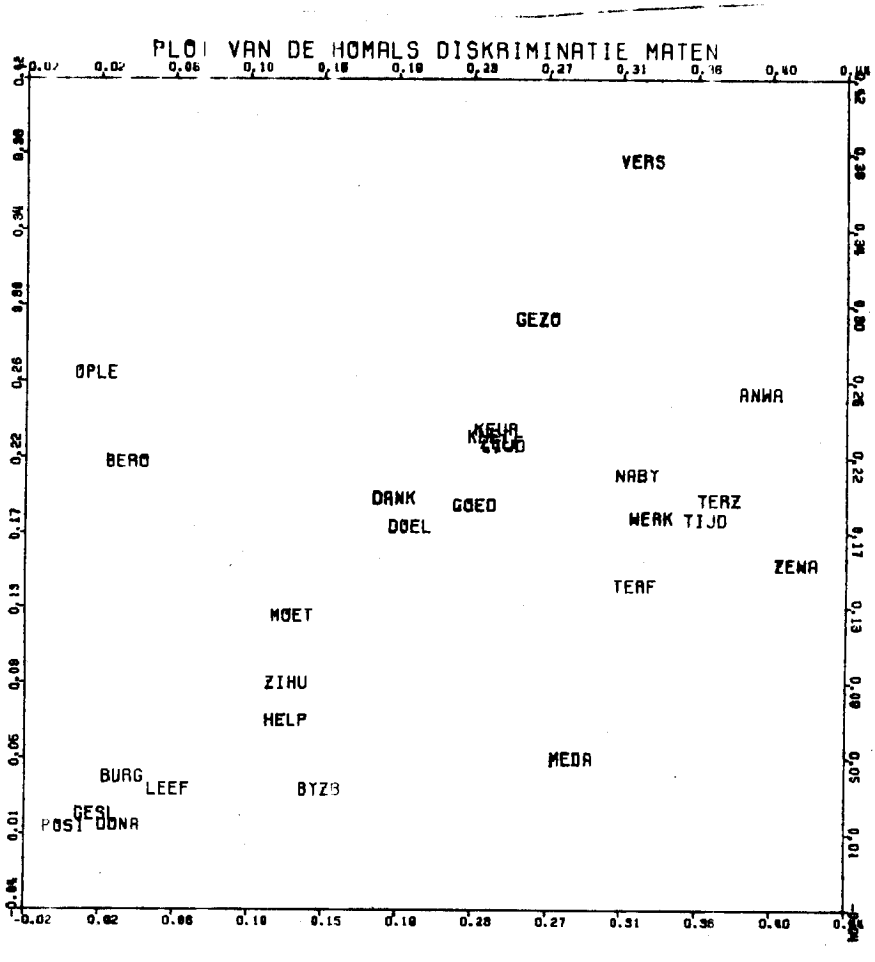
wenselijk is. Wat kunnen ze niet ontkennen? MEDA, TERZ, ANWA en NABY, en pragmatiese motieven zoals WERK, VERS en KEUR. We noemen deze groep de heimelijken, en zij delen met de nobelen 2-en voor LIJD en DOEL. Tenslotte via een stel weet-niet antwoorden naar het laatste blad. Dit loopt van enthousiastelingen (centraal) naar dwangmatigen (perifeer). De eersten geven veel 4 en 5 antwoorden, en de laatsten zeggen bovendien, dat een grote rol speelt: ANWA,

WERK, TIJD, VERS, NABY, ZEWA, TERZ en GEZO. Deze mensen kunnen niet zonder bloed-geven. Ergens ver links boven zit ook nog de man die als enige MEDA5 heeft aangekruist (zie figuur 3.6); hij is uit plot 3.7 weggelaten.

Binnen de diverse blaadjes zijn de motieven niet op dezelfde manier gerangordend, en enige 'motievenstruktuur' kunnen we dan ook niet ontdekken. Het moet ons hier trouwens wel even van het hart, dat de gekozen onderzoeksopzet niet bijster elegant was: het is een mengelmoes van motieven, er zijn geen kriteria aangegeven om ze op te vergelijken, zelfs op het attribuut 'belangrijk' hoeven ze niet ekspliciet vergeleken te worden (de proefpersonen hoeven nergens te kiezen), en ze zijn naïef-positief geformuleerd (er zijn mensen die bijvoorbeeld bloed geven hoewel (zij denken dat) het slecht voor hun gezondheid is, etc.). Het is zo maar een balletje opgooien en de rest aan de proefpersonen en de komputer over willen laten.

De achtergrondsvariabelen zijn apart geplot in figuur 3.8. Beroep en Opleiding doen het meest, en wel in vertikale richting. Dus, mensen met een 'laag' opleidingsnivo en een 'lager' beroep zijn enthousiast, mensen met een 'hoog' opleidingsnivo en een 'hoger' beroep zijn cynies of nobel. De plot van figuur 3.8 is op dezelfde schaal als die van figuur 3.7, wat nog eens laat zien dat behalve de twee genoemde de achtergrondvariabelen weinig diskrimineren.

Als we deze HOMALS oplossing met de VARIMAX Faktor Analyse van de onderzoekers (zie tabel 2.17 in hoofdstuk 2.4) willen vergelijken, is het 't handigst dit te doen aan de hand van een plot van de diskriminatie maten (figuur 3.9), die voor elke variabele aangeven, in hoeverre deze bijdraagt aan de eigenwaarde (in hoeverre zijn kategorien spreiden in elk van de richtingen). Van de variabelen die op de eerste faktor laden diskrimineert in HOMALS eigenlijk alleen LIJD behoorlijk, DANK en DOEL eventueel ook nog, ZIHU en HELP eigenlijk niet. Het zijn trouwens de variabelen die het meest scheef naar links zijn (veel 5-en). De variabelen van de tweede faktor, ANWA, WERK, ZEWA en MEDA diskrimineren in HOMALS vooral langs de eerste as; deze variabelen zijn allemaal scheef naar rechts (veel 1-en). De derde faktor is VERS en GEZO, de twee best diskriminerende HOMALS variabelen. Het zal duidelijk zijn, dat de twee analyses niet bepaald tot gelijkluidende konklusies leiden. Faktor Analyse geeft een zeer

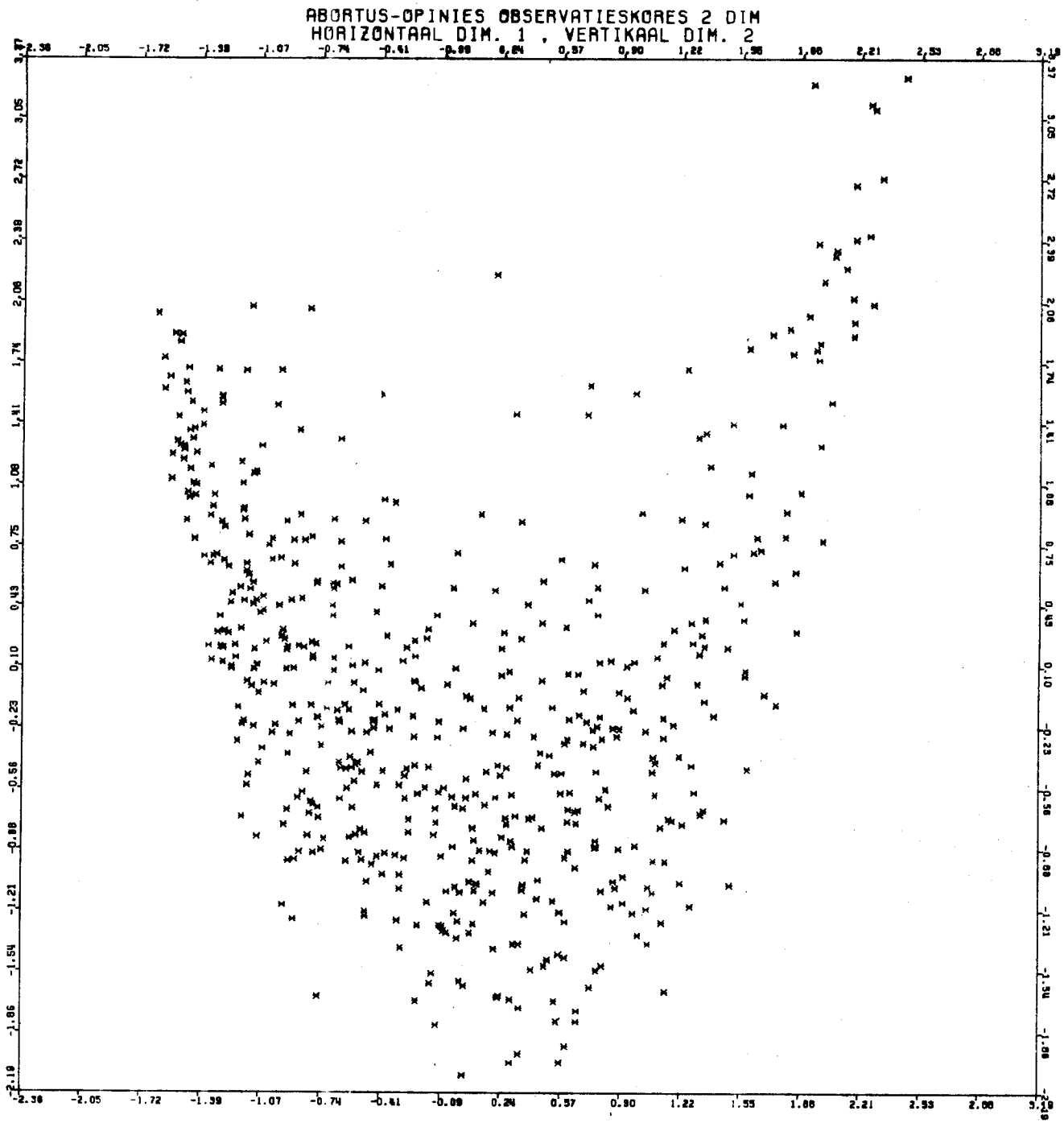


figuur 3.9. Diskriminatiematen Bloedbank; $\lambda_1 = .21$, $\lambda_2 = .16$.

grove trend, HOMALS benadrukt typiese groepen. Faktor Analyse spreidt met de a priori opvattingen van de onderzoeker mee, HOMALS tikt hem op de vingers. De presentatie van de Faktor Analyse verdoezelt dat er mensen zijn die LIJD, DOEL, DANK en HELP onbelangrijk vinden, HOMALS laat zien waar de scheidslijnen tussen belangrijk en onbelangrijk liggen.

Abortus/Sexuele Vrijheid.

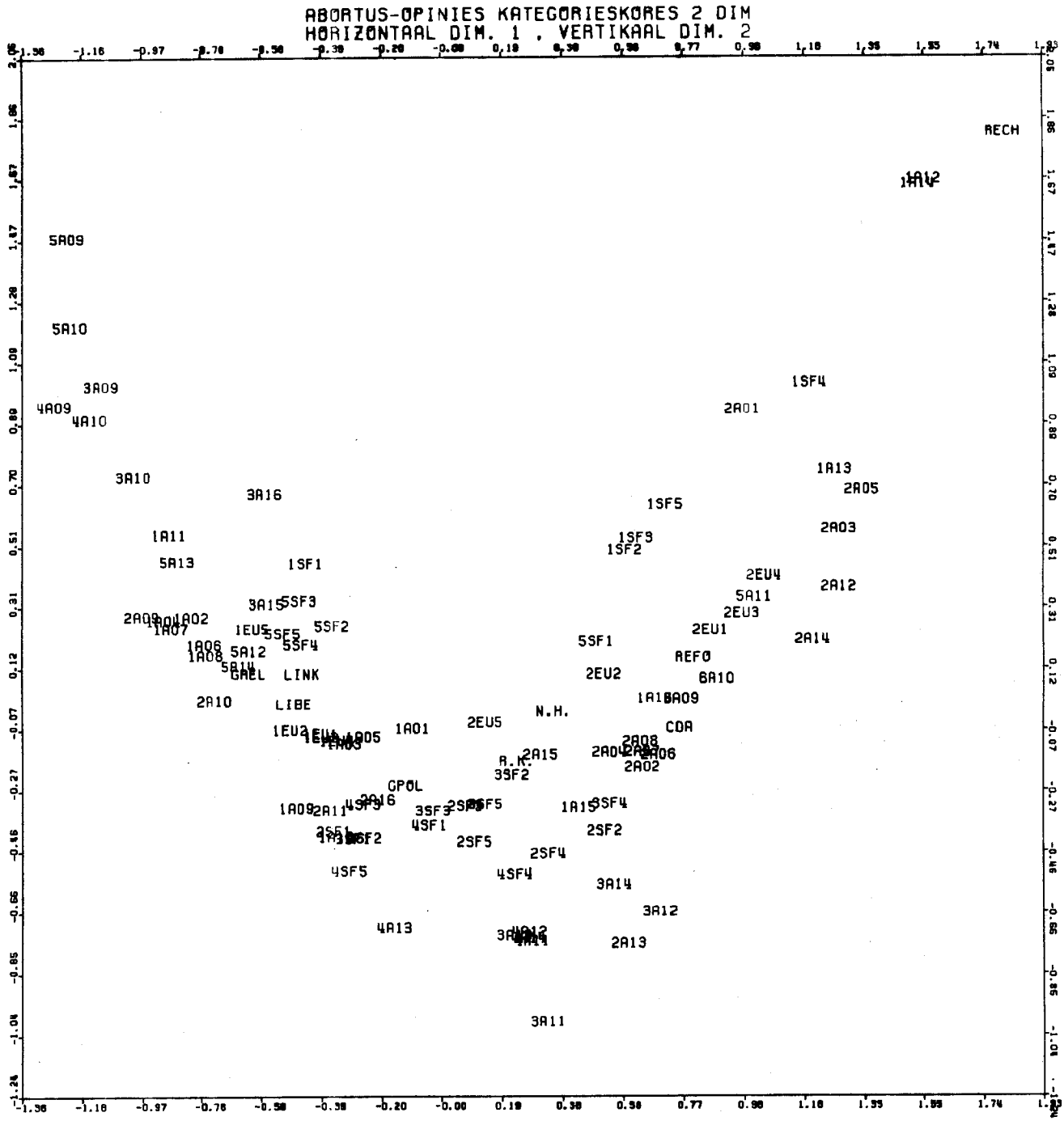
Ook dit voorbeeld zijn we eerder tegengekomen, in hoofdstuk 2.4.3; één-dimensionale HOMALS oplossing gaf mooie monotone transformaties van de kategorie-koderingen, en het was duidelijk dat we met een liberaal-reaktionair kontinuum te maken hadden. Wat er gebeurt wanneer men in een dergelijke situatie toch twee dimensies berekent is afgebeeld in figuur 3.10. Een dergelijke plot is dus een hoefijzer uit het dagelijks leven. Zonder dit precies te hoeven uitrekenen zien we, dat de regressie van x_2 op x_1 kwadraties is. In hoofdstuk 2.1 zijn we al een aantal theoretiese redenen voor dit verschijnsel



figuur 3.10. Individupunten voor Abortus/Sexuele Vrijheid data.

tegengekomen, zodat we hier alleen nog wat praktiese dingen aanstippen:

- het hoefijzer is in het algemeen beter te zien in de gezamenlijke plot van de kategoriepunten (zie figuur 3.11), omdat dit zwaartepunten zijn;
- variabelen die niet goed meedoen komen 'tussen de benen' terecht (vgl. de SF variabelen); variabelen die helemaal niet meedoen



figuur 3.11. Kategoriepunten voor Abortus/Sexuele Vrijheid data.

komen rond de oorsprong terecht (dit is nog het meest het geval voor EU5);

- het is soms raadzaam er op te letten, hoe 'ver' de kategorien plot in de individuen plot steekt (bijv. door ze op een identieke schaal te plotten); hoe verder, hoe beter ze diskrimineren, hoe 'sterker' de schaal;
- als het hoefijzer te dik is, probeer dan een inhoudelijke selektie van variabelen apart (hier bijv. alleen de abortus variabelen).

3.2. Analyse des Correspondances.

3.2.0. Inleiding.

Onder Analyse des Correspondances verstaan we hier een techniek om de rijen en kolommen van een willekeurige matrix met niet-negtieve elementen in de euclidiese ruimte weer te geven. De techniek is ontleend aan Benzécri (1973). De ruimtelijke weergave is zodanig dat identieke rijen (kolommen) op het zelfde punt afgebeeld worden, en naarmate rijen (kolommen) meer met elkaar korresponderen ze dichter bij elkaar komen te liggen.

Analyse des Correspondances is symmetries (hoewel niet-symmetriese generalisaties bestaan), d.w.z. dat de rijen van de matrix hetzelfde weergegeven worden als de kolommen van z'n getransponeerde, en vice-versa. Hieruit volgt, dat als de matrix symmetries is, we voor rijen en kolommen dezelfde weergave vinden.

Speciale matrices waarop we Analyse des Correspondances kunnen toepassen zijn: indikator matrices, twee-dimensionale kruistabellen en matrices van bivariate marginalen. We zullen aantonen dat HOMALS een speciaal geval van Analyse des Correspondances is. Het programma ANACOR implementeert naast bovengenoemde toepassingen bovendien de theorie die in hoofdstuk 4.1 zal worden uiteengezet, namelijk de mogelijkheid om betrouwbaarheidsellipsen om de punten te schatten.

3.2.1. Algemeen model.

We gaan uit van een matrix G, van orde nxm, met niet-negatieve elementen. Ter vereenvoudiging van het model nemen we aan, dat geen enkele rij en geen enkele kolom van G uit louter nullen bestaat. We definieren de vektoren

$$e \triangleq Gu$$
$$d \triangleq G'u$$

Alle elementen van d en e zijn positief. We gebruiken N voor de som van alle elementen van G (dus ook die van d en e):

$$N \triangleq u'Gu = u'e = d'u$$

De afstand tussen de rijen i en k van G definieren we als

$$\delta_{ik}^2 \triangleq N \sum_{j=1}^{m} \left(\frac{g_{ij}}{e_i} - \frac{g_{kj}}{e_k}\right)^2 / d_j$$

en de afstand tussen de kolommen j en ℓ

$$\varepsilon_{j\ell}^2 \triangleq N \sum_{i=1}^{n} \left(\frac{g_{ij}}{d_j} - \frac{g_{k\ell}}{d_\ell}\right)^2 / e_i$$

Deze afstandsdefinitie heeft de volgende achtergrond. Veronderstel dat we op een eindige verzameling $O = \{o_1, o_2, \ldots, o_m\}$ een kansverdeling P op O hebben met $p_j > 0$ en $\Sigma p_j = 1$. Met O en P vast is dan de <u>Benzécri-afstand</u> $d_p(Q,R)$ tussen twee kansverdelingen gedefinieerd als

$$d_p(Q,R) \triangleq \{ \sum_{j=1}^{m} (q_j - r_j)^2 / p_j \}^{\frac{1}{2}} .$$

Als we nu een steekproef ter grootte n uit O trekken, volgens kansverdeling P, en $\underline{f}_j$ is de proportie trekkingen van o_j , dan geldt

$$n\, d_p^2(\underline{F},P) \to \underline{\chi}_{m-1}^2 \quad (n\to\infty)$$

Hier is $\underline{F}$ een schatter van P. Hebben we twee onafhankelijke schatters $\underline{F}_1$ en $\underline{F}_2$ van P, beide uit een steekproef van grootte n, dan

$$n\, d_p^2(\underline{F}_1,\underline{F}_2) \to 2\, \underline{\chi}_{m-1}^2 \quad (n\to\infty)$$

We hebben dus te maken met afstanden tussen kansverdelingen. Beschouwen we iedere rij van G als schatter van d/n, dan volgt de definitie van δ direkt (en, analoog, die van ε).

We bekijken nu eerst alleen de rijen; we willen ze weergeven als vektoren x_i in een euclidiese ruimte en wel zo dat

$$(x_i - x_k)'(x_i - x_k) = \delta_{ik}^2 .$$

Hiertoe schrijven we δ_{ik}^2 in matrixvorm:

$$\delta_{ik}^2 = N \sum_{j=1}^{m} \frac{g_{ij}^2}{d_j e_i^2} - 2N \sum_{j=1}^{m} \frac{g_{ij}g_{kj}}{d_j e_i e_k} + N \sum_{j=1}^{m} \frac{g_{kj}^2}{d_j e_k^2}$$

$$= N\,(E^{-1}GD^{-1}G'E^{-1})_{ii} - 2N\,(E^{-1}GD^{-1}G'E^{-1})_{ik} + N\,(E^{-1}GD^{-1}G'E^{-1})_{kk}$$

We noteren de i'de kolom uit de eenheidsmatrix met u_i en krijgen:

$$\delta^2_{ik} = N\,(u_i - u_k)'E^{-1}GD^{-1}G'E^{-1}(u_i - u_k)$$

We kunnen voor x_i nu nemen $N^{\frac{1}{2}}\,D^{-\frac{1}{2}}G'E^{-1}u_i$ voor alle i, dan geldt het gewenste resultaat, we hebben de rijen voorgesteld als n vektoren in de $\mathbb{R}^m$.

Vooral als m>n is dit een inefficiente manier van representeren. We gaan nu zoeken naar een representatie van de rijen als vektoren x_i in een vektorruimte van zo laag mogelijke dimensie, maar nog steeds met de eis $(x_i-x_k)'(x_i-x_k)=\delta^2_{ik}$ voor alle paren i,k.

In een vektorruimte van dimensie q kunnen we de x_i samennemen in een matrix $X_q=(x_1,\ldots,x_n)'$ van nxq, en dan is $x_i=X'_q u_i$, en dus is

$$(x_i - x_k)'(x_i - x_k) = (u_i - u_k)'X_qX'_q(u_i - u_k)$$

De eis wordt nu

$$(u_i - u_k)'X_qX'_q(u_i - u_k) = N\,(u_i - u_k)'E^{-1}GD^{-1}G'E^{-1}(u_i - u_k)$$

We kunnen hieruit afleiden, dat om q minimaal te maken moet gelden

$$X_qX'_q = N\,E^{-1}GD^{-1}G'E^{-1} - uu'.$$

Dit is een toegestane representatie, immers $u'(u_i-u_k)=0$; en het is de representatie die er voor zorgt dat q=rang(G)-1, en dit is de kleinste q die mogelijk is. Geometries geinterpreteerd hebben we van alle x_i een zelfde vektor afgetrokken, wat de Benzécri afstanden toch niet beinvloed.

We zouden nu de eigenwaarden-ontbinding van $N\,E^{-1}GD^{-1}G'E^{-1} - uu'$ kunnen bepalen, en hieruit X_q. Wat we echter ook kunnen doen, is uitgaan van een andere schrijfwijze van $X_qX'_q$, nl:

$$X_qX'_q = N\,E^{-\frac{1}{2}}(E^{-\frac{1}{2}}GD^{-1}G'E^{-\frac{1}{2}} - \frac{E^{\frac{1}{2}}uu'E^{\frac{1}{2}}}{N})E^{-\frac{1}{2}},$$

en uitgaan van de eigenwaarden-ontbinding van

$$H_1 \triangleq E^{-\frac{1}{2}}GD^{-1}G'E^{-\frac{1}{2}} - \frac{E^{\frac{1}{2}}uu'E^{\frac{1}{2}}}{N}$$

Om te zien, waarom we dit doen, gaan we nu ook de kolommen in het verhaal betrekken. Omdat alle afleidingen volkomen analoog zijn, geven we allen de resultaten. Met

$$y_j = N^{\frac{1}{2}} E^{-\frac{1}{2}} G D^{-1} u_j$$

is

$$(y_j - y_\ell)'(y_j - y_\ell) = \varepsilon^2_{j\ell}$$

en zijn de kolommen voorgesteld als m vektoren in $\mathbb{R}^n$. Willen we schrijven $y_j = Y_q' u_j$, dan vinden we

$$Y_q Y_q' = N D^{-\frac{1}{2}} H_2 D^{-\frac{1}{2}}$$

met

$$H_2 \triangleq D^{-\frac{1}{2}} G' E^{-1} G D^{-\frac{1}{2}} - \frac{D^{\frac{1}{2}} uu' D^{\frac{1}{2}}}{N}$$

Nu voeren we in de matrix

$$H_3 \triangleq E^{-\frac{1}{2}} G D^{-\frac{1}{2}} - \frac{E^{\frac{1}{2}} uu' D^{\frac{1}{2}}}{N} \qquad (1a)$$

Simpel rekenwerk levert op dat $H_1 = H_3 H_3'$ en $H_2 = H_3' H_3$, en dus hebben H_1 en H_2 dezelfde eigenwaarden. De singuliere waarden dekompositie

$$H_3 = K \Lambda L' \quad \text{met } K'K = I,\ L'L = I \text{ en } \Lambda \text{ diagonaal, geordend} \qquad (1b)$$

levert gelijk op $H_1 = K \Lambda K'$ en $H_2 = L \Lambda L'$, en we kunnen dus schrijven

$$X_q X_q' = N E^{-\frac{1}{2}} K \Lambda^2 K' E^{-\frac{1}{2}}$$

zodat de rijen van

$$X_q = N^{\frac{1}{2}} E^{-\frac{1}{2}} K \Lambda$$

de gezochte representatie in $\mathbb{R}^q$ leveren. Het zelfde geldt, naar analogie, voor de rijen van

$$Y_q = N^{\frac{1}{2}} D^{-\frac{1}{2}} L \Lambda .$$

Nu is een representatie in $\mathbb{R}^p$ $(p<q)$ ook gemakkelijk te vinden, nl door alleen de p grootste eigenwaarden en de bijbehorende kolommen van K en L (eigenvektoren) te nemen. Dit betekent dat $(x_i - x_k)'(x_i - x_k)$ een benadering is van ξ^2_{ik} en $(y_j - y_\ell)'(y_j - y_\ell)$ van $\varepsilon^2_{j\ell}$, en wel op de zelfde manier, omdat de dezelfde eigenwaarden weglaten. Dit is een van de redenen waarom we uitgingen van de

eigenwaarden-ontbinding van H_1 i.p.v. $E^{-\frac{1}{2}}H_1E^{-\frac{1}{2}}$? Een andere reden hiervoor is, dat we X_q en Y_q in elkaar kunnen uitdrukken, waarvoor we nog een paar eigenschappen van X_q, Y_q, K en L nodig hebben, i.e.

$$H_3D^{\frac{1}{2}}u = 0$$
$$H_3'E^{\frac{1}{2}}u = 0$$
$$H_1E^{\frac{1}{2}}u = 0$$
$$H_2D^{\frac{1}{2}}u = 0$$
$$K'E^{\frac{1}{2}}u = 0 \qquad (2a)$$
$$L'D^{\frac{1}{2}}u = 0 \qquad (2b)$$
$$X_q'u = 0 \qquad (3a)$$
$$Y_q'u = 0 \qquad (3b)$$

wat betekent dat de x_i en de y_j gecentreerd zijn. Verder geldt nu

$$Y_q = D^{-1}G'X_q\Lambda^{-1} \qquad (4a)$$
$$X_q = E^{-1}G\,Y_q\Lambda^{-1} \qquad (4b)$$

en dus, als we de rij- of kolomkwantifikaties kennen, kunnen we de andere direkt via (4a,b) afleiden. Dit is één manier om X_q en Y_q ten opzichte van elkaar te schalen. De operatoren $D^{-1}G'$ en $E^{-1}G$ nemen gewogen gemiddelden van de rijen van X_q en Y_q, en als we daarom nemen, bijvoorbeeld

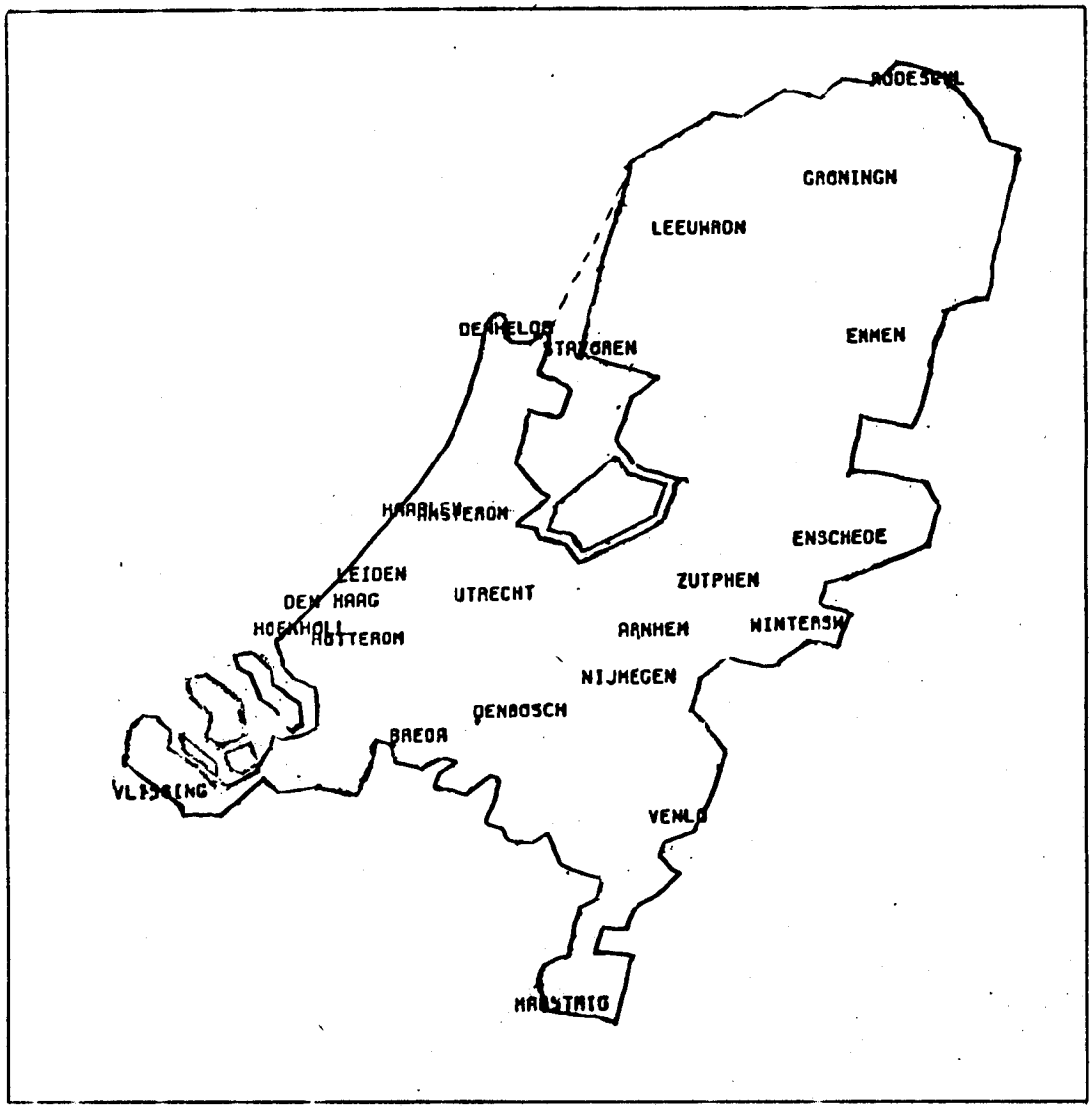
$$X = N^{\frac{1}{2}}\,E^{-\frac{1}{2}}K\Lambda$$
$$Y = N^{\frac{1}{2}}\,D^{-\frac{1}{2}}L$$

dan gelden andere 'formules de transition':

$$Y = D^{-1}G'X\Lambda^2$$
$$X = E^{-1}G\,Y$$

We zorgen er nu dus voor, dat de rij-punten gewogen kombinaties zijn van kolom-punten, dit is weer het 'principe barycentrique' wat we al eerder tegenkwamen en bijv. bij de korrespondentie-analyse van MVA boeken zo gebruikt hebben.

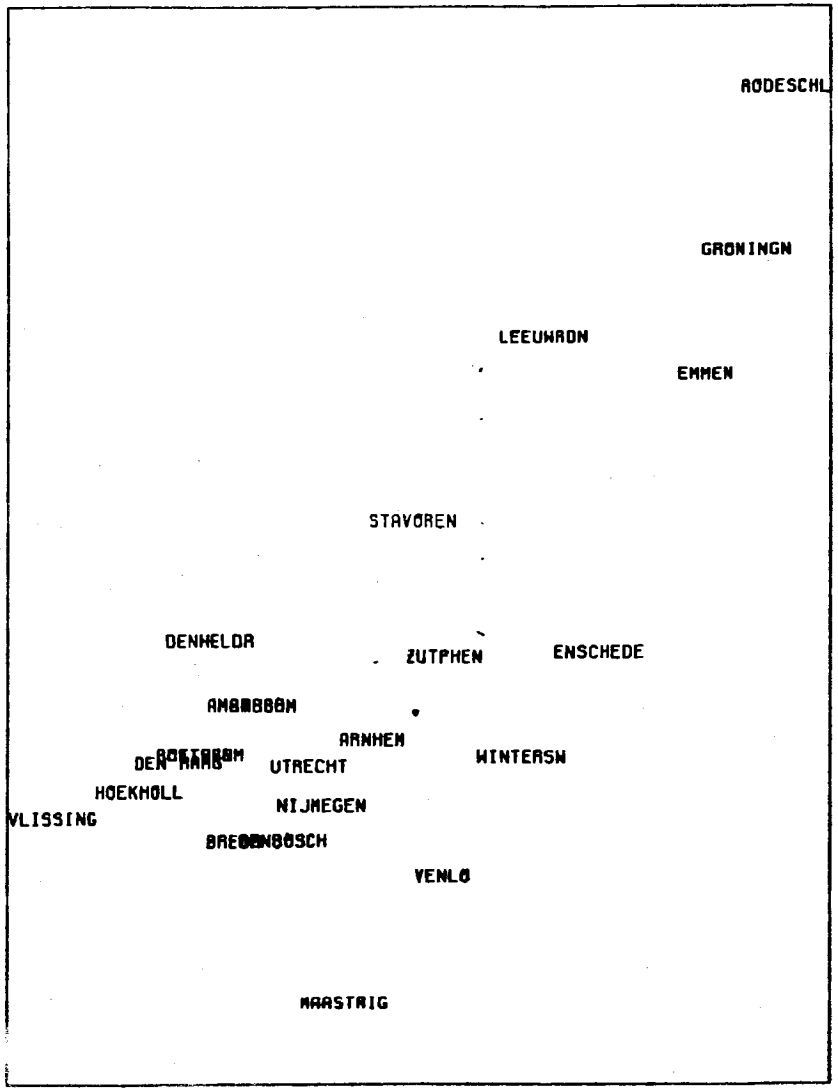
We geven nu een voorbeeld van een toepassing op een symmetriese matrix. Daartoe zijn de afstanden tussen 23 plaatsen in Nederland genomen, op drie manieren: vogelvluchtafstanden, vogelvluchtafstan-



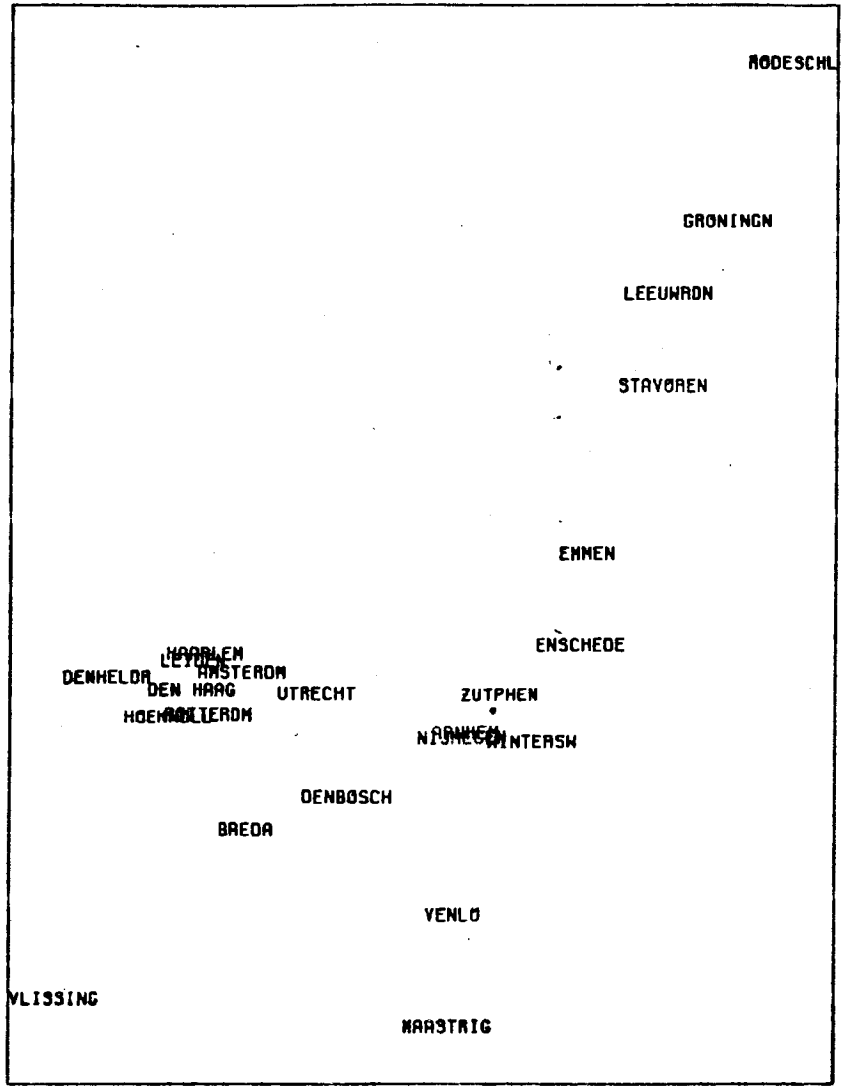
figuur 3.12. Vogelvlucht-afstanden.

den gedichotomiseerd op de mediaanafstand en de afstanden via de kortste railverbinding. Om ze geschikt te maken als input voor het programma ANACOR zijn alle afstanden per matrix van hun grootste element afgetrokken.

In het geval van vogelvlucht-afstanden verwachten we dat het programma de kaart van Nederland natuurgetrouw reproduceert. Dat dit inderdaad het geval is, zien we in figuur 3.12. We kunnen dit resultaat nu als uitgangspunt nemen voor de twee andere gevallen. De gedichotomiseerde afstanden geven een aanwijzing voor de robuustheid van ANACOR. De rekonstruktie (zie figuur 3.13) blijft opmerkelijk goed; wel is er natuurlijk sprake van enige klustering in de Randstad en omstreken. Het resultaat voor de rail-afstanden geeft



figuur 3.13. Vogelvlucht-afstanden gedichotomiseerd op de mediaan.



figuur 3.14. Rail-afstanden.

weer wat we al wisten: er is geen railverbinding over de afsluitdijk. Dit maakt dat profielen in de Randstad en profielen in het Noorden meer op elkaar gaan lijken.

Alle kaartjes zijn overigens 45^0 linksom geroteerd t.o.v. hun principale assen stand. Bij de oorspronkelijke stand horen de volgende eigenwaarden:

	3.12	3.13	3.14
1^e as	.0641	.4155	.0589
2^e as	.0238	.2590	.0143
3^e as	.0020	.0909	.0059

Voor een twee-dimensionaal probleem als het onderhavige verwachten we inderdaad een scherpe daling van de derde eigenwaarde t.o.v. de eerste twee. Dat in het eerste voorbeeld de derde eigenwaarde niet identiek nul is komt overeen met de lichte vertekening die figuur 3.12 te zien geeft.

3.2.2. Toepassing op indikator matrixen.

We nemen nu voor G een $n \times \Sigma k_j$ indikator super matrix, behorende bij n observaties van m nominale variabelen met $k_1,\ldots,k_m$ kategorieën. De kolommen van G representeren kategorieën, de rijen individuen. Elk individu skoort voor elke variabel in een bepaalde kategorie, dus uit de eerder gegeven definitie van indikator matrices volgt dat in iedere rij van G precies m enen staan. Dus

$$e = Gu = mu$$

Hieruit volgt

$$E = mI \qquad \text{en} \qquad N = u'Gu = m\,u'u = mn$$

De rij-totalen zijn allemaal m, er zijn n rijen, dus de som over alle elementen van G is mn. De vektor van kolom-totalen d splitsen we als volgt:

$$d = (d_1,\ldots,d_m)' \quad \text{met } d_j = (d_1^j,\ldots,d_{k_j}^j)' \quad j = 1,\ldots,m$$

Hierin is d_r^j het aantal enen in de r-de kolom van G_j (de j-de submatrix van G), en dus het totaal aantal individuen dat in

kategorie r van variabele j valt. De reeds eerder ingevoerde matrix D kunnen we overeenkomstig opsplitsen in diagonale submatrixen D_j met elementen $d_1^j,\ldots,d_{kj}^j$. De representatie in $\mathbb{R}^p$ volgt uit (1a,b), welke we nu kunnen schrijven als

$$m^{-\frac{1}{2}} GD^{-\frac{1}{2}} - \frac{m^{\frac{1}{2}}}{mn} uu'D^{\frac{1}{2}} = K\Lambda L' \quad \text{met } K'K=I,\ L'L=I \text{ en } \Lambda \text{ diagonaal, geordend}$$

Getransponeerd en met zich zelf vermenigvuldigd levert dit

$$m^{-1} D^{-\frac{1}{2}}G'GD^{-\frac{1}{2}} - \frac{1}{mn} D^{\frac{1}{2}}uu'D^{\frac{1}{2}} = L\Lambda^2L' \qquad (5)$$

Dit kunnen we schrijven in de vorm die we al eerder tegenkwamen; met $L'L=I$ en $u'D^{\frac{1}{2}}L=0$ (vgl 2b) vinden we

$$G'GD^{-\frac{1}{2}}L = m\, D^{\frac{1}{2}}L\Lambda^2$$

Noem nu

$$Y \triangleq m^{\frac{1}{2}}n^{\frac{1}{2}} D^{-\frac{1}{2}}L \qquad \text{zodat} \quad Y'DY=mnL'L=mnI \text{ en } u'DY=m^{\frac{1}{2}}n^{\frac{1}{2}}u'D^{\frac{1}{2}}L=0$$

en substitueer $C=G'G$, dan is

$$CY = m\, DY\Lambda^2 \qquad (6)$$

en we zijn terug bij HOMALS. In 3.2.1. zagen we dat $Y_q=N^{\frac{1}{2}}D^{-\frac{1}{2}}L\Lambda$, in ons geval is $N=nm$, dus

$$Y_q = m^{\frac{1}{2}}n^{\frac{1}{2}} D^{-\frac{1}{2}}L\Lambda = Y\Lambda \qquad (7)$$

We bekijken Y nog nader. Het is een matrix van afmetingen $\Sigma k_j \times q$. Hij kan geschreven worden als $Y=(Y_1,\ldots,Y_m)'$ met Y_j afmetingen $k_j \times q$ en y_{rs}^j is de kwantifikatie van de r-de kategorie van variabele j in dimensie s. Via (7) zien we het verband tussen de kategorie-kwantifikaties Y en de Benzécri representaties Y_q. Analoog aan (7) nemen we als individu skores

$$X \triangleq X_q\Lambda^{-1} \qquad (8)$$

Omdat we met (4b) het verband tussen X_q en Y_q kennen, kunnen we ook het verband tussen X en Y vinden:

$$X = X_q\Lambda^{-1} = E^{-1}GY_q\Lambda^{-2} = m^{-1}\, GY\Lambda^{-1} = K \qquad (9)$$

We hebben nu een q-dimensionale representatie. Nemen we alleen de eerste p singuliere waarden en vektoren, dan hebben we HOMALS

in p dimensies. Dit resultaat is leuk, maar niet nuttig, omdat er een HOMALS algorithme bestaat dat efficiënter is dan de SVD (of EVD) die we in ANACOR gebruiken. Het wordt pas interessant als we de theorie naar iets efficiënters kunnen vertalen. Dit is het geval als er veel identieke observaties zijn.

We gaan uit van een G als boven, maar gaan deze vóór toepassing van SVD 'in elkaar drukken': we tellen identieke rijen van G bij elkaar op. Als G bijvoorbeeld de linker matrix hieronder is, dan zijn we geinteresseerd in de rechter matrix:

1	0	0	1	0	0	1	0		3	0	0	3	0	0	3	0
0	1	0	0	1	0	0	1		0	1	0	0	1	0	0	1
1	0	0	0	0	1	0	1		1	0	0	0	0	1	0	1
1	0	0	1	0	0	1	0		0	1	0	1	0	0	0	1
0	1	0	1	0	0	0	1		0	0	2	2	0	0	0	2
0	0	1	1	0	0	0	1		2	0	0	2	0	0	0	2
1	0	0	1	0	0	1	0									
1	0	0	1	0	0	0	1									
0	0	1	1	0	0	0	1									
1	0	0	1	0	0	0	1									

Wiskundig geformuleerd: zij A een indikator matrix van nxv, waarbij v gelijk is aan het aantal <u>verschillende</u> profielen in G, en waarvoor geldt

$a_{ij} = 1$ als en alleen als individu i skoort profiel j,

waarbij j dit keer loopt van 1 tot v. We gaan nu in plaats van op G onze techniek toepassen op de matrix A'G. Hierboven hebben we G en A'G al gegeven. Ter verduidelijking geven we ook nog A:

1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
1	0	0	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
1	0	0	0	0	0
0	0	0	0	0	1
0	0	0	0	1	0
0	0	0	0	0	1

Merk op dat de rijsom van rij j van A'G gelijk is aan m maal het aantal individuen dat profiel j skoort, dus m maal het aantal enen in de j-de kolom van A, ofwel E=mA'A. Door optellen van rijen verandert er niets aan kolomtotalen, dus D blijft als boven, en ook blijft N=nm. Nu wordt (1):

$$m^{-\frac{1}{2}}\,(A'A)^{-\frac{1}{2}}A'GD^{-\frac{1}{2}} - \frac{m^{\frac{1}{2}}}{mn}\,(A'A)^{\frac{1}{2}}uu'D^{\frac{1}{2}} = \tilde{K}\,\tilde{\Lambda}\tilde{L}'$$

met $\tilde{K}'\tilde{K}=I$, $\tilde{L}'\tilde{L}=I$, $\tilde{\Lambda}$ diagonaal pxp, geordend. We geven met $\sim$ aan dat de matrices niet noodzakelijk gelijk hoeven te zijn aan de eerdergenoemde. Transponeren en met zich zelf vermenigvuldigen levert

$$m^{-1}\,D^{-\frac{1}{2}}G'A(A'A)^{-1}A'GD^{-\frac{1}{2}} - \frac{1}{mn}\,D^{\frac{1}{2}}uu'D^{\frac{1}{2}} = \tilde{L}\tilde{\Lambda}^2\tilde{L}' \qquad (10)$$

We willen nu bewijzen dat dit hetzelfde is als (5). Om dat te doen, merken we allereerst op, dat $G=A\tilde{G}$, waarbij $\tilde{G}$ een indikator matrix is, die alle v profielen precies één maal bevat. Hoe kunnen we $\tilde{G}$ in G uitdrukken? Door identieke rijen op te tellen, en dan door hun aantal te delen. In formule

$$\tilde{G} = (A'A)^{-1}A'G$$

Hieruit volgt

$$G = A\tilde{G} = A(A'A)^{-1}A'G \qquad \text{dus} \qquad G'A(A'A)^{-1}A'G = G'G$$

En hieruit volgt dat van (5) en (10) de linkerleden aan elkaar gelijk zijn, en dus ook de rechterleden, waarmee bewezen is dat $\tilde{\Lambda}=\Lambda$, $\tilde{L}=L$ (op eventuele spiegelingen na). Wat geldt er voor de individu-skores? We willen dat de skores die nu aan de profielen toegekend worden, dezelfde zijn als die welke eerst aan de bij deze profielen behorende individuen toegekend werden. Uit 3.2.1 volgt:

$$K = m^{-\frac{1}{2}}\,GD^{-\frac{1}{2}}L\Lambda^{-1} \quad ; \quad \tilde{K} = m^{-\frac{1}{2}}\,(A'A)^{-\frac{1}{2}}A'GD^{-\frac{1}{2}}\tilde{L}\tilde{\Lambda}^{-1}$$

In deze uitdrukking voor $\tilde{K}$ vervangen we $\tilde{L}\tilde{\Lambda}^{-1}$ door $L\Lambda^{-1}$, en dan substitueren we K:

$$\tilde{K} = (A'A)^{-\frac{1}{2}}A'K \quad ; \quad (A'A)^{-\frac{1}{2}}\tilde{K} = (A'A)^{-1}A'K$$

De uitdrukking rechts mogen we lezen als: tel identieke rijen op, en deel door het aantal. Gevolg is, dat de rijen van $(A'A)^{-\frac{1}{2}}\tilde{K} = m^{\frac{1}{2}}\,E^{-\frac{1}{2}}\tilde{K}$ dezelfde zijn als die van K; K gaf de individu-skores, dus $m^{\frac{1}{2}}\,E^{-\frac{1}{2}}\tilde{K}$ geeft de profiel-skores.

Als n>>v is de benadering via de v profielen (ANACO) efficiënter dan de benadering via n individuen (HOMALS). Ter illustratie: het nu volgende voorbeeld gaat over zes binaire variabelen, die

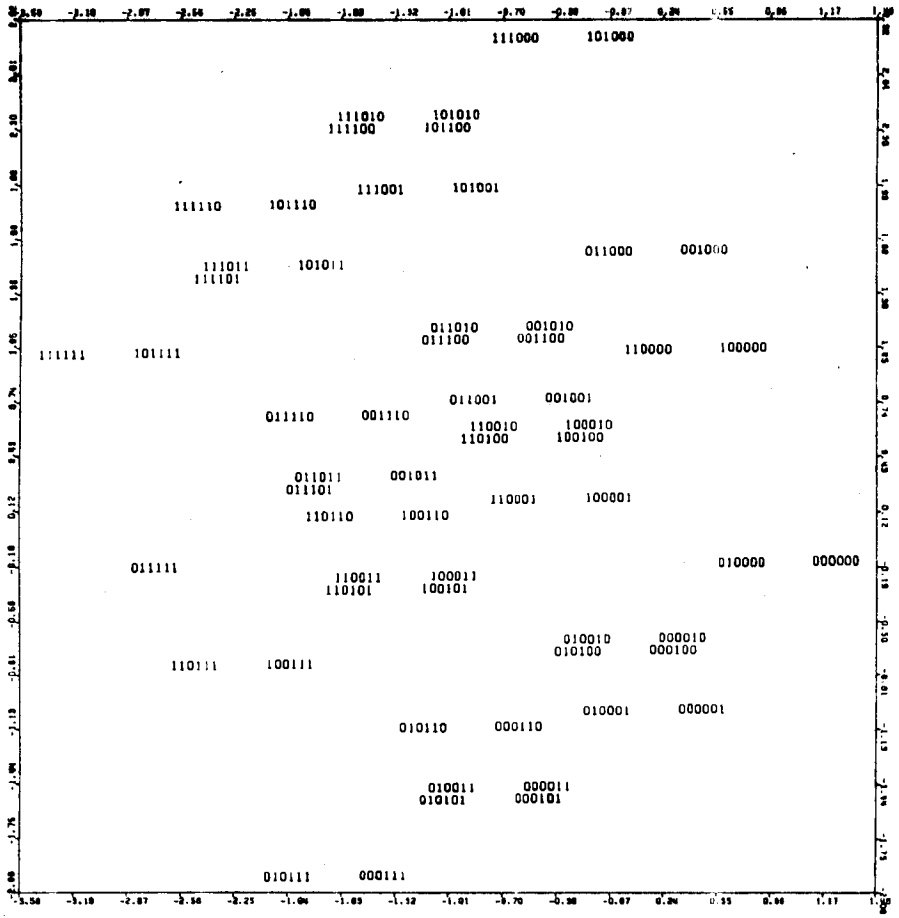
1	1	1	1	1	1	042		0	1	1	1	1	1	011
1	1	1	1	1	0	033		0	1	1	1	1	0	007
1	1	1	1	0	1	006		0	1	1	1	0	1	002
1	1	1	1	0	0	017		0	1	1	1	0	0	005
1	1	1	0	1	1	012		0	1	1	0	1	1	004
1	1	1	0	1	0	029		0	1	1	0	1	0	008
1	1	1	0	0	1	008		0	1	1	0	0	1	004
1	1	1	0	0	0	082		0	1	1	0	0	0	044
1	1	0	1	1	1	051		0	1	0	1	1	1	072
1	1	0	1	1	0	069		0	1	0	1	1	0	126
1	1	0	1	0	1	020		0	1	0	1	0	1	045
1	1	0	1	0	0	054		0	1	0	1	0	0	142
1	1	0	0	1	1	034		0	1	0	0	1	1	080
1	1	0	0	1	0	124		0	1	0	0	1	0	258
1	1	0	0	0	1	027		0	1	0	0	0	1	137
1	1	0	0	0	0	317		0	1	0	0	0	0	760
1	0	1	1	1	1	001		0	0	1	1	1	1	000
1	0	1	1	1	0	002		0	0	1	1	1	0	002
1	0	1	1	0	1	000		0	0	1	1	0	1	000
1	0	1	1	0	0	009		0	0	1	1	0	0	004
1	0	1	0	1	1	001		0	0	1	0	1	1	004
1	0	1	0	1	0	011		0	0	1	0	1	0	003
1	0	1	0	0	1	007		0	0	1	0	0	1	006
1	0	1	0	0	0	059		0	0	1	0	0	0	030
1	0	0	1	1	1	008		0	0	0	1	1	1	033
1	0	0	1	1	0	023		0	0	0	1	1	0	048
1	0	0	1	0	1	007		0	0	0	1	0	1	038
1	0	0	1	0	0	035		0	0	0	1	0	0	064
1	0	0	0	1	1	010		0	0	0	0	1	1	042
1	0	0	0	1	0	055		0	0	0	0	1	0	096
1	0	0	0	0	1	013		0	0	0	0	0	1	090
1	0	0	0	0	0	194		0	0	0	0	0	0	718

tabel 3.3. Sugiyama data (godsdienst in Japan) links profielen, rechts frekwenties.

1. PRACT Do you make it a rule to practice religious conduct, such as attending religious services, religious worship and missionary works and do you occasionally offer prayers or chant sutras?
2. GRAVE Do you visit a grave once or twice a year?
3. BOOKS Do you occasionally read religious books, such as the Bible or the Buddhist Scriptures?
4. SUCCE Do you visit shrines and temples to pray for business prosperity, succes in an entrance examination and so forth?
5. MASCO Do you keep a talisman, such as an amulet, charm or mascot near you?
6. FORTU Did you draw a fortune, consult a diviner or had you your fortune told within the last years?

dus 2^6 = 64 profielen kunnen hebben (in feite zijn het er, doordat enkele marginale frekwenties nul zijn, 61), met 4243 individuen. De rekentijd met HOMALS-GS was 70 sec., met ANACOR 3 sec.

De zes variabelen betreffen een aantal godsdienstige praktijken in Japan; de frekwenties waarmee de profielen in de steekproef voorkwamen, alsmede de inhoudelijke omschrijving van de praktijken staan in tabel 3.3. De plot van de profielpunten in figuur 3.15. De eigenwaarden voor deze eerste twee assen zijn .2692 en .2037. We zien in de plot dat van rechts onder naar links



figuur 3.15. Profiel-punten voor Sugiyama data.

boven de 'religieuze' items 2, 1, 3 zijn die steeds vaker met 'ja' beantwoord worden, terwijl van rechts boven naar links onder de 'pragmatiese' items 5, 4, 6 vaker beaamd worden. Een plot van de kategorieskores is te vinden in hoofdstuk 4.1., waar nader op de stabiliteit van de punten wordt ingegaan.

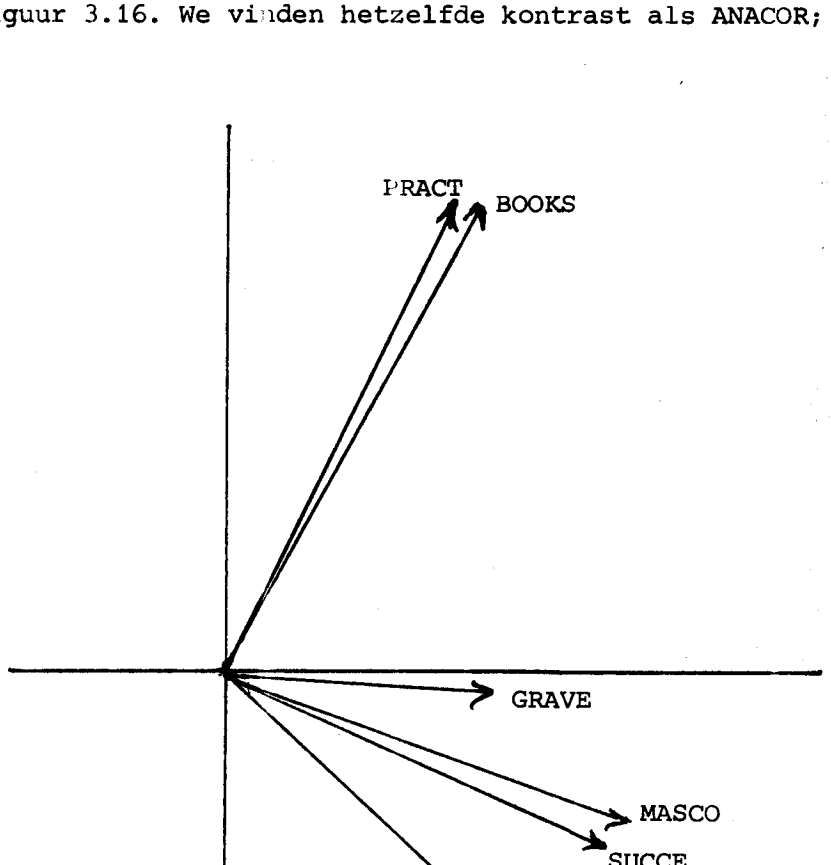
Zoals we al in hoofdstuk 2 gezien hebben is het bijzondere van binaire gegevens, dat we elke kategorie een willekeurige schaalwaarde mogen toekennen (weliswaar twee kategorieën van dezelfde variabele niet dezelfde schaalwaarde), en aan de hand van de observaties de korrelatiematrix van de variabelen kunnen uitrekenen.

Deze korrelatiematrix is onafhankelijk van de gekozen schaalwaarden. In ons voorbeeld wordt de korrelatiematrix gegeven door tabel 3.4. Bij andere

1.0000					
.0853	1.0000				
.2842	.0522	1.0000			
.0770	.1119	.0667	1.0000		
.0971	.1631	.0519	.2785	1.0000	
-.0182	.0614	.0407	.2111	.2018	1.0000

tabel 3.4. Korrelaties Sugiyama.

schaling is het alleen mogelijk dat overeenkomstige rijen en kolommen van van teken veranderen. We zien aan de korrelatiematrix dat 1 en 3 enerzijds en 4,5 en 6 anderzijds relatief hoog korreleren. Toepassing van PCA in twee dimensies levert eigenwaarden 1.6149 en 1.2221, de eigenvektoren zijn geplot in figuur 3.16. We vinden hetzelfde kontrast als ANACOR; omdat de gewogen



figuur 3.16. PCA oplossing Sugiyama.

kategorie kwantifikaties van elke variabele op elke as optellen tot 1, kunnen kunnen we bij binaire gegevens in feite volstaan met voor iedere variabele één kategorie kwantifikatie. Door andere normalisatie leveren PCA en ANACOR niet exact hetzelfde beeld. We kunnen het verband weergeven als

$$\mu_q = m\,\lambda^2_q \qquad (q=1,\ldots,m)$$
$$mn\,(x^q_j)^2 = D^j_1(y^j_{1q})^2 + D^j_2(y^j_{2q})^2 \qquad (q=1,\ldots,m;\ j=1,\ldots,m)$$
$$D^j_1\,y^j_{1q} + D^j_2\,y^j_{2q} = 0 \qquad (j,q=1,\ldots,m)$$

We kunnen, afgezien van het teken, x^q_j enerzijds en y^j_{1q} , y^j_{2q} anderzijds in elkaar uitdrukken.

3.2.3. Twee-dimensionale kruistabellen.

We kunnen met Analyse des Correspondances de samenhang van twee nominale variabelen onderzoeken. We gaan uit van een frekwentietabel F van afmetingen $k_1 x k_2$ (aantal kategorieën van de eerste resp. de tweede variabele), met elementen f_{ij} die gelijk zijn aan het aantal individuen dat zowel in kategorie i van variabele 1 als in kategorie j van variabele 2 skoort. Wat we in 3.2.1 rij-objekten noemden, zijn nu de kategorieën van de eerste variabele. Kolom-objekten zijn kategorieën van de tweede variabele. We substitueren in (1) voor G nu F, voor de rij-totalen E(e) de marginalen van de eerste variabele $D_1(d_1)$, voor de kolom-totalen D(d) de marginalen van de tweede variabele $D_2(d_2)$ en voor de som van alle matrix-elementen N het aantal observaties n. Voor (1) kunnen we nu schrijven

$$D_1^{-\frac{1}{2}}FD_2^{-\frac{1}{2}} - 1/n\ D_1^{-\frac{1}{2}}uu'D_2^{-\frac{1}{2}} = K\Lambda L' \qquad K'K=I,\ L'L=I,\ \Lambda \text{ diagonaal, geordend} \qquad (11a)$$

Voor een weergave van de kategorieën in p dimensies gebruiken we weer K_p, die de eerste p kolommem van K bevat, L_p analoog voor L en Λ_p. Net als in 3.2.1. nemen we $X_p = n^{\frac{1}{2}}\,D_1^{-\frac{1}{2}}K_p\Lambda_p$ en $Y_p = n^{\frac{1}{2}}\,D_2^{-\frac{1}{2}}L_p\Lambda_p$ en definieren als kwantifikaties van de kategorieën van de 1^e resp. 2^e variabele

$$Y_1 \triangleq X_p\Lambda_p^{-1} = n^{\frac{1}{2}}D_1^{-\frac{1}{2}}K_p \qquad (11b)$$
$$Y_2 \triangleq Y_p\Lambda_p^{-1} = n^{\frac{1}{2}}D_2^{-\frac{1}{2}}L_p \qquad (11c)$$

Uit (4) volgt

$$Y_1 = D_1^{-1}FY_2\Lambda_p^{-1} \qquad (12a)$$
$$Y_2 = D_2^{-1}F'Y_1\Lambda_p^{-1} \qquad (12b)$$

Formules (12) zijn interessant, omdat de matrices $D_1^{-1}F$ en $D_2^{-1}F'$, die Y_1 en Y_2 relateren, probabilisties geinterpreteerd kunnen worden, nl.

$$(D_1^{-1}F)_{ij} = f_{ij}/d_i^1$$

We kunnen $(1/n)f_{ij}$ zien als benadering van de kans dat een individu zowel in kategorie i van variabele 1 als in kategorie j van variabele 2 valt, en $(1/n)d_i^1$ als benadering van de kans dat een individu in kategorie i van variabele 1 valt. Daarmee is f_{ij}/d_i^1 een benadering van de voorwaardelijke kans op i én j, gegeven i. Het omgekeerde geldt voor $D_2^{-1}F'$. De skores van de kategorieën van de variabelen hangen dus samen via deze voorwaardelijke kansen, als gegeven in (12).

We kunnen de samenhang van twee nominale variabelen ook, zoals in 3.2.2, nagaan via de bijbehorende indikator super matrix. Zijn de resultaten dan hetzelfde als hierboven? We gaan dit als volgt na. Stel $G=(G_1\ G_2)$ is een indikator super matrix van $n\times(k_1+k_2)$. Met F, D_1 en D_2 als boven is

$$G'G = \begin{bmatrix} G_1'G_1 & G_1'G_2 \\ G_2'G_1 & G_2'G_2 \end{bmatrix} = \begin{bmatrix} D_1 & F \\ F' & D_2 \end{bmatrix} \qquad (13)$$

immers in $G_1'G_1$ staan de aantallenindividuen die in de verschillende kategorieën van variabele 1 vallen, evenals in D_1; idem voor $G_2'G_2$ en D_2; en in $G_1'G_2$ staat voor ieder paar kategorieën het aantal individuen, dat in dit paar valt, en dat staat ook in F. We laten nu zien, dat (5) en (11) ekwivalent zijn. Eerst herschrijven we (13):

$$D^{-\frac{1}{2}}G'GD^{-\frac{1}{2}} = \begin{bmatrix} I & D_1^{-\frac{1}{2}}FD_2^{-\frac{1}{2}} \\ D_2^{-\frac{1}{2}}F'D_1^{-\frac{1}{2}} & I \end{bmatrix}$$

en hierin substitueren we de resultaten van (11)

$$D^{-\frac{1}{2}}G'GD^{-\frac{1}{2}} = \begin{bmatrix} I & K\Lambda L' \\ L\Lambda K' & I \end{bmatrix} + 1/n\ D^{\frac{1}{2}}uu'D^{\frac{1}{2}} - \begin{bmatrix} 1/n\ D_1^{\frac{1}{2}}uu'D_1^{\frac{1}{2}} & \\ & 1/n\ D_1^{\frac{1}{2}}uu'D_1^{\frac{1}{2}} \end{bmatrix}$$

dus (vgl (5)) moeten we de eigenwaardenontbinding bepalen van

$$1/2\ D^{-\frac{1}{2}}G'GD^{-\frac{1}{2}} - 1/2n\ D^{\frac{1}{2}}uu'D^{\frac{1}{2}} = 1/2\begin{bmatrix} I & K\Lambda L' \\ L\Lambda K' & I \end{bmatrix} - 1/2n\begin{bmatrix} D_1^{\frac{1}{2}}uu'D_1^{\frac{1}{2}} & \\ & D_1^{\frac{1}{2}}uu'D_1^{\frac{1}{2}} \end{bmatrix}$$

Deze eigenwaarden ontbinding is

$$\begin{bmatrix} \frac{1}{2}\sqrt{2}\ K & \frac{1}{2}\sqrt{2}\ K \\ \frac{1}{2}\sqrt{2}\ L & -\frac{1}{2}\sqrt{2}\ L \end{bmatrix}\begin{bmatrix} \frac{1}{2}(I+\Lambda) & \\ & \frac{1}{2}(I-\Lambda) \end{bmatrix}\begin{bmatrix} \frac{1}{2}\sqrt{2}\ K' & \frac{1}{2}\sqrt{2}\ L' \\ \frac{1}{2}\sqrt{2}\ K' & -\frac{1}{2}\sqrt{2}\ L' \end{bmatrix}$$

en analoog aan 3.2.2 is

$$Y = \sqrt{2}\sqrt{n} \begin{bmatrix} D_1^{-\frac{1}{2}} & \\ & D_2^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{2}\sqrt{2}\ K & \frac{1}{2}\sqrt{2}\ K \\ \frac{1}{2}\sqrt{2}\ L & -\frac{1}{2}\sqrt{2}\ L \end{bmatrix} = \sqrt{n} \begin{bmatrix} D_1^{-\frac{1}{2}}K & D_1^{-\frac{1}{2}}K \\ D_2^{-\frac{1}{2}}L & -D_2^{-\frac{1}{2}}L \end{bmatrix}$$

en dit is juist de schalingvoor de kategorieën die we uit (11) afgeleid hebben. We kunnen konkluderen dat Analyse des Correspondances op de kruistabel ekwivalent is aan HOMALS bij twee variabelen. Behandeling van de kruistabel heeft als voordeel dat er een kleiner probleem geanalyseerd wordt, dus sneller werkt, als nadeel dat de individu-skores niet gevonden worden. Ter illustratie van de snelheidswinst: het nu volgende voorbeeld gaat over twee nominale variabelen, elk met 7 kategorieën. Er zijn 49 profielen (waarvan 2 leeg). Analyse van de 47x14 matrix (met individu-skores) kost 2.77 sec., analyse van de 7x7 matrix (zonder individu-skores) kost .75 sec. Voor de grootste eigenwaarden van de analyse van G en F geldt $\lambda_G^2 = (1+\lambda_F)/2$

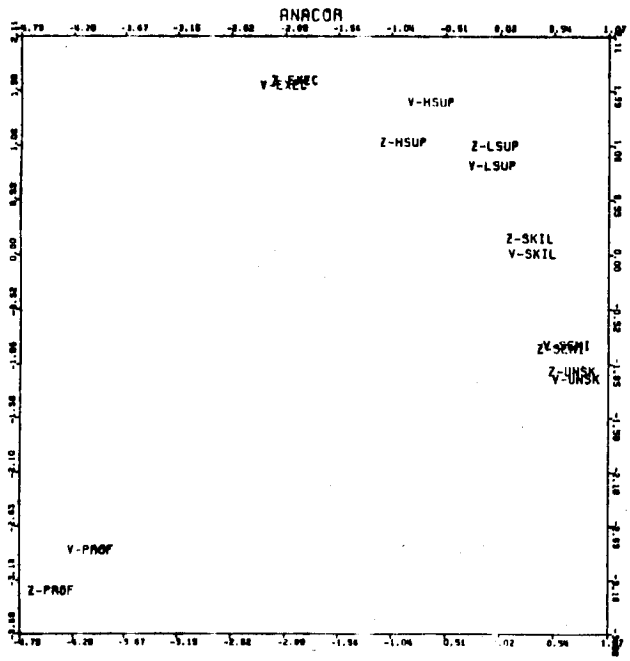
Het voorbeeld betreft de analyse van de bekende sociale mobiliteitstabel van Glass (1954), die ook in Goodman (1965, 1969), Haberman (1974) en Bishop e.a (1975) uitvoerig besproken wordt (zie tabel 3.5.).

beroep vader \ beroep zoon	1	2	3	4	5	6	7
1	50	19	26	8	18	6	2
2	16	40	34	18	31	8	3
3	12	35	65	66	123	23	21
4	11	20	58	110	223	64	32
5	14	36	114	185	714	258	189
6	0	6	19	40	179	143	71
7	0	3	14	32	141	91	106

tabel 3.5. Sociale Mobiliteit

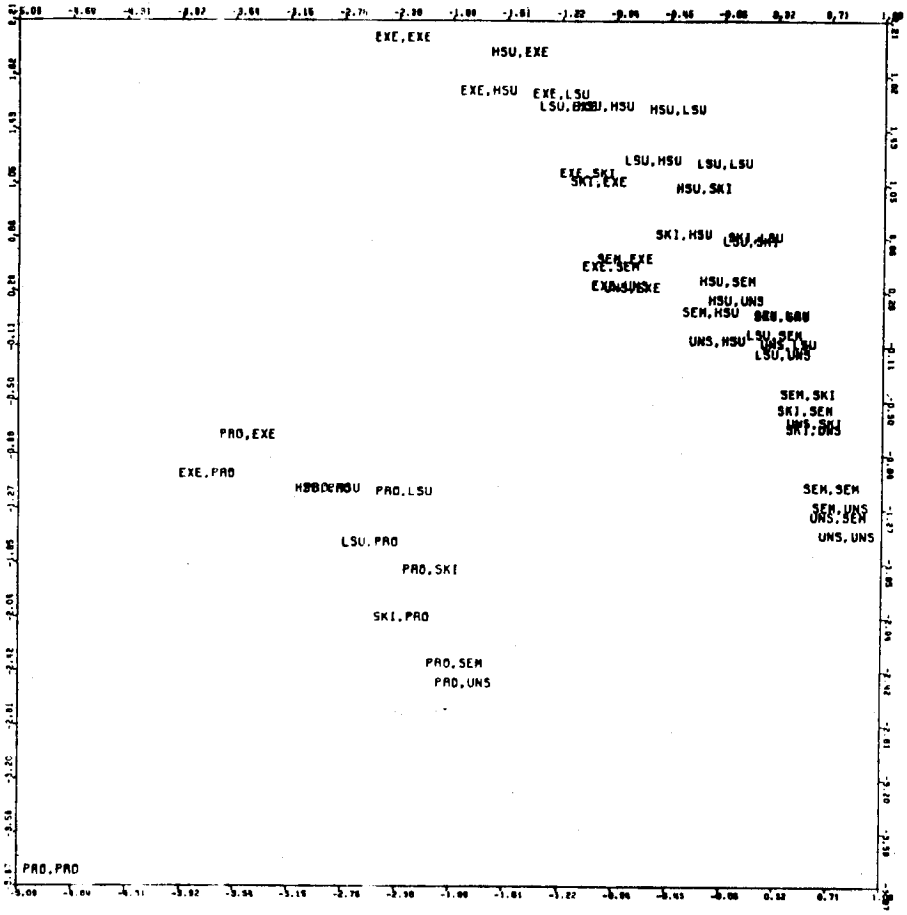
1.PROF: professional and high admin.
2.EXEC: managerial and executive
3.HSUP: higher supervisory
4.LSUP: lower supervisory
5.SKIL: skilled manual and routine non-manual
6.SEMI: semi-skilled manual
7.UNSK: unskilled manual

De tabel geeft aan, voor een steekproef van 3497 Engelsen, in welke beroepsgroep de vader valt en in welke de zoon. De observatie-eenheden zijn dus eigenlijk gezinnen. Toepassing van ANACOR levert de resultaten van figuur 3.17. De eigenwaarden zijn .5256 en .2674. We zien dat op de eerste as de ordening van de beroepsgroepen loopt van 'hoog' naar 'laag', en dat er een groot effekt van asymmetrie is (de schaling voor vaders en zonen ontlopen elkaar nauwelijks). Er is wel een sterk kontrast tussen gezinnen waar of vader of



figuur 3.17. Kategorie-punten voor Sociale Mobiliteit.

zoon PROF is en de andere gezinnen. Dit heeft tot effekt dat in figuur 3.18, waar de gezinnen zijn geplot, tussen PROF,PROF en de rest alle gezinnen liggen waarvan minstens één van de twee de hoogste status PROF heeft.



figuur 3.18. Individu-punten Sociale Mobiliteit.

3.2.4. Matrices van bivariate marginalen.

Over matrices van bivariate marginalen C van afmetingen Σk_j x Σk_j kunnen we kort zijn. Het is snel in te zien dat de analyse van deze matrices hetzelfde oplevert als analyse van indikatormatrixen, maar dat we nu geen individu-skores vinden.

Stel D is de bij C horende diaginaalmatrix van marginalen. We merken eerst op, dat $Cu=C'u=m\ Du$, en $N=u'Cu=m\ u'Du=m^2n$. In woorden: elke rij en kolom van C sommeert tot m maal het bijbehorende diagonaalelement, en in de m^2 blokken C_{ij} staan steeds n observaties. Volgens (1) moeten we SVD toepassen op

$$1/m\ D^{-\frac{1}{2}}CD^{-\frac{1}{2}} - 1/mn\ D^{\frac{1}{2}}uu'D^{\frac{1}{2}}$$

Dit is een symmetriese matrix, dus SVD is eigenwaardenontbinding. Stel het resultaat is

$$L_c \Lambda_c L_c' \quad \text{met } L_c'L_c=I,\ \Lambda_c \text{ diagonaal, geordend, positief.}$$

We kijken nu wat we bij indikatormatrices deden, en we zien dat we hierboven juist formule (5) herhaald hebben, en dus dezelfde analyse als in 3.2.2 uitvoeren.

We besluiten dit hoofdstuk met een voorbeeld dat een aantal eigenaardigheden vertoont die we in 't algemeen met name kunnen verwachten bij gestratificeerde steekproeven. De Nederlandse Studenten Raad verzamelde in 1968 de eerste keuzen tussen vijf politieke partijen van 1616 studenten in het hele land. De steekproef was gestratificeerd over 12 universiteiten (cq.TH's) en 13 fakulteiten. De gegevens (zie ook Lammers, 1969) staan in tabel 3.6. De volgorde van de partijen is: CDA (d.w.z. KVP, ARP, CHU, GPV, SGP); VVD; PVDA; PACO (d.w.z. de kleine linkse partijen); D'66.

De eerste vier assen van ANACOR staan geplot in figuur 3.19. De eerste as (met eigenwaarde .6718) vertelt ons, dat techniese wetenschappen gestudeerd worden in Delft, Drienenoord en Eindhoven; de tweede (met eigenwaarde .6579), dat er een landbouwhogeschool is in Wageningen; en de derde (met eigenwaarde .5385) zegt dat er in Rotterdam en Tilburg ekonomiese universiteiten zijn.

Dit zijn geen hemelschokkende resultaten, en veel hebben we er niet aan. Het zou voorbarig zijn, nu te konkluderen dat 'politieke voorkeur' niets 'doet'. Dit heeft te maken met de aard van één van de drie kruistabellen. De bivariate marginalen van 'Universiteiten' maal 'studierichtingen' kunnen eigenlijk niet als random worden opgevat, een groot aantal kombinaties kunnen niet voorkomen,

	leiden					a'dam vu					a'dam gu				
rechten	13	32	9	2	18	1	0	0	1	1	3	14	12	2	10
medicijn	3	20	7	4	10	7	1	3	1	2	4	6	4	7	8
wis&nat	2	15	4	5	15	6	0	1	0	1	1	4	13	4	19
soc.wet.	5	12	4	3	14	2	0	1	0	2	1	8	28	11	19
letteren	5	8	6	3	5	2	0	0	1	0	7	8	16	8	16
techn.wet.															
pol.-soc.											1	1	5	0	4
veearts															
tandheel															
theologie	1	1	3	2	0	5	0	0	0	0	1	0	0	0	0
landbouw															
filosofie	0	0	1	2	0	0	0	0	0	1	2	2	6	5	7
economie						4	0	0	0	2	7	16	8	6	20

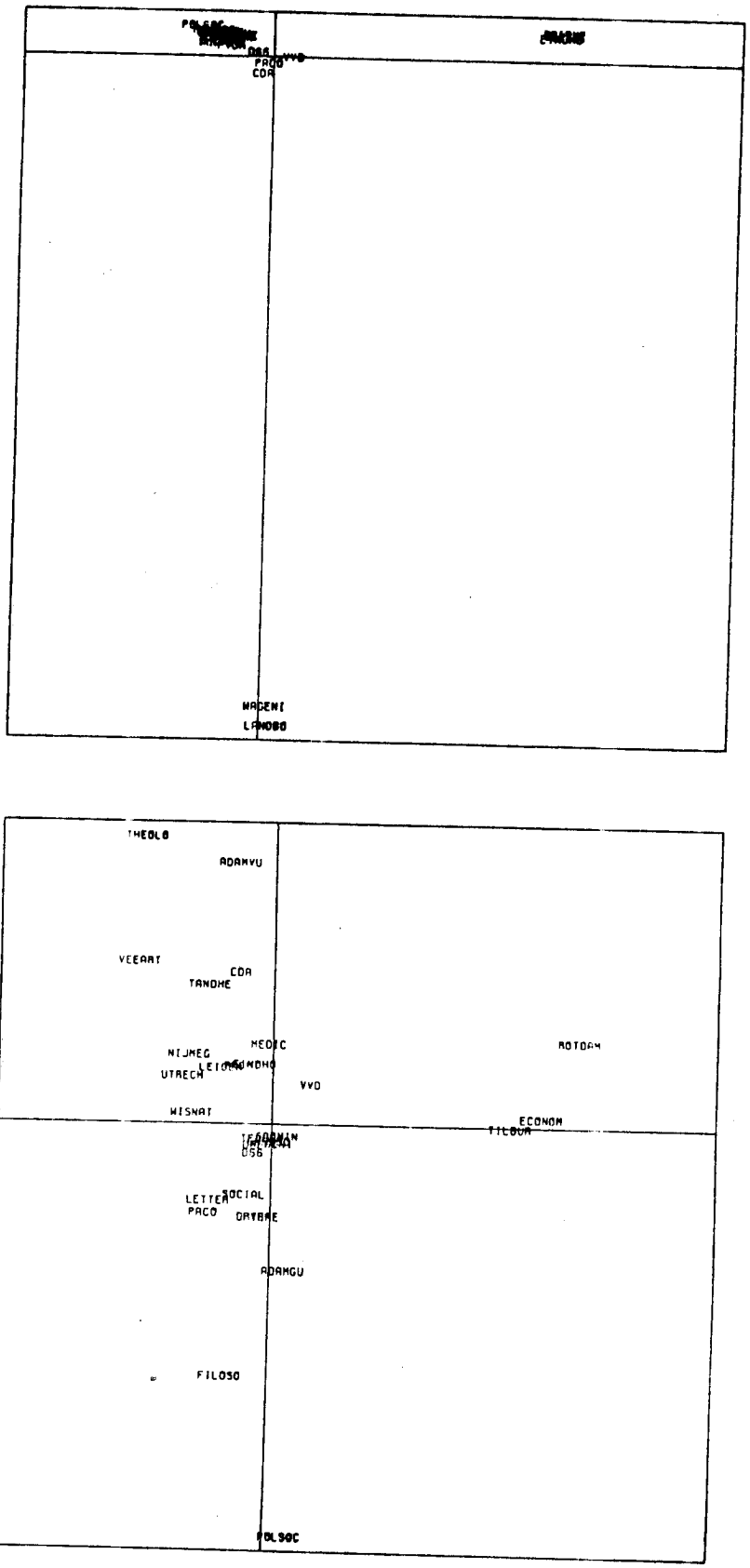
	utrecht					wageningen					groningen				
rechten	7	16	1	0	5						2	8	6	0	7
medicijn	6	11	9	2	11						3	7	3	2	11
wis&nat	10	14	16	7	21						7	5	9	2	12
soc.wet.	6	4	5	10	11	0	1	1	0	0	4	5	11	6	10
letteren	3	10	4	4	13						1	5	6	2	6
techn.wet.															
pol.-soc.															
veearts	3	6	1	1	5										
tandheel	0	3	0	0	1						1	2	0	1	1
theologie	5	0	0	2	1						1	0	2	0	1
landbouw						11	14	7	6	14					
filosofie	0	1	0	1	4						0	1	1	1	0
economie						0	0	0	0	1	1	13	5	0	7

	drienenoord					nijmegen					tilburg				
rechten						5	7	3	1	6	0	0	1	0	2
medicijn						6	9	3	2	5					
wis&nat						7	2	2	4	6					
soc.wet.						4	2	5	4	22	5	1	0	3	6
letteren						4	0	8	2	4					
techn.wet.	0	7	3	4	5										
pol.-soc.															
veearts															
tandheel						1	2	0	0	2					
theologie						6	0	0	0	5					
landbouw															
filosofie						0	0	1	0	0					
economie											3	11	2	0	9

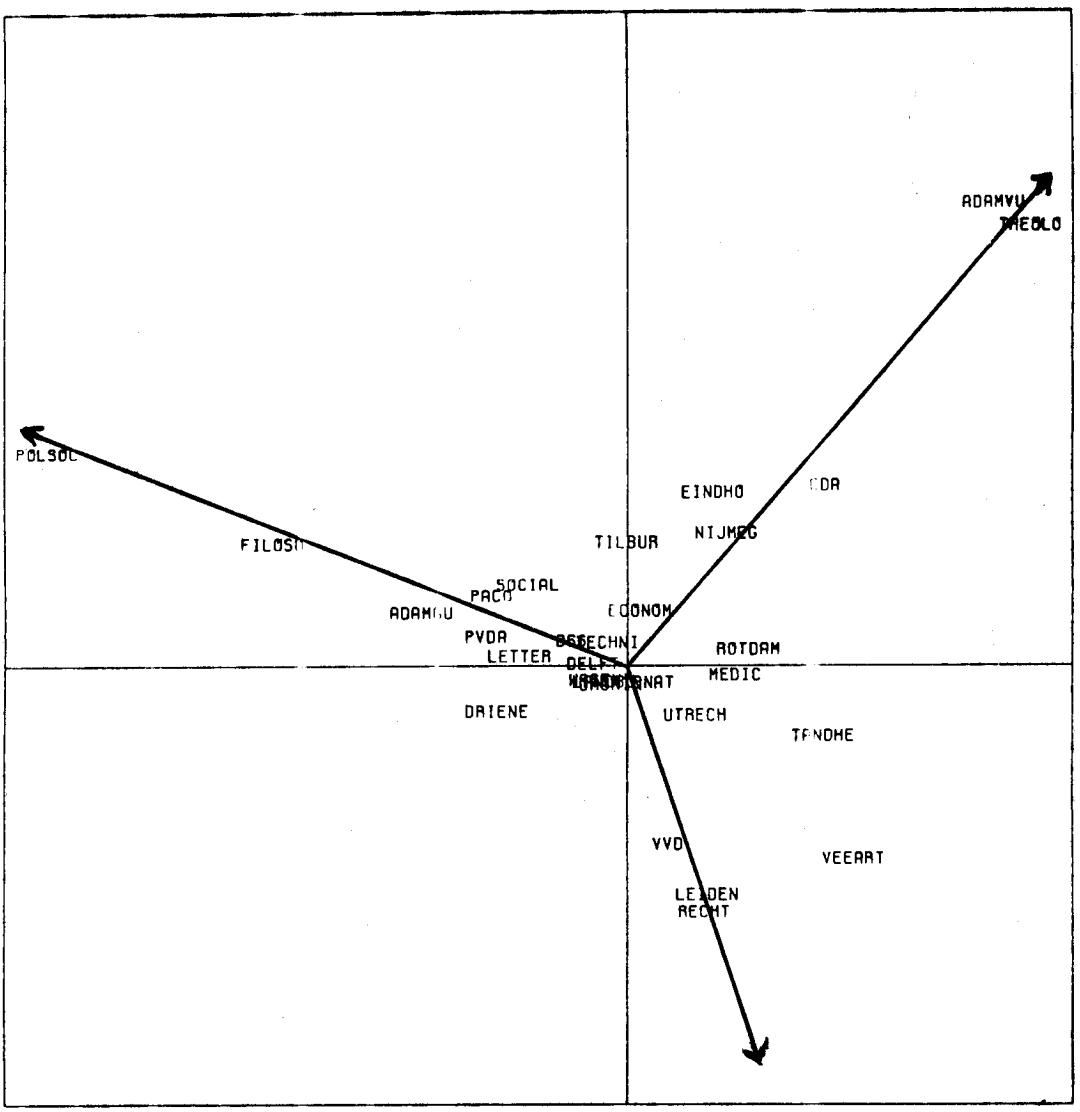
	rotterdam					eindhoven					delft				
rechten	2	2	1	0	0										
medicijn	8	7	1	1	7										
wis&nat															
soc.wet.	1	1	3	1	1										
letteren															
techn.wet.						12	7	3	0	13	24	66	22	20	50
pol.-soc.															
veearts															
tandheel															
theologie															
landbouw															
filosofie															
economie	9	33	15	2	16										

tabel 3.6. Aantallen eerste keuzes op CDA, VVD, PVDA, PACO en D66 van studenten uit 1968.

de meeste andere zijn 'fixed'. Nu kunnen we bovendien in 't algemeen verwachten, dat ANACOR/HOMALS één perfekte as kan fitten in het geval van één kruistabel waarvoor geldt dat er minstens één rij i en één kolom j bestaat, zodanig dat alle cellen behalve (i,j) nul zijn. Als in ons geval van drie kruistabellen er één is die voor één rij en één kolom aan het bovenstaande voldoet, worden



figuur 3.19. Assen 1 vs 2 (boven) en 3 vs 4 (onder) van ANACOR op NSR data.



figuur 3.20. Vierde en vijfde as van ANACOR op NSR data.

worden op die perfekte as de betreffende twee variabelen maximaal homogeen en de derde in 't geheel niet gerepresenteerd; we verwachten dan een eigenwaarde van 2/3 voor die as. Als er nog een rij en een kolom te vinden is, waarvoor alle cellen op één na leeg zijn, krijgen we een tweede as waar weer twee van de drie variabelen perfekt op gerepresenteerd worden met eigenwaarde 2/3. Data-analyties gezien kunnen we nu twee kanten op: een variabele weggooien of kategorieën samennemen, dan wel naar meer assen kijken. Dit laatste doen we hier (zie figuur 3.20). Het is duidelijk dat ook voor studenten in 1968 Nederland 'driestromenland' is, en het is ook duidelijk in welke plaatsen en in welke studierichtingen de drie stromen het sterkst vertegenwoordigd zijn.

4.1 Betrouwbaarheid van HOMALS oplossingen

4.1.0 Inleiding

Tot nu toe hebben we met HOMALS of ANACOR getalwaarden toegekend aan kategorieën en (groepen) individuen. We gaan nu de stabiliteit van deze resultaten onderzoeken, ofwel de gevoeligheid voor kleine veranderingen in de data-set. Dit doen we op twee manieren, namelijk de in hoofdstuk 9 te behandelen bootstrap methode, en de nu aan de orde zijnde statistisch-analytische benadering. We nemen in deze paragraaf aan dat de data een aselekte steekproef vormen getrokken uit een oneindig grote populatie (of, wat ekwievalent is, een steekproef met teruglegging uit een eventueel eindig grote populatie). De HOMALS oplossing bestaat uit schatters van populatiegrootheden, en we gaan eigenschappen van deze schatters bekijken. Verwante literatuur is De Leeuw (1973, paragraaf 6.8), Lebart (1976), O'Neill (1978a, 1978b), Davis (1977).

4.1.1 Populatie-grootheden

We bekijken de matriks die we in hoofdstuk 3 A'G noemden eens nader. Voor elke rij van deze matriks geldt dat de plaats van de niet-nul elementen alleen bepaald wordt door het profiel dat bij deze rij hoort, en de grootte van deze elementen alleen door het aantal individuen dat bij dit profiel hoort. Noem dit aantal $\underline{n}_\upsilon$ (stochastisch, want we praten nu over steekproeven) voor de υ^e rij van A'G, dan kunnen we de rij schrijven als $\underline{n}_\upsilon \times$ (een vaste rij). Stel N_{obs} is het aantal observaties, dan is $\underline{\pi}_\upsilon \overset{\Delta}{=} \underline{n}_\upsilon / N_{obs}$ de steekproefkans op rij υ. Verzamelen we de vaste rijen nu in een matriks, die we bij gebrek aan symbolen maar weer G noemen, en definieren we $\underline{\Pi} \overset{\Delta}{=} \text{diag}(\underline{\pi}_1,\ldots,\underline{\pi}_n)$, met n het aantal rijen van G, dan is de te analyseren matriks $N_{obs}\underline{\Pi}G$. Hierin is G een indikator-supermatriks waarin alle profielen staan. De elementen van iedere rij van G sommeren tot m, het aantal variabelen, ofwel $Gu = mu$.

We maken nu de stap van HOMALS naar ANACOR, ofwel van indikator matriks naar willekeurige matriks met niet-negatieve elementen, door de volgende generalisatie. Zij G een willekeurige $n \times k$-matriks met niet-negatieve elementen en vaste rijsom μ: $Gu = \mu u$. We stappen ook over van steekproef naar populatie: de kans op trekking van rij υ van G is $\pi_\upsilon > 0$. We voeren enkele symbolen in:

$$\Pi \overset{\Delta}{=} \text{diag}(\pi_1,\ldots,\pi_n),$$
$$\pi \overset{\Delta}{=} \Pi u.$$

De som van alle kansen is 1, ofwel $\pi'u = 1$. We gaan de matriks G analyseren met de methode uit 3.2.1 (bij een steekproef analyseren we $N_{obs}\Pi G$, maar de konstante N_{obs} heeft geen invloed op de uitkomsten. De randfrekwenties van G zijn

$$e = \Pi G u = \mu \Pi u = \mu\pi, \text{ dus } E = \mu\Pi,$$
$$d = G'\Pi u = G'\pi.$$

(Lezen we voor G even een indikator-supermatriks, dan staat hier d_υ^j is de kans dat een trekking voor variabele j in kategorie υ valt). De som van alle elementen van ΠG is

$$T = u' Gu = \quad .$$

We hebben nu voldoende om formules (1a) en (1b) uit 3.2.1 te kunnen herschrijven:

$$\mu^{-\frac{1}{2}}\Pi^{\frac{1}{2}}GD^{-\frac{1}{2}} - \mu^{-\frac{1}{2}}\Pi^{\frac{1}{2}}uu'D^{\frac{1}{2}} = K\Lambda L' \tag{1}$$

$$K'K = I,\ L'L = I,\ \Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_r),\ \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0.$$

Transponeren en met zichzelf vermenigvuldigen levert

$$\mu^{-1}D^{-\frac{1}{2}}G'\Pi GD^{-\frac{1}{2}} - \mu^{-1}D^{\frac{1}{2}}uu'D^{\frac{1}{2}} = L\Lambda^2L'. \tag{2}$$

Geheel analoog aan 3.2.1 zijn de kolom-skores

$$Y = \mu^{\frac{1}{2}}D^{-\frac{1}{2}}L.$$

Ze voldoen aan

$$G'\Pi GY = \mu DY\Lambda^2;\ Y'DY = \mu I;\ u'DY = 0. \tag{3}$$

Met G een indikator-supermatriks is dit de populatieversie van HOMALS. De populatie-rijskores zijn

$$X = \mu^{\frac{1}{2}}E^{-\frac{1}{2}}K = \Pi^{-\frac{1}{2}}K = GY\Lambda^{-1}. \tag{4}$$

4.1.2 Steekproeftrekking

We gaan in drie stappen van populatie naar steekproef.

a: Wat gebeurt er met X, Y, en Λ^2 wanneer we Π "storen" ?

b: Wat is het gedrag van een schatter van Π ?

c: Wat is het gedrag van schatters van X, Y, en Λ^2 (dat wil zeggen: kombineer a en b) ?

a: Zij $\tilde{\Pi}$ een diagonaalmatriks waarvan alle elementen hooguit infinitesimaal verschillen van Π. In plaats van (3) willen we nu oplossen

$$G'\tilde{\Pi}G\tilde{Y} = \mu\tilde{D}\tilde{Y}\tilde{\Lambda}^2;\ \tilde{Y}'\tilde{D}\tilde{Y} = \mu I;\ u'\tilde{D}\tilde{Y} = 0; \tag{5}$$

en daarna

$$\tilde{X} = G\tilde{Y}\tilde{\Lambda}^{-1}. \tag{6}$$

Hier zijn $\tilde{D}$, $\tilde{Y}$, $\tilde{\Lambda}^2$, en $\tilde{X}$ funkties van $\tilde{\Pi}$, die we hier niet nader specificeren.

We geven de globale lijn van de afleiding van $\tilde{Y}$ en $\tilde{\Lambda}^2$. Voor $\tilde{X}$ hebben we (6). Zij $1 \leq j \leq n$. Notatieafspraak: G_j^D is de diagonaalmatriks waarvan de elementen in de j^e rij van G staan; $g_j \triangleq G_j^D u$ (dus g_j' is de j^e rij van G).

We differentieren de drie vergelijkingen (5) naar $\tilde{\pi}_j$ (zie bijvoorbeeld Wilkinson, 1965, hoofdstuk 2, of Kato, 1966). De afgeleiden in het punt $\tilde{\Pi} = \Pi$ geven we aan met .j:

$$G'f_jf_j'GY + G'\Pi GY_{.j} = \mu D_{.j}Y\Lambda^2 + \mu DY_{.j}\Lambda^2 + \mu DY\Lambda^2_{.j},$$

$$Y'_{.j}DY + Y'D_{.j}Y + Y'DY_{.j} = 0,$$

$u'D_{.j}Y + u'DY_{.j} = 0.$

Hierin is $D_{.j} = G_j^D$ en $G'f_j = g_j$. Met $Y'DY = I$ volgt uit de eerste vergelijking

$$\Lambda^2_{.j} = \mathrm{diag}(\mu^{-2}Y'g_jg_j'Y - \mu^{-1}Y'G_j^DY\Lambda^2).$$

Voor de s^e kolom van de eerste vergelijking geldt

$$(G'\Pi G - \mu\lambda_s^2 D)y_{s.j} = \mu\lambda_s^2 G_j^D y_s - g_j g_j' y_s - \mu\lambda^2_{s.j} Dy_s.$$

Met gebruikmaking van de tweede en derde vergelijking kan $y_{s.j}$ hieruit opgelost worden.

b: Uit een steekproef ter grootte N_{obs} kunnen we Π schatten met $\underline{\Pi} = \mathrm{diag}(\underline{\pi}_1,\ldots,\underline{\pi}_n)$, waarin $\underline{\pi}_\upsilon$ de fraktie observaties is die gelijk is aan rij υ van G. Van $\underline{\Pi}$ is bekend dat

$$E(\underline{\pi}_\upsilon) = \pi_\upsilon,$$

$$V(\underline{\pi}_\upsilon) = (\pi_\upsilon - \pi_\upsilon^2)/N_{obs},$$

$$C(\underline{\pi}_{\upsilon_1},\underline{\pi}_{\upsilon_2}) = -\pi_{\upsilon_1}\pi_{\upsilon_2}/N_{obs}.$$

In matriks-notatie

$$E(\underline{\Pi}) = \Pi,$$

$$V(\underline{\Pi}) = (\Pi - \pi\pi')/N_{obs}.$$

c: We moeten nu het gedrag bepalen van de oplossingen van

$$G'\underline{\Pi}G\underline{Y} = \mu D\underline{Y}\Lambda^2;\ \underline{Y}'D\underline{Y} = \mu I;\ u'D\underline{Y} = 0. \tag{7}$$

en

$$\underline{X} = G\underline{Y}\Lambda^{-1}. \tag{8}$$

Asymptotisch, voor $N_{obs} \to \infty$ geldt

$$E(\underline{Y}) \to Y, \tag{9a}$$

$$E(\underline{\Lambda}^2) \to \Lambda^2, \tag{9b}$$

$$E(\underline{X}) \to X, \tag{9c}$$

$$N_{obs}\, C(\underline{\lambda}_s^2,\underline{\lambda}_t^2) \to \sum_{\ell=1}^{n} \pi_\ell \lambda^2_{s.\ell}\lambda^2_{t.\ell} \qquad (s,t=1,\ldots,r) \tag{10a}$$

$$N_{obs}\, C(\underline{y}_{is},\underline{y}_{it}) \to \sum_{\ell=1}^{n} \pi_\ell y_{is.\ell}y_{it.\ell} - \tfrac{1}{4}y_{is}y_{it} \qquad (s,t=1,\ldots,r;\ i=1,\ldots,K) \tag{10b}$$

$$N_{obs}\, C(\underline{x}_{js},\underline{x}_{jt}) \to \sum_{\ell=1}^{n} \pi_\ell x_{js.\ell}x_{jt.\ell} - \tfrac{1}{4}x_{js}x_{jt} \qquad (s,t=1,\ldots,r;\ j=1,\ldots,n) \tag{10c}$$

Verdere uitwerkingen laten we vanwege de kompleksiteit van de formules achterwege.

4.1.3 Betrouwbaarheidsgebieden van HOMALS skores

We nemen nu voor G weer een n × (Σk_j) indikator-matriks. Stel we beschouwen de 2-vektor $y_i^j = (y_{1i}^j \; y_{2i}^j)'$ van de kategorieskores voor de i^e kategorie van variabele j. De overeenkomstige steekproefgrootheid $\underline{y}_i^j = (\underline{y}_{1i}^j \; \underline{y}_{2i}^j)'$ is een konsistente schatter van y_i^j volgens (9a). Met behulp van (10b) vinden we nu schatters $\hat{\underline{\Sigma}}_i^j$ van Σ_i^j, de 2 × 2 variantie-kovariantiematriks van de kategorieskores voor de i^e kategorie van variabele j. Ook deze schatters zijn als gevolg van (9) konsistent. Volgens de centrale grenswaardestelling geldt dat

$$N_{obs}^{\frac{1}{2}} (\underline{y}_i^j - y_i^j) \xrightarrow{L} N(0, \Sigma_i^j) \text{ voor } N_{obs} \to \infty,$$

en dus

$$N_{obs} (\underline{y}_i^j - y_i^j)' (\hat{\underline{\Sigma}}_i^j)^{-1} (\underline{y}_i^j - y_i^j) \xrightarrow{L} \chi_2^2 \text{ voor } N_{obs} \to \infty.$$

Het symbool $\xrightarrow{L}$ wordt gebruikt voor konvergentie in verdeling, N(0,Σ) is een multinormaalverdeling, en χ_2^2 is chi-kwadraat met twee vrijheidsgraden. Met behulp van het laatste resultaat kunnen we voor elke kategorie een betrouwbaarheidsgebied voor de kategorieskores berekenen, als er een betrouwbaarheidsgrens gegeven is. Een analoge afleiding geldt voor de individu-skores.

4.1.4 De praktijk: ANACOR

Het programma ANACOR levert als output de geschatte varianties en kovarianties. Bij de behandeling van indikator-matriksen wordt de fraktie observaties van een bepaald profiel gebruikt als schatter van de populatiekans. Hierop zijn de berekende (ko)varianties gebaseerd.

Maar ook als we een willekeurige n × K matriks met niet-negatieve elementen als invoer gebruiken, vinden we (ko)varianties van rij en kolomskores. Waarop zijn deze gebaseerd, ofwel, in termen van 4.1.1, we analyseren nu wel ΠG, maar wat is Π en wat is G ? Dit is in te zien door ons uit 4.1.2 te herinneren dat Gu = μu, μ = u'ΠGu, en ΠGu = μπ. De schatters van de rijkansen vinden we door de rijsom te delen door de totale som:

$$\pi = \frac{\Pi G u}{u' \Pi G u}.$$

De rijen van G, die als konstant worden beschouwd, vinden we door deling door π.

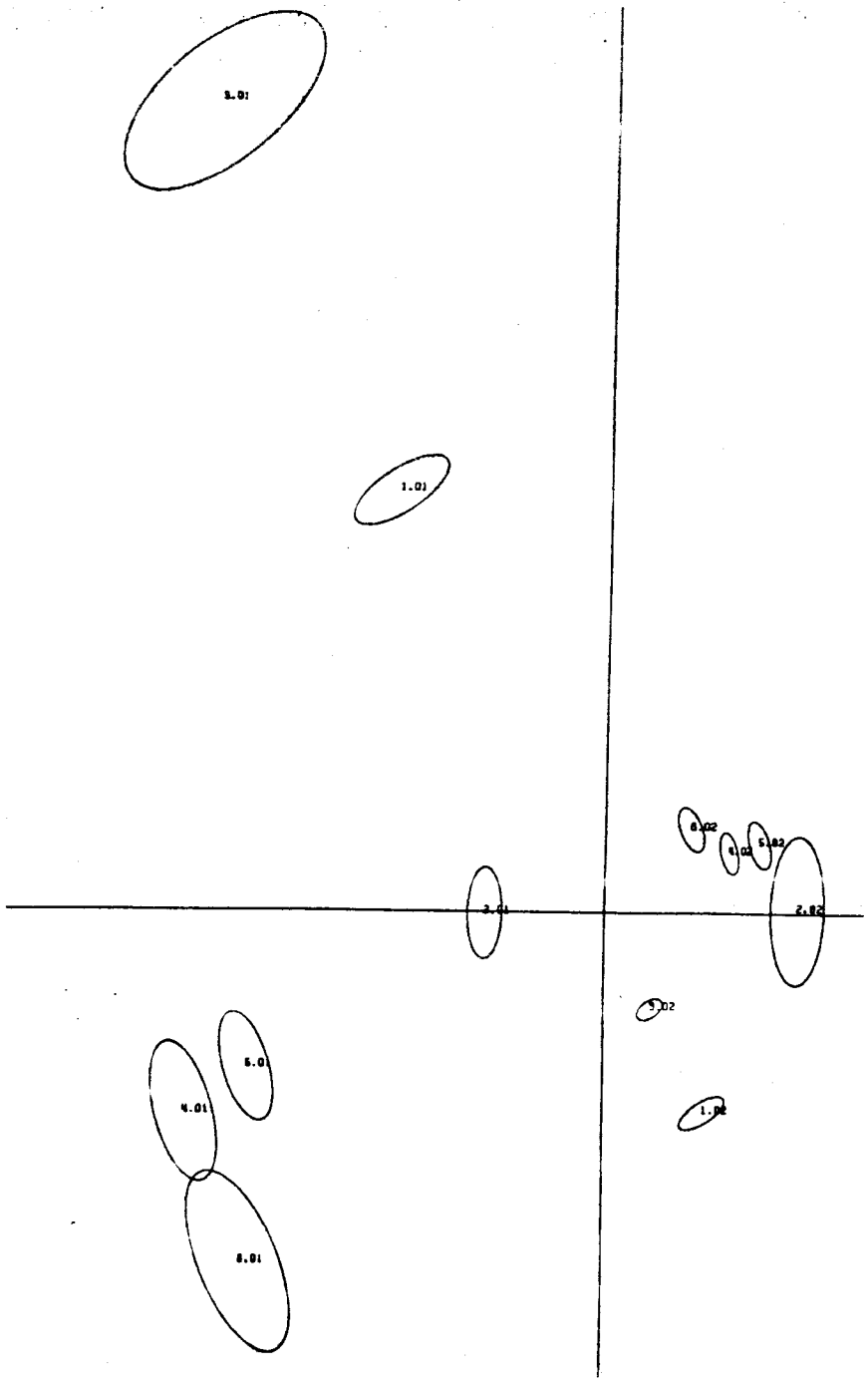
Een voorbeeld: van de matriks

1.0 0.5 0.3
0.5 1.0 0.2
0.3 0.2 1.0

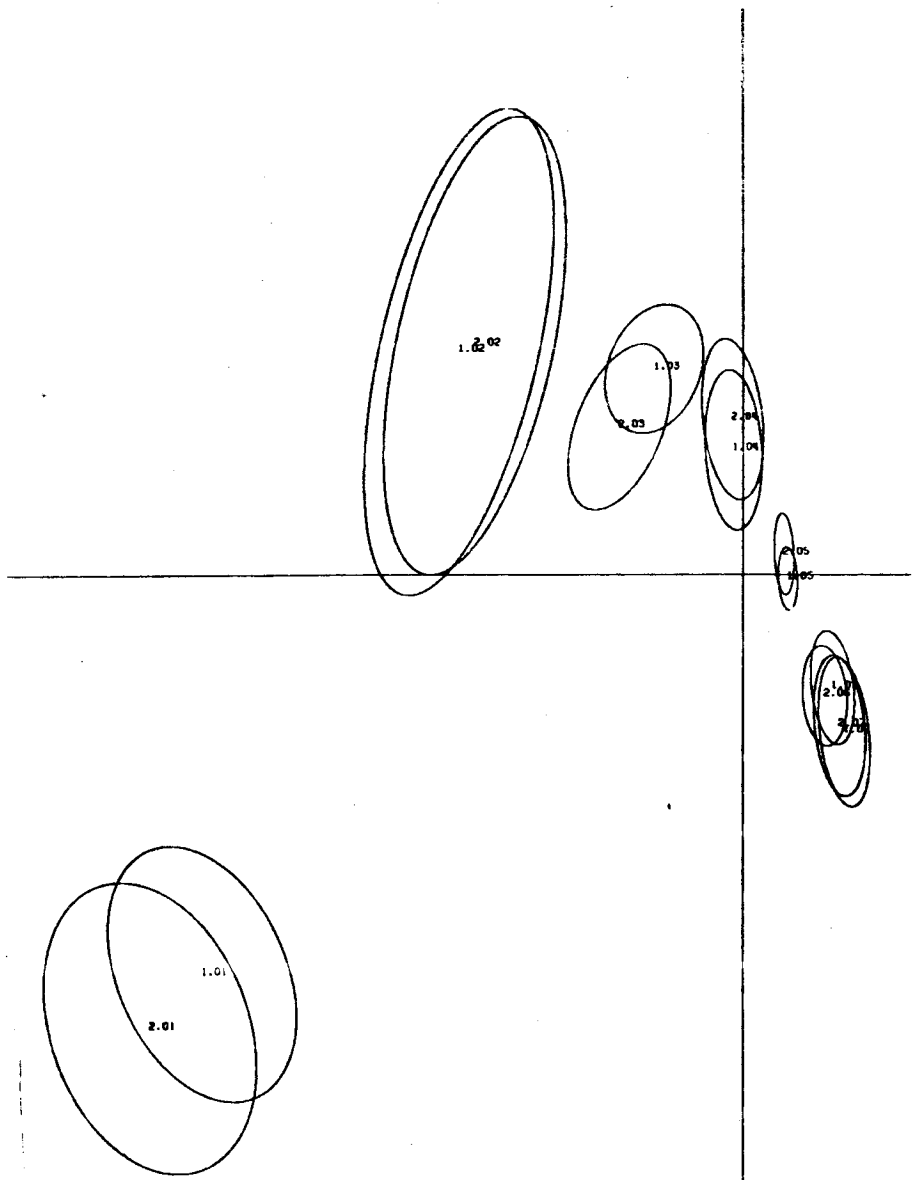
zijn de rijsommen 1.8, 1.7, en 1.5. De totale som is 5. De geschatte kansen zijn dan .36, .34, en .30, dit definieert de diagonale matriks Π. De matriks G is

2.78 1.39 0.83
1.47 2.94 0.59
1.00 0.67 3.33

en deze wordt als konstante behandeld. Omdat bij berekening van de varianties het aantal onservaties een rol speelt (zie formule (10)), moet bij het gebruik van



figuur 4.1: Japanners in ellipsen.



figuur 4.2: sociale mobiliteit in ellipsen.

ANACOR-S voor het analyseren van willekeurige matrices dit aantal als parameter worden meegegeven. Bij het analyseren van indikator-matrices volgt het aantal observaties uit de data.

4.1.5 Voorbeelden

a: Godsdienst in Japan

In figuur 4.1 zijn de 95%-betrouwbaarheidsgebieden geplot, gebaseerd op een χ^2_2-benadering. Alleen de kategorieen zijn geplot met als label: voor de punt het nummer van de variabele, na de punt het nummer van de kategorie. We zien aan de figuur:

- de betrouwbaarheidsgebieden van de twee kategorieen van iedere variabele zijn vermenigvuldigingen ten opzichte van de oorsprong van elkaar. Dat wil zeggen: neem een punt (z_{11},z_{12}) in het betrouwbaarheidsgebeid van kategorie 1 van een variabele, dan is $(-z_{11}\alpha_2/\alpha_1,-z_{12}\alpha_2/\alpha_1)$ het overeenkomstige punt in het betrouwbaarheidsgebied van kategorie 2. Hierin is α_1 (α_2) de mariginale frekwentie van kategorie 1 (2). Anders gesteld: van de betrouwbaarheids-ellipsen van een variabele zijn de assen parallel, en de lengtes van de assen zijn omgekeerd evenredig met de marginale frekwenties.
- De voornaamste konklusie uit de betrouwbaarheden is dat het samengaan van variabelen 1 en 3 enerzijds, en van 4,5, en 6 anderszijds, dat we in hoofdstuk 3.2 al uit een analyse van de kategoriepunten destilleerden, stabiel is.

b: Sociale mobiliteit

Ook hier zijn in 4.2 de 95% betrouwbaarheidsgebieden geplot. We zien ook hier dat zelfs als punten instabiel liggen (beroepsgroep 2) toch de globale struktuur uit de oorspronkelijke analyse behouden blijft. Alleen overlappen de betrouw-baarheidsgebieden van groep 6 en 7 elkaar. Zouden we het aantal beroepsgroepen kleiner willen maken, dan komt samenvoeging van 6 en 7 het eerst in aanmerking.

4.2 Diskretisering van kontinue variabelen

4.2.1 HOMALS en PCA

Stel we hebben m kontinue stochastische variabelen $\underline{x}_1,\ldots,\underline{x}_m$ met verwachting 0 en variantie 1; hun korrelatiematriks is R. Bij PCA zoeken we oplossingen van de vergelijking

$$Rz = m\lambda z,\ z'z = 1. \tag{11}$$

We gaan nu $\underline{x}_1,\ldots,\underline{x}_m$ diskretiseren. Hieronder verstaan we: voor $j=1,\ldots,m$ kiezen we een aantal kategorieën k_j en getallen $a^j_1 < \ldots < a^j_{k_j-1}$. We noteren verder $a^j_0 \triangleq -\infty$ en $a^j_{k_j} \triangleq +\infty$. De kans dat variabele j in interval υ valt is

$$d^j_\upsilon \triangleq P(a^j_{\upsilon-1} < \underline{x}_j \leq a^j_\upsilon) \quad (j=1,\ldots,m;\ \upsilon=1,\ldots,k_j)$$

en de kans dat simultaan geldt dat variabele j in interval υ en variabele ℓ in interval ω valt is

$$c^{j\ell}_{\upsilon\omega} \triangleq P(a^j_{\upsilon-1} < \underline{x}_j \leq a^j_\upsilon \ \&\ a^\ell_{\omega-1} < \underline{x}_\ell \leq a^\ell_\omega)$$

$$(j,\ell=1,\ldots,m;\ \upsilon=1,\ldots,k_j;\ \omega=1,\ldots,k_\ell).$$

Merk op dat als $j = \ell$ geldt $c_{\upsilon\upsilon}^{jj} = d_\upsilon^j$ en $c_{\upsilon\omega}^{jj} = 0$ als $\upsilon \neq \omega$. We nemen de $c_{\upsilon\omega}^{j\ell}$ samen in matriksen $C_{j\ell}$, de d_υ^j in diagonaalmatriksen D_j, de matriksen C_j in de supermatriks C met als diagonaalblokken de D_j, en de D_j in een diagonaalmatriks D. Op C en D kunnen we HOMALS toepassen: bepaal de oplossingen van

$$Cy = m\mu Dy,\ y'Dy = 1,\ y'Du = 0, \tag{12}$$

met y een vektor van k_j elementen; $y = (y_1,\ldots,y_m)$ en $y_j = (y_1^j,\ldots,y_{k_j}^j)$ voor $j=1,\ldots,m$. We zoeken nu het verband tussen PCA (11) en HOMALS (12).

Met $p_j(x)$ geven we de dichtheid van de verdeling van de j^e variabele in punt x aan, met $p_{j\ell}(x,y)$ de dichtheid van de bivariate verdeling van variabele j en ℓ in het punt (x,y). We gaan alle p_j en $p_{j\ell}$ benaderen door stapfunkties. Kies getallen u_υ^j in de intervallen $(a_{\upsilon-1}^j, a_\upsilon^j)$ voor alle j en υ zo, dat met $\delta u_\upsilon^j \triangleq a_\upsilon^j - a_{\upsilon-1}^j$ $(\upsilon\neq 1,\ \upsilon\neq k_j)$ de volgende benaderingen gelden

$$c_{\upsilon\omega}^{j\ell} \sim p_{j\ell}(u_\upsilon^j, u_\omega^\ell)\delta u_\upsilon^j \delta u_\omega^\ell, \tag{13a}$$

$$d_\upsilon^j \sim p_j(u_\upsilon^j)\delta u_\upsilon^j. \tag{13b}$$

We kiezen verder ook getallen $u_1^j < a_1^j$, $u_{k_j}^j > a_{k_j-1}^j$, δu_1^j, en $\delta u_{k_j+1}^j$ zodanig dat (13) ook voor de staarten van de verdelingen geldt. Voor k_j voldoende groot, de intervalgrenzen goed gekozen, en de u^j goed gekozen (wat "goed" hier betekent zullen we later zien) kunnen we de samenhang van (11) en (12) laten zien. Zij z een oplossing van (11). Neem aan dat de regressie lineair is, en definieer

$$y_\upsilon^j \triangleq z_j u_\upsilon^j \qquad (j=1,\ldots,m;\ \upsilon=1,\ldots,k_j)$$

Als we dit invullen in (12), en we gebruiken (13) vrijuit, dan vinden we

$$Cy \sim m\mu Dy, \text{ waarbij } \mu = \lambda.$$

Door substitutie van $z_j u^j$ voor y^j is gemakkelijk na te gaan dat $y'Dy = m$, en $y'Du = 0$. We kunnen zo m oplossingen vinden van (12), door uit te gaan van m oplossingen van (11). Maar (12) heeft meer dan m oplossingen. Deze hangen af van de verdeling van $\underline{x}_1,\ldots,\underline{x}_m$.

We bekijken één bepaalde verdeling, die met name in hoofdstuk 2 al uitvoerig geanalyseerd is. Stel $\underline{x}_1,\ldots,\underline{x}_m$ zijn multinormaal verdeeld. Definieer de matriks $R^{(2)}$ als volgt

$$r_{j\ell}^{(2)} = r_{j\ell}^2 \qquad (j,\ell=1,\ldots,m).$$

Als $R^{(2)}z^{(2)} = m\lambda z^{(2)}$, $z^{(2)\prime}z^{(2)} = 1$, dan is y, met $y_\upsilon^j = \frac{1}{2}\sqrt{2}((u_\upsilon^j)^2 - 1)z_j$ een oplossing van (12) met $\mu = \lambda$. Het bewijs loopt geheel analoog. Het verband tussen y en u worden gegeven door de Hermite-Chebyshev polynoom van de graad twee. Bij $R^{(s)}$, met elementen $r_{j\ell}^{(s)} = r_{j\ell}^s$ wordt het verband gegeven door het k-de Hermite Chebyshev polynoom.

4.2.2 Onder- en bovengrens voor de grootste eigenwaarde

Voor de grootste eigenwaarde bij (11) geldt

$$m\lambda = \max \{z'Rz : z'z = 1\},$$

en voor de grootste eigenwaarde bij (12) geldt

$$m\mu = \max \{y'Cy : y'Dy = 1 \ \& \ y'Du = 0\}.$$

Bij een goede diskretisering is μ een benadering voor λ, maar wat is een goede diskretisering, en hoe goed is de benadering ? Voor de beantwoording van deze vraag hebben we wat niet helemaal elementaire wiskunde nodig. Mensen die op resultaten belust zijn kunnen het inspringende stukje gevoeglijk overslaan.

Zij H een Hilbert-ruimte, S een lineaire deelruimte van H, A en B begrensde lineaire operatoren op H. We noteren het inwendig produkt op H als $(x,y) = (y,x)$; de bilineaire funkties α en β worden gedefinieerd door

$\alpha(x,y) \triangleq (x,Ay) = (Ax,y)$,

$\beta(x,y) \triangleq (x,By) = (Bx,y)$.

We nemen aan dat voor alle $x \in H$, $x \neq 0$, geldt dat $\beta(x,x) > 0$. Zij

$$\lambda \triangleq \max \{\alpha(x,x)/\beta(x,x) : x \in H\}, \tag{14}$$

$$\mu \triangleq \max \{\alpha(x,x)/\beta(x,x) : x \in S\}. \tag{15}$$

We nemen verder aan dat het eerste maximum bereikt wordt door de vektor z, en dat $\beta(z,z) = 1$. Zij

$$\varepsilon \triangleq \min \{\beta^{\frac{1}{2}}(x-z,x-z) : x \in S\}.$$

Dit minimum wordt aangenomen in u. Van u is bekend dat

$$\beta(u,u) = \beta(u,z). \tag{16}$$

Hiermee is

$$\varepsilon^2 = \beta(u-z,u-z) = \beta(u,u) - 2\beta(u,z) + 1 = 1 - \beta(u,u),$$

dus $\beta(u,u) = 1 - \varepsilon^2$. Zij $v \triangleq u/(1 - \varepsilon^2)$, dan is $\beta(v,v) = 1/(1 - \varepsilon^2)$, en $\beta(z,v-z) = 0$. Bovendien kunnen we $v = z + (v - z)$ gebruiken in

$$\alpha(v,v) = \alpha(z,z) + 2\alpha(z,v-z) + \alpha(v-z;v-z).$$

Een laatste definitie

$$\tau \triangleq \inf \{\alpha(x,x)/\beta(x,x) : x \in H\}.$$

We bekijken nu $\alpha(v,v)$ termsgewijs.

$\alpha(z,z) = \lambda\beta(z,z) = \lambda$,

$\alpha(z,v-z) = \lambda\beta(z,v-z) = 0$,

$\alpha(v-z,v-z) \geq \tau\beta(v-z,v-z)$.

Hierin is

$$\beta(v-z,v-z) = \beta(v-u,v-u) + 2\beta(v-u,u-z) + \beta(u-z,u-z) =$$
$$= \varepsilon^4/(1 - \varepsilon^2) + 0 + \varepsilon^2 =$$
$$= \varepsilon^2/(1 - \varepsilon^2).$$

Samengevat

$\alpha(v,v) \geq \lambda + \tau\varepsilon^2/(1 - \varepsilon^2)$,

en dus

$\mu \geq \alpha(v,v)/\beta(v,v) \geq \lambda(1 - \varepsilon^2) + \tau\varepsilon^2 = \lambda - (\lambda - \tau)\varepsilon^2$.

Omdat S een deelruimte is van H geldt vanzelfsprekend dat $\mu \leq \lambda$ Als we boven- en benedengrens kombineren vinden we

$0 \leq \lambda - \mu \leq (\lambda - \tau)\varepsilon^2$.

Stel nu weer, als in 4.2.1, dat p_j en $p_{j\ell}$ dichtheden van (bivariate) verdelingen zijn van m stochasten met verwachting 0 en variantie 1. Voor H nemen we de verzameling van alle vektoren $\Phi = (\phi_1,\ldots,\phi_m)'$ met $\phi_j: \mathbb{R} \to \mathbb{R}$, waarvoor $\int\phi_j^2(x)p_j(x)dx < \infty$ $(j=1,\ldots,m)$. Als S nemen we de verzameling van alle vektoren $\Psi = (\psi_1,\ldots,\psi_m)'$ met $\psi_j: \mathbb{R} \to \mathbb{R}$ een trapfunktie, konstant op de intervallen $(a_{\upsilon-1}^j, a_\upsilon^j]$, waarbij we de a^j voorlopig vast nemen $(j=1,\ldots,m;\ \upsilon=1,\ldots,k_j-1)$. Trapfunkties zijn kwadratisch integreerbaar, dus S is een deelruimte van H. Voor B nemen we de identiteit, en A is gedefinieerd door

$$(A\Phi)_j(x) = \phi_j(x) + p_j^{-1}(x) \sum_{\ell\neq j}^{m} \int\phi_\ell(x)p_{j\ell}(x,y)dy.$$

We zijn met dit instrumentarium in staat de grootste eigenwaarden van (11) en (12) te vergelijken. Uit de eerder gegeven definities van α en β volgt

$$\beta(\Phi,\Phi) = \sum_{j=1}^{m} \int\phi_j^2(x)p_j(x)dx,$$

$$\alpha(\Phi,\Phi) = \sum_{j=1}^{m} \int\phi_j^2(x)p_j(x)dx + \sum_{j=1}^{m}\sum_{\ell\neq j}^{m} \iint\phi_j(x)\phi_\ell(y)p_{j\ell}(x,y)dxdy.$$

Neem R, z, en λ als in (11), met λ maksimaal. Neem $\phi_j(x) = z_j x$ $(j=1,\ldots,m)$. Dan is

$\beta(\Phi,\Phi) = 1$,

$\alpha(\Phi,\Phi) = z'Rz = m\lambda$.

Omdat de a_υ^j vast verondersteld zijn kunnen we de trapfunkties weergeven door een supervektor $y = (y_1,\ldots,y_m)'$, met $y_j = (y_1^j,\ldots,y_{k_j}^j)'$, waarin y_υ^j de funktiewaarde op het interval $(a_{\upsilon-1}^j, a_\upsilon^j]$ is $(j=1,\ldots,m;\ \upsilon=1,\ldots,k_j)$. Geven we de vektor van trapfunkties in S die bij y hoort aan met Ψ_y, dan moeten we om ε te vinden het minimum over y zoeken van

$$\beta(\Phi-\Psi_y,\Phi-\Psi_y) = \sum_{j=1}^{m}\sum_{\upsilon=1}^{k_j} \int_{a_{\upsilon-1}^j}^{a_\upsilon^j} (z_j x - y_\upsilon^j)^2 p_j(x)dx.$$

De mensen die het inspringende stuk overgeslagen hebben heten we hier weer van harte welkom. We beginnen met een korte samenvatting van wat ze gemist hebben.

Bij een gegeven diskretisering (dus zowel aantal kategorieen k_j als diskretisatiepunten a_υ^j gegeven) bepalen we

$$\varepsilon^2 = \min \{ \sum_{j=1}^{m} \sum_{\upsilon=1}^{k_j} \int_{a_{\upsilon-1}^j}^{a_\upsilon^j} (z_j x - y_\upsilon^j)^2 p_j(x)\,dx : y\} \qquad (17)$$

en dan is

$$0 \leq \lambda - \mu \leq \lambda\varepsilon^2.$$

Is alleen het aantal kategorieen per variabele gegeven dan minimaliseren we (17) zowel over de y_υ^j als over de diskretisatiepunten a_υ^j, door afwisselend te minimaliseren over y en a (zowaar een alternerend kleinste kwadraten algoritme). Voor vaste a_υ^j is de optimale y_υ^j gegeven door

$$y_\upsilon^j = z_j E_\upsilon(\underline{x}_j),$$

waarbij $E_\upsilon(\underline{x}_j)$ de verwachte waarde van $\underline{x}_j$ is, gegeven dat $\underline{x}_j$ tussen a_υ^j en $a_{\upsilon-1}^j$ ligt (een voorwaardelijke verwachting dus). Voor vaste y_υ^j zijn de minimaliserende a_υ^j gegeven door

$$a_\upsilon^j = (y_\upsilon^j + y_{\upsilon+1}^j)/2z_j.$$

Voordat we wat voorbeelden geven moet we eerst iets zeggen over de literatuur. Die valt in drie delen uiteen. Er is statistische literatuur zoals Sheppard (1898) en Cox (1957). Daarnaast is er zeer veel literatuur over 'Quantization', dat wil zeggen over het diskretiseren van kontinue signalen, in de kommunikatie literatuur. We geven een paar van de belangrijkste referenties: Max (1960), Roe (1964), Wood (1969), Elias (1970), Gish en Pierce (1968), Sharma (1978), Gersho (1979). Een groot deel van de literatuur gaat over de asymptotische benadering van ε^2 als het aantal kategorieen zeer groot wordt. Voor onze doeleinden is het voldoende te konstateren dat als $k_j \to \infty$ dan $k_j^2\varepsilon^2 \to C$, een konstante die van $p(x)$ afhangt. Het derde type literatuur dat een groot aantal relevante resultaten bevat is de literatuur over numerieke integratie.

Voorbeeld: voor een standaardnormaalverdeelde variabele die in k kategorieen verdeeld wordt is het rekenwerk al gedaan door Max. Hij geeft tabellen waarin de optimale a_υ^j en y_υ^j, tezamen met de optimale ε^2, vermeld staan. In die tabellen vinden we bijvoorbeeld

k	ε^2/z^2	diskretisatiepunten
2	.3634	0
3	.1902	$\pm$ .6120
4	.1175	0, $\pm$ .9816
5	.0799	$\pm$ .3823, $\pm$ 1.244
10	.0229	0, $\pm$ .4047, $\pm$ .8339, $\pm$ 1.3246, $\pm$ 1.9678.

Voor dit speciale geval geven we ook even de asymptotische formules, die al voor zeer kleine k aardig op blijken te gaan. Asymptotisch is ε^2/z^2 gelijk aan $2.73k/(k + .853)^3$, en is y_υ^j gelijk aan $2.449\ \mathrm{erf}^{-1}(2\upsilon - k)/(k + .853)$, erf(x) is de fout-funktie (Abramowitz en Segun, 1965, hoofdstuk 7).

Hebben we nu m standaardnormaal verdeelde variabelen $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_m$, die we allemaal optimaal in k kategorieen verdelen, en is $Rz = m\lambda z$, $z'z = 1$, dan volgt uit het bovenstaande dat de som van de $\varepsilon_k^2 z_j^2$ onafhankelijk is van z, immers

$$\varepsilon_k^2 = \sum_{j=1}^{m} \varepsilon_k^2 z_j^2.$$

Voor de grootste eigenwaarde van het HOMALS probleem geldt dus

$$0 \leq \lambda - \hat{\mu} \leq \lambda\varepsilon_k^2,$$

waarbij ε_k^2 direkt uit de tabellen van Max kan worden afgelezen.

Bij 2 variabelen en 3 intervallen hebben we HOMALS gedraaid met waarden van λ gelijk aan .55, .75, en .95, door de variabelen een korrelatie van .1, .5, en .9 te geven. PCA met twee variabelen is natuurlijk triviaal. We geven de ppopulatiewaarde van μ, en de steekproefwaarde $\hat{\mu}$ bij een steekproef van 100 observaties.

r	λ	$\lambda\varepsilon_3^2$	μ	$\lambda - \mu$	$\hat{\mu}$
.1	.55	.2092	.5406	.0094	.5744
.5	.75	.2853	.7060	.0440	.7461
.9	.95	.3614	.8904	.0596	.8984

Nog een voorbeeld. In feite het eenvoudigste wat maar denkbaar is. We veronderstellen dat x uniform verdeeld is. Indit geval maakt het alternerende kleinste kwadraten algoritme y_u^j steeds gelijk aan het middelpunt van het interval, terwijl a_u^j in de andere stap ook gelijk gemaakt wordt aan het gemiddelde van de twee omliggende elementen van y_j. Uiteindelijk konvergeert dit dus naar een situatie waarin de intervallen allemaal gelijk zijn, en waarin de waarden van y_j precies in het midden van de intervallen liggen. We vinden bovendien $\varepsilon_k^2/z^2 = 1/k^2$.

4.2.3 Het samennemen van kategorieen.

Wat gebeurt er met de grootste eigenwaarde en met de kategorie-skores als we verschillende kategorieen samennemen. In de matriks algebra is er veel over gelijksoortige onderwerpen gepubliceerd, en de resultaten zijn op PCA en ANACOR toegepast door Escofier en Le Roux (1972, 1977). Wij passen hier de analyse van de vorige paragraaf toe op het voorbeeld "sociale mobiliteit", waarin we zagen dat beroepsgroepen 6 en 7 dicht bij elkaar lagen. We bekijken eerst wat er gebeurt als we eisen dat twee kategorieen gelijke skore hebben. Daarvoor springen we weer even in.

We refereren aan de afleiding in 4.2.2 over lineaire operatoren in Hilbert ruimten. Als H nemen we nu de verzameling vektoren met Σk_j elementen waarvoor $y'Du = 0$, als S de deelverzameling van H waarvoor geldt dat $y_u^j = y_\omega^\ell$ (niet noodzakelijk $j = \ell$). Voor β nemen we het inwendig produkt gewogen voor D

$$\beta(y,y) = y'Dy,$$

en analoog

$$\alpha(y,y) = \frac{1}{m} y'Cy.$$

Dus

$$m\lambda = \max \{y'Cy : y'Dy = 1 \ \& \ y \in H\},$$

$$m\mu = \max \{y'Cy : y'Dy = 1 \ \& \ y \in S\}.$$

Stel $\hat{y}$ is de optimale oplossing van het eerste probleem. We moeten om ε^2 uit te kunnen rekenen $(y - \hat{y})'D(y - \hat{y})$ minimaliseren over y in S, dus met $y_\upsilon^j = y_\omega^\ell$. Als het minimum wordt bereuikt in u, dan zijn

$$u_\upsilon^j = u_\omega^\ell = \frac{d_\upsilon^j \hat{y}_\upsilon^j + d_\omega^\ell \hat{y}_\omega^\ell}{d_\upsilon^j + d_\omega^\ell};$$

en de overige elementen van u zijn gelijk aan de overeenkomstige van $\hat{y}$. Het minimum is

$$\varepsilon^2 = \frac{d_\upsilon^j d_\omega^\ell}{d_\upsilon^j + d_\omega^\ell} (\hat{y}_\upsilon^j - \hat{y}_\omega^\ell)^2.$$

Even samenvatten. Stel $\hat{y}$ is de optimale oplossing van

$$Cy = m\lambda Dy, \ y'Dy = 1, \ y'Du = 0, \ \lambda \text{ maksimaal.}$$

We zoeken de oplossing van

$$Cy = m\mu Dy, \ y'Dy = 1, \ y'Du = 0, \ y^j = y \ , \ \mu \text{ maksimaal.} \qquad (18)$$

Dan geldt met

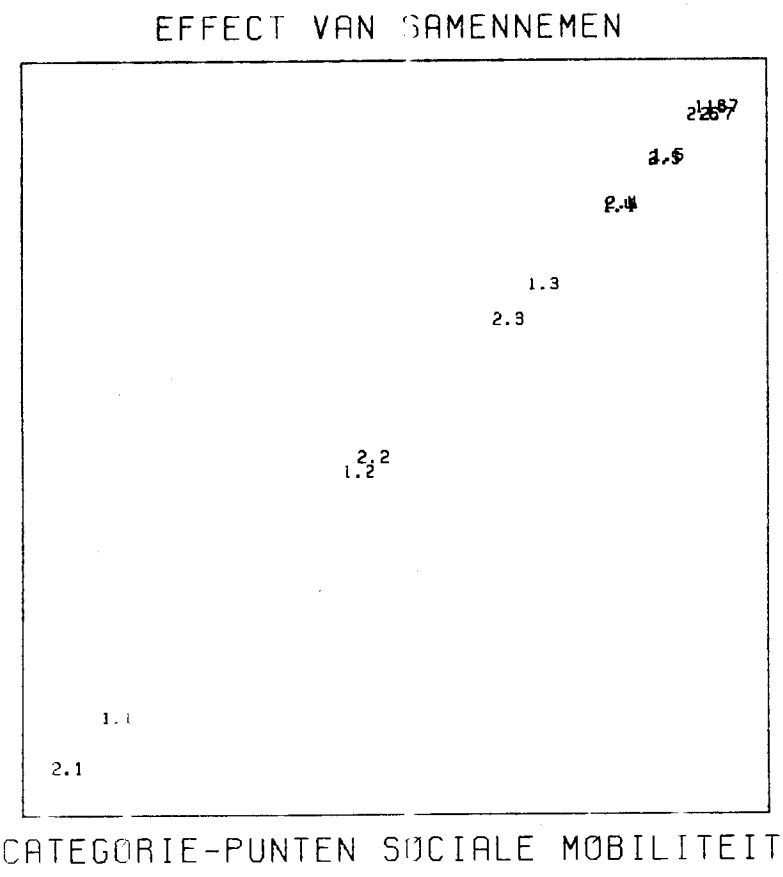
$$\varepsilon^2 = \frac{d_\upsilon^j d_\omega^\ell}{d_\upsilon^j + d_\omega^\ell} (\hat{y}_\upsilon^j - \hat{y}_\omega^\ell)^2,$$

dat

$$0 \leq \lambda - \mu \leq \lambda\varepsilon^2.$$

Als we in (18) $j = \ell$, $\upsilon \neq \omega$, nemen, dan eisen we dat twee kategorieen van één variabele gelijke skoring krijgen. Dit is niet identiek aan het samennemen van twee kategorieen tot één, omdat in het laatste geval C en D veranderen (ze krijgen bijvoorbeeld een rij en een kolom minder). De oplossingen zijn echter wel ekwievalent.

In het voorbeeld "sociale mobiliteit" is $\lambda = .7628$. We kunnen afleiden dat $\varepsilon^2 = .00065$, dus $0 \leq \lambda - \mu \leq .0005$, ofwel $.7623 \leq \mu \leq .7628$. De analyse op de samengevoegde data gaf $\mu = .7627$. We vergelijken ook de kategorieskores. In figuur 4.3 staan horizontaal de kategorieskores bij 7 kategorieen, en vertikaal de kategorieskores bij 6 kategorieen.


EFFECT VAN SAMENNEMEN
1.3
2.3
1.2
2.2
1.1
2.1
CATEGORIE-PUNTEN SOCIALE MOBILITEIT

figuur 4.3

5. PRINCALS

5.1. PRINCALS geometries

5.1.1. Inleiding en terminologie

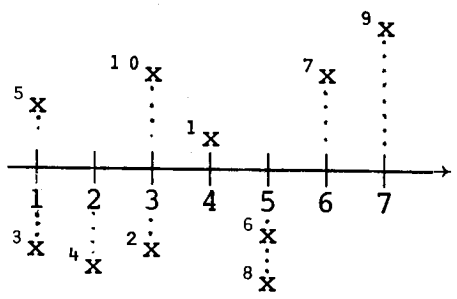
De geometrie van Niet Lineaire Principale Componenten Analyse en van Niet Metrische Principale Componenten Analyse (NLPCA & NMPCA) laat zich gemakkelijk formuleren aan de hand van geometrische representaties van één variabele en n individupunten in een gemeenschappelijke euklidische ruimte. Deze afbeeldingen stellen per variabele de optimale representatie volgens het model voor en kunnen op dezelfde wijze voor elke variabele gemaakt worden.

Er wordt steeds onderscheid gemaakt tussen perfekte fit, de situatie waarin de modeleisen perfekt op de gegevens passen, en imperfekte fit, waarbij de modeleisen zo goed mogelijk benaderd worden. Dit benaderen gaat gepaard met verlies, gemeten door een verliesfunktie. We gaan uit van kategorische gegevens. Dus elke variabele kan slechts een beperkt aantal diskrete waarden aannemen. Deze waarden kunnen gediskretiseerde intervallen op een kontinuum zijn zoals bv. inkomensklassen, maar ook echte diskrete kategorieën zoals bv. sex, religie of beroep. Mochten meerdere individuen in dezelfde kategorie vallen of skoren dan is er sprake van een " tie ".

5.1.2. Principale Componenten Analyse

PCA geeft de datamatrix weer als n punten en m vektoren in een p-dimensionale ruimte ($p << m$). De n punten korresponderen met n individuen of rijen en de m vektoren met m variabelen of kolommen. Dit is van huid uit een m-dimensionaal probleem ($n>m$), waarbij de datamatrix gereduceerd wordt tot rang p, zodat de n individupunten in een p-dimensionale deelruimte liggen. De relatie tussen de individupunten en de variabelerichtingen is zodanig, dat de projekties van de individupunten op zo'n variabelerichting proportioneel zijn met de overeenkomstige kolom van de datamatrix. Zie figuur 5.1.

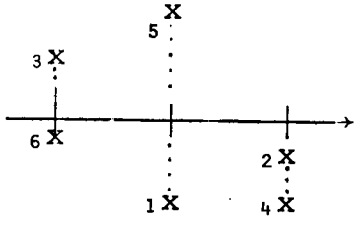
Een dergelijke relatie is gedefinieerd voor elke variabele ten opzichte van alle n individupunten. In geval van perfekte fit en aaneensluitende datawaarden houdt dit in, dat de projekties op de variabelerichtingen voor elke variabele, i.e. kolom uit de datamatrix, onderling op overeenkomstige afstanden van elkaar in de volgorde van de korresponderende datavektor moeten liggen.



$$\begin{bmatrix} 4 \\ 3 \\ 1 \\ 2 \\ 1 \\ 5 \\ 6 \\ 5 \\ 7 \\ 3 \end{bmatrix}$$

figuur 5.1 PCA representatie van een datavektor.

In die situaties waar geen perfekte fit mogelijk is, hetgeen bijna altijd het geval is, wordt de proportionaliteits-eis geweld aangedaan. Voorzover ze verkeerd staan is er verlies. Zijn er ties in de datavektor dan vallen de overeenkomstige projekties samen. In de klassieke PCA situatie zijn er dikwijls evenveel kategorieën als individuen en deze kategorieën worden als projekties van individuen enkelvoudig weergegeven. Bij kategorische data zijn er meestal veel minder kategorieën dan individuen. Er zijn dan veel ties en dus samenvallende projekties. Dit houdt in dat overeenkomstige individupunten op een hypervlak moeten liggen en elk zo'n hypervlak representeert een " tie-block ", ofwel een kategorie. Deze hypervlakken zijn parallel en zij staan loodrecht op de bijbehorende variabele richting.



$$\begin{bmatrix} 2 \\ 3 \\ 1 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

figuur 5.2 PCA hypervlak representatie van een datavektor.

De eisen gesteld aan de projekties van individupunten, worden bij data met veel ties aan deze hypervlakken gesteld. De snijpunten van hypervlakken met de variabelerichting moeten proportioneel zijn met de vektor van kategorie-nummers. De genoemde snijpunten vatten wij op als kategoriekwantifikaties.

5.1.3. HOMALS

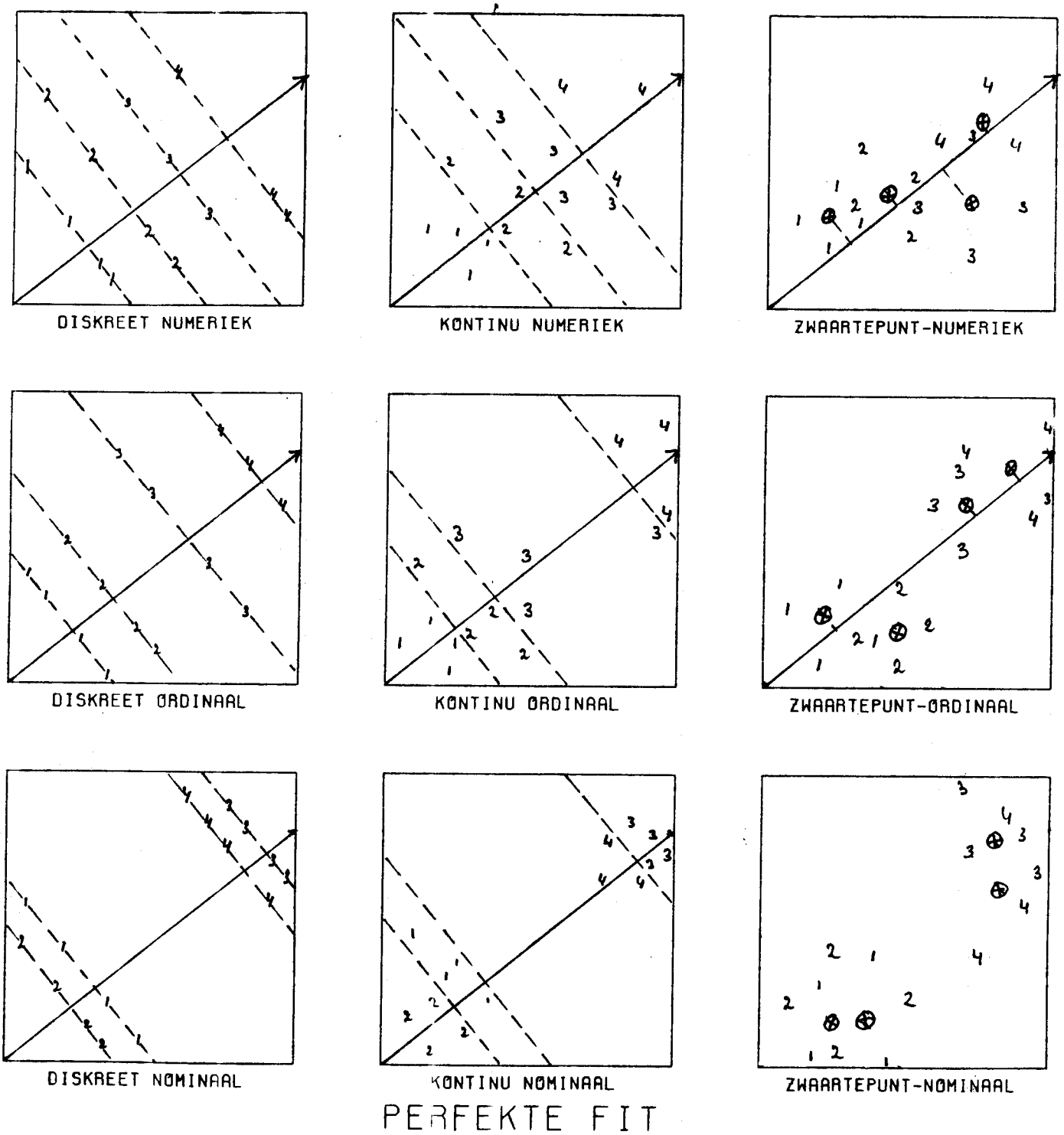
Voor de volledigheid herhalen wij hier enige punten uit hoofdstuk 3.1.

HOMALS geeft een kategorische datamatriks weer als n individupunten en Σ_{k_j} kategoriepunten in een p-dimensionale euklidische ruimte met $p << (\Sigma_{k_j} - m)$. De dimensionaliteit van dit probleem is $\Sigma_{k_j} - m$ en de datamatriks wordt benaderd door een matriks met lagere rang p. Kategoriepunten en individupunten moeten voldoen aan de eis dat een kategoriepunt het zwaartepunt is van de punten van individuen die in die kategorie skoren. Als een dergelijke groep van individupunten samenvalt met het bijbehorende kategoriepunt is er sprake van perfekte fit. Liggen de individupunten er om heen dan is er sprake van imperfekte fit en het verlies is de som van de gekwadrateerde afstanden van de individupunten tot het kategoriepunt. In HOMALS gaan wij er impliciet van uit dat er per variabele veel minder kategorieën dan observaties zijn en dat deze kategorieën meervoudig diskreet gekwantificeerd worden. Er worden geen variabelen gekwantificeerd, maar kategorieën, die korresponderen met groepen van individuen. Als er evenveel individuen als kategorieën zijn voor een variabele, wordt er altijd aan de HOMALS eis voldaan omdat de kategoriepunten dan gelijk zijn aan de individupunten.

5.1.4. PRINCALS = HOMALS & PCA

Niet Lineaire PCA, zoals tot nu toe behandeld, wordt gekenmerkt door meervoudige kategoriekwantifikatie, waarvan HOMALS een speciaal geval is met een identiteitsrelatie tussen kategorieën en hun kwantifikaties. Wat anderen Niet Metrische PCA noemen (Kruskal en Shepard, 1974) betekent het afzwakken van de proportionaliteitseis, terwijl de enkelvoudige kwantifikatie van kategorieën wordt toegepast. (De Leeuw en Van Rijckevorsel, 1980) PRINCALS is de naam van een komputerprogramma dat de diskrete opties van zowel NMPCA als NLPCA in één algoritme verenigt.
In de hierna volgende paragrafen leggen wij uit hoe de geometrie van PCA met zwakkere restrikties leidt tot speciale gevallen van HOMALS. Of, anders geformuleerd, hoe de geometrie van HOMALS met sterkere restrikties leidt tot NMPCA. Tevens wordt uitgelegd hoe dit afzwakken respektievelijk versterken van restrikties geometries voorgesteld kan worden.

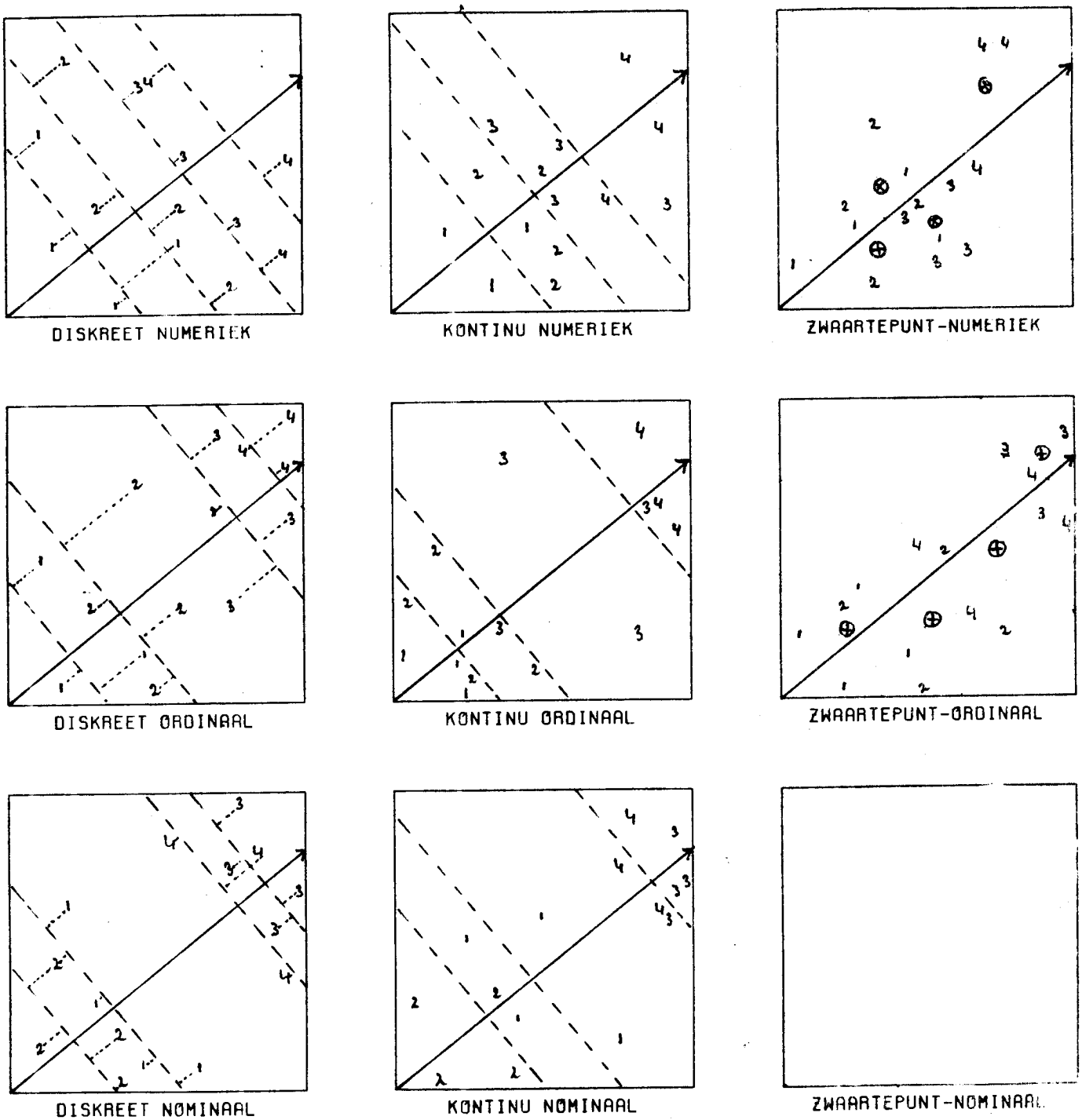
We kunnen de proportionaliteits-eis van PCA zwakker maken, zodat het model algemener wordt, door te eisen dat per variabele de kategorie-kwantifikaties slechts monotoon met de datavektor moeten zijn en dat zij niet op gelijke afstanden van elkaar hoeven te liggen. Een dergelijk meetnivo noemen wij ordinaal. Zo geformuleerd is het meetnivo in de klassieke PCA numeriek. Numerieke oplossingen voldoen altijd aan de ordinale eisen, vandaar dat deze ordinale restrikties zwakker genoemd worden. Nog verder generaliserend kunnen we ook de ordinaliteits-eis laten vallen en alleen maar identiteit eisen: punten van individuen in dezelfde kategorie moeten op een hypervlak liggen. De kategorieën worden afgebeeld op een enkelvoudige nominale schaal. Het meetnivo is nominaal. Ordinale en numerieke oplossingen voldoen altijd aan de laatstgenoemde restrikties die daarom de zwakste genoemd worden.
Let op, dat de eis dat individuen die bij dezelfde kategorie horen op een hypervlak moeten liggen, onverminderd van kracht is binnen al de drie benaderingen. De afgezwakte proportionaliteits-restrikties hebben allen betrekking op relaties tussen kategorieën. De behandeling van individupunten binnen een kategorie, i.e. de behandeling van ties, is tot nu toe onveranderd gebleven, de getiede individupunten moeten nl. op het bijbehorende hypervlak liggen. Dit heet de diskrete benadering. Ook deze restriktie kunnen we afzwakken door niet te eisen dat de getiede individupunten op een hypervlak moeten liggen, maar in een gebied begrensd door twee hypervlakken. Elke kategorie wordt gekwantificeerd in de vorm van een boven- en ondergrens op de variabelerichting. Een dergelijke behandeling van ties heet de kontinue benadering. Vaak is het voor de hand liggend om aan te nemen dat de kategorieën van de datavektor aaneensluitend zijn, hetgeen betekent dat de kategoriekwantifikaties dit ook moeten zijn. De ondergrens van de ene kategorie moet dan samenvallen met de bovengrens van de volgende kategorie. We kunnen ook op deze intervallen numerieke, ordinale of nominale schaalnivo's definiëren; de voorwaarden, die we bij diskrete schaling aan de hypervlakken oplegden, gelden nu voor de genoemde intervallen. De intervallen moeten proportioneel met de korresponderende datavektor zijn. Het is duidelijk dat de diskrete behandeling van ties altijd voldoet aan deze, daarom zwakker genoemde, kontinue eisen.



figuur 5.3 Enkelvoudige kwantifikaties, perfekte fit

De meest vrije manier van tie behandeling is de zogenaamde zwaartepunts-schaling (De Leeuw, 1977). Hierbij moeten de zwaartepunten van de individuen met dezelfde datawaarde proportioneel op de variabelerichting liggen. Puntenwolken van individuen die in verschillende kategorieën vallen mogen elkaar overlappen, als hun

gemiddelden maar aan de aangenomen schaaleisen voldoen. Het is vanzelfsprekend dat de diskrete en kontinue oplossingen altijd aan deze eisen voldoen. Zie figuur 5.3. Wij hebben in bovenstaand verhaal steeds perfekte fit voor ogen gehad om zodoende aan te kunnen tonen hoe de verschillende meetnivo's en vormen van tie behandeling formeel van elkaar



figuur 5.4 Enkelvoudige kwantifikaties, imperfekte fit.

verschillen. Bij imperfekte fit, zoals weergegeven in figuur 5.4, worden de genoemde eisen steeds benaderd, hetgeen bij de zwakkere restrikties erg veel vrijheid schept.

Bij de nominale zwaartepunts-schaling dient zich de overeenkomst met HOMALS aan. In HOMALS worden immers ook gemiddelden van getiede individuen geschaald. Alleen geldt in HOMALS de meervoudige diskrete eis dat de individupunten moeten samenvallen met het overeenkomstige kategoriepunt, terwijl bij deze vorm van PCA de enkelvoudige eis geldt dat de zwaartepunten van getiede individupunten op de goede manier op een vektor liggen. Het is duidelijk dat HOMALS voldoet aan de norm van nominale zwaartepunts-schaling in PCA.

5.2 PRINCALS & HOMALS analyties

5.2.1. HOMALS

Zoals we gezien hebben in hoofdstuk 2 en 3 krijgt in HOMALS elk individu een p-dimensionale skore, gelijk aan de som van de p-dimensionale kwantifikaties van de kategorieën waarin dit individu valt. Als er meer variabelen zijn dan willen we submatriksen van kategorie-kwantifikaties Y_j, $j=1,\ldots,m$, vinden, zodanig dat de korresponderende submatriksen van individu-skores G_jY_j, $j=1,\ldots,m$, zo homogeen mogelijk zijn. De afleiding van de definitie van homogeniteit en de maximalisatie ervan als eigenprobleem is in paragraaf 2.1 uitvoerig behandeld.
Een belangrijke eigenschap is dat perfekte fit, de situatie waarin alle variabelen de individuen indelen in dezelfde groepen, normaal gesproken niet voorkomt. De maximale homogeniteit wordt benaderd met behulp van een kwadratische verliesfunktie,

$$\sigma(X;Y) = \frac{1}{m} \Sigma_{j=1}^{m} \text{SSQ } (X - G_jY_j).$$

Vanwege het meervoudige karakter van de kategorie-kwantifikaties Y noemen we het verlies σ hier voortaan de meervoudige stress.

5.2.2. PCA

In klassieke PCA willen we de gestandariseerde datamatrix H (n x m) benaderen met een produkt van matriksen X (n x p) en A (m x p) van

lagere volle rang p met p << min(n,m)). Als p = min(n,m) dan is

$$H = XA', \qquad X'X = I \text{ en } u'X = 0.$$

Deze benadering kunnen wij in de vorm van een verliesfunktie gieten

$$\sigma_1 = \frac{1}{m} \text{SSQ } (H - XA'), \qquad X'X = I \text{ en } u'X = 0.$$

De gebruikelijke manier om onder de restriktie X'X = I dit verlies te minimaliseren is de datamatriks H te schrijven als een singuliere waarde dekompositie (SVD)

$$H = K\Lambda L',$$

met K'K = I, L'L = I en Λ diagonaal. De oplossingen voor X en A zijn

$$X = K \text{ en } A = \Lambda L'.$$

Maar als de afmetingen van H erg groot zijn is dit een kostbare manier om X en A te vinden. Bovendien zijn niet-lineaire en/of niet-metrische generalisaties niet gemakkelijk in te bouwen. De vorm van de HOMALS verliesfunktie maakt het interessant om deze als alternatieve PCA verliesfunktie te bekijken

$$\sigma_2 = \frac{1}{m} \Sigma_{j=1}^{m} \text{SSQ } (X - h_j a_j'), \quad X'X = I \text{ en } u'X = 0.$$

We kunnen laten zien dat onder de gebruikte normalisaties geldt dat

$$\sigma_2 = \sigma_1 + (p - 1),$$

dit wil zeggen dat minimaliseren van beide verliesfunkties bij PCA dezelfde oplossing geeft, omdat de verliesfunkties alleen maar een konstante verschillen.
In de uiteindelijke oplossing zijn de elementen van A korrelaties (de " ladingen ") van de m variabelen met de p principale komponenten; geometries zijn het vektoren (richtingen) in een p-dimensionale ruimte. De individu-skores liggen als punten in diezelfde ruimte. De afbeeldingen van de individuen op de richtingen in A moeten per variabele proportioneel zijn met de data.

5.2.3. PRINCALS

Het voordeel van het gebruik van σ_2 in plaats van σ_1 is dat de

parameters op een wat nettere manier gescheiden zijn, en dat we PRINCALS als een andere specifikatie van HOMALS kunnen behandelen. Anders gezegd: Takane, Young en De Leeuw (1978) gaan bij PRINCIpals uit van PCA en dus van σ_1, wij daarentegen gaan bij PRINCALS uit van HOMALS en dus van σ_2. Tot nu toe hebben we niets gezegd over optimaal schalen en wij zij er vanuit gegaan dat de matriks H een vaste niet veranderende matriks is. Wij kunnen een kolom van H, h_j, schrijven als het produkt $G_j z_j$ van de indikator matriks G_j en een vektor van kategorie-kwantifikaties z_j,

$$h_j = G_j z_j, \qquad j=1,\ldots,m.$$

De verliesfunktie ziet er als volgt uit

$$\sigma = \frac{1}{m} \Sigma_{j=1}^{m} \text{SSQ } (X - G_j z_j a_j'),$$

met de restrikties $X'X = I$, $u'X = 0$ en $z_j'D_j z_j = 1$, $j=1,\ldots,m$.

De standarisering van H hebben wij nu geformuleerd als een restriktie op z_j $(j=1,\ldots,m)$. De impliciet toegestane lineaire transformaties van de data in het klassieke PCA geval kunnen wij hier als de expliciete restrikties op de vektoren z_j, voor iedere j, formuleren dat deze vektoren in een lineaire deelruimte C_j van dimensie 1 moeten liggen, namelijk de ruimte van alle vektoren proportioneel met de kategorienummers z_j. Niet-metrische generalisaties zijn mogelijk door aan de ruimte waarin de vektoren z_j moeten liggen andere eigenschappen toe te kennen. Dit voegt de ekstra restriktie toe dat z_j in een of andere kegel C_j moet liggen, waarbij C_j bepaald wordt door het meetnivo van de data

$$z_j \in C_j, \qquad j=1,\ldots,m.$$

In de rechterterm van de verliesfunktie kunnen we $z_j a_j'$ ook schrijven als de matriks Y_j van kategorie kwantifikaties. Als we toestaan dat de rang van Y_j groter kan zijn dan 1 en we substitueren dan Y_j in de PRINCALS verliesfunktie dan hebben we de HOMALS verliesfunktie met de meervoudige kategorie-kwantifikaties Y_j.

$$\sigma_H = \frac{1}{m} \Sigma_{j=1}^{m} \text{SSQ } (X - G_j Y_j), \quad X'X = I, \; u'X = 0.$$

Indien Y_j van rang 1 moet zijn, d.w.z. geschreven kan worden als

het rang 1 produkt $z_j a_j'$, hebben we dus de PRINCALS verliesfunktie met de enkelvoudige kategorie-kwantifikaties z_j.
We zullen nu laten zien hoe we de PRINCALS verliesfunktie opsplitsen kunnen in additieve komponenten. We doen dit voor iedere variabele apart, zodat we geen indeks j nodig hebben. We beschouwen ook X zolang als vast, zodat het verlies alleen nog maar een funktie is van de twee vektoren z en a. Dus

$$\sigma(z;a) = SSQ\ (X - Gza').$$

We definieren nu $U = D^{-1}G'X$, en schrijven za' als $za' = U + (za' - U)$. Dit geeft de partitionering

$$\sigma(z;a) = SSQ\ (X - GU) + SSQ_D\ (U - za'),$$

waarbij SSQ_D de kwadratensom gewogen met de elementen van de diagonale matriks D is.
Het aantrekkelijke van deze partitionering is dat het totale PRINCALS verlies opgesplitst wordt in een linker deel gelijk aan het meervoudige HOMALS verlies en een rechterdeel wat er bij komt als we de enkelvoudige lineaire PCA restrikties toepassen op de HOMALS oplossing. Voor de meervoudige komponent geldt bovendien

$$SSQ\ (X - GU) = p - SSQ_D\ (U).$$

De enkelvoudige komponent kan nog verder opgesplitst worden. Definieer $\bar{z} = Ua/a'a$, en schrijf $z = \bar{z} + (z - \bar{z})$. Dan geldt

$$SSQ_D\ (U - za') = SSQ_D\ (U - \bar{z}a') + SSQ\ (a)SSQ_D\ (z - \bar{z}).$$

Uit deze partitionering volgt hoe we $\sigma(z;a)$ moeten minimaliseren voor vaste X en a over z zodanig dat

$$z'Dz = 1,$$

en

$$z \in C.$$

We hoeven namelijk alleen maar de term $SSQ_D\ (z - \bar{z})$ onder deze restrikties te minimaliseren. In de De Leeuw (1977) wordt aangetoond dat we dat kunnen doen door $SSQ_D\ (z - \bar{z})$ te minimaliseren over $z \in C$, en dan vervolgens de gevonden oplossing te normaliseren vol-

gens z'Dz = 1. Projekteren van z op C wordt besproken in hoofdstuk 8 onder de titel "optimale schaling".
Een analoge partitionering is mogelijk om te laten zien hoe we moeten minimaliseren over a voor gegeven X en z. Definieer $\bar{a}$ $\bar{a} = U'Dz = X'Gz$ en schrijf a als $a = \bar{a} + (a - \bar{a})$. Dan

$$SSQ_D\ (U - za') = SSQ_D\ (U - z\bar{a}') + SSQ\ (a - \bar{a}),$$

omdat $SSQ_D\ (z) = 1$. Minimaliseren over a doen we gewoon door a gelijk te maken aan $\bar{a}$. Als we deze resultaten invullen in

$$\sigma(z;*) \triangleq \min\ \{\sigma(z,a)\ :\ a\},$$

dan krijgen we

$$\sigma(z;*) = p - SSQ\ (\bar{a}),$$

waarbij $\bar{a} = U'Dz = X'Gz$. In

$$\sigma(*;a) \triangleq \min\ \{\sigma(z,a)\ :\ z'Dz = 1\ \&\ z \in C\}$$

krijgen we het iets ingewikkelder resultaat (door de normalisatie z'Dz = 1, en door de projektie op de kegel)

$$\sigma(z;*) = p + SSQ(a)(1 - 2\ SSQ_D^{\frac{1}{2}}(\hat{z}),$$

waarbij $\hat{z}$ de projektie van $\bar{z} = Ua/a'a$ op de kegel C is.
Hoe we moeten minimaliseren over X voor vaste a en z volgt uit de ongepartitioneerde verliesfunktie zelf. Definieer $\bar{X} = Gza'$. We moeten de term $SSQ\ (X - \bar{X})$ minimaliseren onder de restricties X'X = I en u'X = 0. Dit definieert een ortogonaal procrustes probleem (Cliff,1966).
Wij hebben om de overeenkomst met HOMALS te benadrukken de kwadratische vormen voor a en z via de HOMALS kategorie-kwantifikaties Y afgeleid. Voor kontinue niet-kategorische benaderingen waarbij de indicator matriks notatie niet van toepassing is, is een direkte afleiding van a en z noodzakelijk. Het PRINCALS verlies ziet er dan als volgt uit, voor vaste X uiteraard,

$$\sigma(q;a) = SSQ\ (X - qa'),$$

waarbij Q een matriks van toegestane transformaties van de datamatriks H is.

Definieer

$$\bar{q} = Xa/a'a,$$

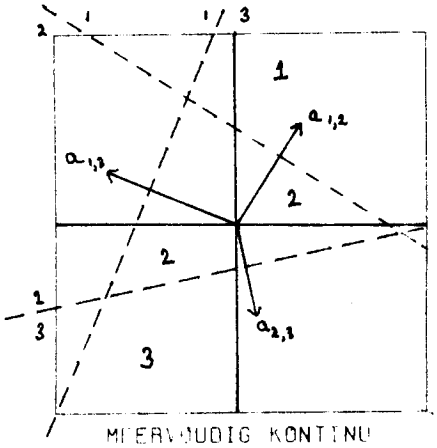
dan is de partitionering die korrespondeert met het optimale schalingsverlies voor vaste X en a met X'X = I en u'X = 0

$$\sigma(q;a) = SSQ\ (X - \bar{q}a') + SSQ\ (a)SSQ\ (q - \bar{q}).$$

De eerste term rechts van het gelijkteken wordt geminimaliseerd onder de voorwaarden q'q = 1 en q ϵ C. De laatste restriktie definieert dezelfde soort optimale schalingsproblemen als in het diskrete geval. Het zijn hier echter niet de kategorie-kwantifikaties die in een of andere deelruimte moeten liggen maar de getransformeerde datavektor q met i.h.a. veel meer elementen dan z. De overeenkomstige regressie problemen zijn dan ook van de orde n. Op dezelfde manier kunnen we $\bar{a} = X'q$ definieren, dan is de partitionering naar het verlies van de gewichten a voor vaste X en q met X'X = I, u'X = 0, q'q = 1 en q ϵ C

$$\sigma(q;a) = SSQ\ (X - a\bar{q}') + SSQ\ (a - \bar{a}).$$

Omdat op a geen restrikties rusten is het verlies weer minimaal voor $a = \bar{a}$.
De twee schalingsmogelijkheden waar we het nog niet over gehad hebben zijn de enkelvoudige zwaartepunts-schaling en de meervoudige kontinue schaling.
Meervoudige kontinue schaling kan bijvoorbeeld niet met behulp van de HOMALS verliesfunktie beschreven worden en daarom gebruiken wij de PRINCALS notatie.



MEERVOUDIG KONTINU

Indikator matriks

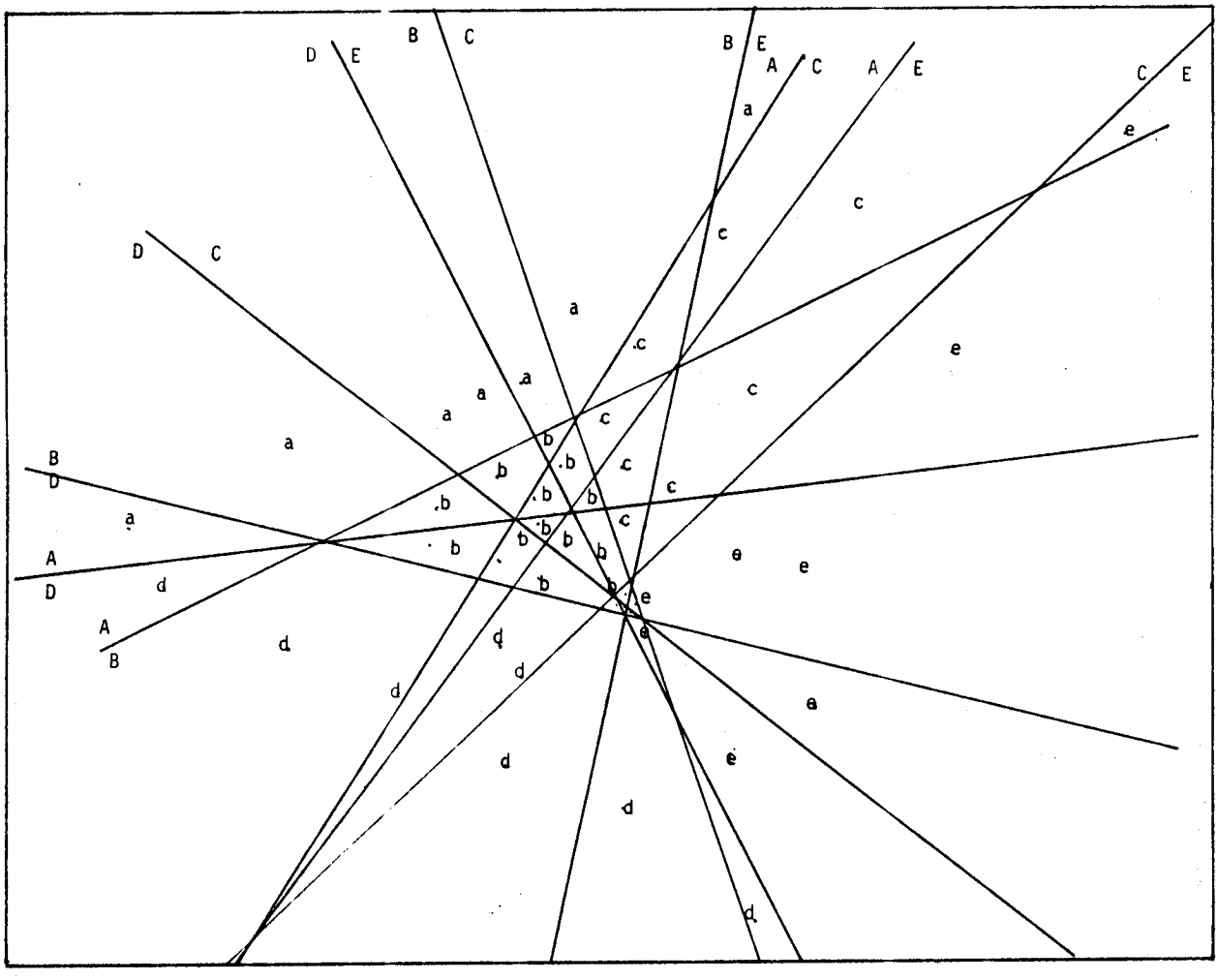
1	2	3
1	0	0
0	1	0
0	1	0
0	0	1

Paren matriks

1,2	1,3	2,3
1	1	0
1	0	1
1	0	1
0	1	1

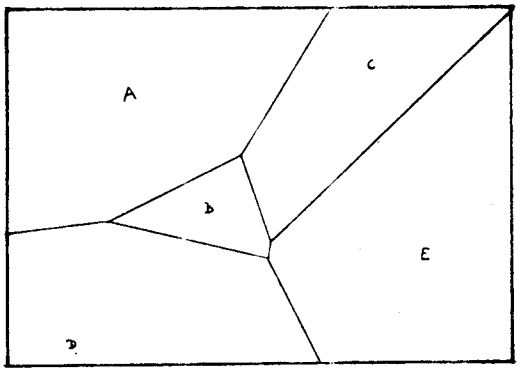
figuur 5.5 Meervoudige kontinue representatie van een variabele

Wij delen een k-kategorie variabele op in ½k(k - 1) binaire variabelen schalen. Elk paar van kategorieën is dus een variabele met twee kategorieën geworden. Individuen die niet in een van de twee kategorieën van een dergelijk paar vallen worden door dat paar in hun modelskores niet beperkt. In figuur 5.5 zien wij de geometrie van meervoudig kontinue schalen aan de hand van een voorbeeld met een variabele van drie kategorieën. Rechts in fig. 5.5 staan de gewone indikator matriks en ter verduidelijking ook een indikator matriks voor de paren. De scheidende hypervlakken zijn aangegeven met gestippelde lijnen en de richting per paar is aangegeven met een pijl. Het gaat natuurlijk om de hypervlakken. Gekombineerd met elkaar definieren zij niet-overlappende gebieden, waarin per gebied die punten liggen die horen bij individuen die in dezelfde kategorie vallen. Dit wordt wat duidelijker als we kijken naar figuur 5.6, waar we een variabele met meer kategorieën schalen en waar tevens meer individuen in elke kategorie vallen.



figuur 5.6 Meervoudige kontinue representatie van een variabele met vijf kategorieën.

Wij gebruiken hiervoor een konfiguratie die in hoofdstuk 10 in een andere kontekst eveneens beschreven wordt. Stel wij hebben een indeling in vijf beroepsgroepen, gekodeerd A, B, C, D en E, met in elke beroepsgroep evenveel individuen. We hebben $\binom{5}{2} = 10$ hypervlakken nodig om de 40 punten naar beroepsgroep te scheiden. Of, met andere woorden, er zijn 10 paren van beroepen waarvan de beoefenaars door een hypervlak van elkaar gescheiden moeten worden. Deze 10 hypervlakken definieren 5 elkaar niet-overlappende gebieden.



figuur 5.7 Niet-overlappende gebieden per kategorie met meervoudig kontinue schalen.

In figuur 5.7 zien we de gebieden die door de hypervlakken in figuur 5.6 begrensd worden, in kaart gebracht. De verliesfunktie die bij deze benadering hoort is de ordinaal kontinue enkelvoudige verliesfunktie waar elk paar kategorieën een binaire variabele is en waarbij de individuen die in geen van beide kategorieën vallen als "missing data" opgevat worden (en dus wat dat paar betreft niet homogeen hoeven te zijn).

5.3 Het algoritme van HOMALS en PRINCALS

Ook in termen van algoritmes lijken HOMALS en PRINCALS veel op elkaar. Eerder in dit hoofdstuk is al aangetoond dat de normalisatie van X in beide gevallen tot dezelfde berekeningen leidt. In paragraaf 5.2.3. zijn de updates voor z en a met behulp van de meervoudige kategorie-kwantifikaties geformuleerd. Het is dus niet zo verwonderlijk noch toevallig dat de algoritmes op analoge wijze in elkaar gestoken zijn. Het grote voordeel van deze manier van handelen is dat de relatieve snelheid en efficient geheugenbeslag van het HOMALS algoritme gebruikt wordt. De PCA representatie vindt plaats als een extra bewerking van de meervoudige kategorie-kwantifikaties. Het HOMALS algoritme heeft deze gunstige eigenschappen omdat gebruik gemaakt wordt van de "sparseness" van indikator matriksen, waardoor vermenigvuldigen optellen wordt, terwijl deze matriksen in hun gereduceerde vorm opgeslagen worden.

Een tweede reden is de manier, waarop de normalisering van X wordt opgelost: namelijk als een iteratieve benadering van singuliere vectoren (=direkte iteratie). Dit in tegenstelling tot soortgelijke modellen, waar per iteratie een kompleet eigenwaarden probleem wordt opgelost, als het al iteratief wordt aangepakt. Qua rekenmethodes werd vroeger meestal, ondanks Richardson, een komplete EVD van de matriks van geschaalde bivariate marginalen uitgevoerd. Als de som van het aantal kategorieën wat groter werd was het probleem al snel onoplosbaar, cf. Lingoes,1968.
Thurstone opperde in 1947 reeds het idee dat NMPCA een bruikbare techniek zou kunnen zijn en in 1959 kwam Guttman met de noodzakelijke vergelijkingen om het probleem op te lossen. In dit licht doet het wat vreemd aan dat andere auteurs rond 1970 elkaar de eer betwisten de eerste te zijn die NMPCA mogelijk maakte.
De grootste bekendheid op dit gebied hebben de programmaas van Kruskal en Shepard(1974), Young,1972 (POLYCON), en Takane, Young en De Leeuw,1978 (PRINCIPALS). Minder bekende broeders zijn PRINQUAL (Tenenhaus, 1977) en LINEA (Roskam, 1968). PRINCALS lijkt het meest op PRINCIPALS en niet alleen in naam. De voornaamste verschillen zijn dat PRINCALS wel meervoudig en niet kontinue kan schalen en dat PRINCIPALS niet meervoudig en wel kontinue kan schalen. Een ander opvallend verschil is dat PRINCALS dankzij de specifieke structuur van het algoritme sneller en efficienter rekent dan PRINCIPALS.

In de beschrijving van de algoritmes is het aantrekkelijk HOMALS als een beperkt geval van PRINCALS te beschouwen. Daarom behandelen we eerst het PRINCALS algoritme.
Alternerend wordt er geminimaliseerd over X, Z en A onder de kondities $X'X = I$, $u'X = 0$, $z'Dz = 1$ en $z \in C$. De laatste twee restrikties gelden per variabele. Het is een iteratief algoritme; de waarden van de voorafgaande iteratie hebben de hoge indeks ^{0} en de updates in de volgende iteratie hebben de indeks $^{\times}$. We beginnen met aanvangs-schattingen voor X en A.
De volgende stappen worden alternerend berekend totdat het absolute verschil in stress tussen twee opvolgende iteraties verwaarloosbaar klein wordt.

$$(1) \qquad U_j^0 = D_j^{-1} G_j' X^0$$

$$(2) \qquad z_j^0 = P_{C_j}(U_j^0 a_j^0)$$

$$(3) \qquad z_j^\times = z_j^0 (z_j^{0\prime} D_j z_j^0)^{-\frac{1}{2}}$$

$$(4) \qquad a_j^\times = z_j^\times U_j^0$$

$$(5) \qquad X^\times = \overline{X}(\overline{X}'\overline{X})^{-\frac{1}{2}}$$

met $$\overline{X} = \frac{1}{m} \sum_{j=1}^{m} G_j z_j^\times a_j^\times,$$

en in stap (2) is z_j^0 de <u>projektie</u> van $U_j^0 a_j^0$ op de kegel C_j.

Stap (1) tot en met (4) worden succesievelijk voor elke variabele uitgevoerd. De stappen (2), (3) en (4) worden alternerend berekend totdat het absolute verschil in stress, veroorzaakt door deze twee parameters, binnen de hoofd iteratie stabiel is. In de regel is een zo'n interne iteratie voldoende. De schatting van X en A is afkomstig van een SVD van de ruwe datamatriks. Deze SVD is ook met het bovenstaand algoritme uitgevoerd en de initiele configuratie hiervan is een geortogonaliseerde verzameling van willekeurig gekozen getallen tussen 0 en 1. Het algoritme is versneld door stap (5) niet iedere iteratie uit te voeren maar om de drie iteraties. Het is nog een onderwerp voor verder onderzoek dit aantal niet-ortogonaliserende iteraties te optimaliseren, maar blijkens onze ervaring, is drie in ieder geval aan de veilige kant.
Voor die variabelen die wij alleen meervoudig willen schalen slaan we stap (2), (3) en (4) over. Als we deze drie stappen totaal verwijderen uit het algoritme dan houden we het HOMALS algoritme over dat er dan als volgt uit ziet:

We beginnen met een schatting van X, die bestaat uit een geortogonaliseerde configuratie van willekeurig gekozen getallen tussen 0 en 1. De volgende twee stappen worden alternerend uitgevoerd totdat het absolute verschil in stress tussen twee opvolgende iteraties kleiner dan een van te voren afgesproken klein getal wordt.

$$(1) \qquad Y_j^0 = D_j^{-1} G_j X^0$$

$$(2) \qquad X^{\times} = \overline{X}(\overline{X}'\overline{X})^{-\frac{1}{2}}$$

met $$\overline{X} = \Sigma_{j=1}^{m} G_j Y_j^0$$

Er wordt alternerend geminimaliseerd over X en Y onder de restrikties X'X = I en u'X = 0. Ook hier worden steeds drie niet-ortogonaliserende iteraties uitgevoerd tegenover een iteratie met stap (2).

Zowel in PRINCALS als in HOMALS wordt na de laatste iteratie, d.w.z. als de stress zich gestabiliseerd heeft, de matriks X geroteerd naar de stand van de principale komponenten en de hiermee korresponderende eigenwaarden. Hierna worden in PRINCALS de stappen (1) tot en met (4) en in HOMALS stap (1) nog een keer berekend om zodoende de andere parameters ook te roteren.

HOMALS en PRINCALS zijn programmaas geschreven in ANSI Fortran IV. Zij zijn speciaal gemaakt om, behoudens snel en efficient te rekenen, gebruikt te kunnen worden op verschillende merken komputers. De struktuur is zodanig dat een programmeur of gemotiveerde leek, de flow gemakkelijk kan volgen, zowel door het uitvoerige kommentaar in het programma zelf als door de eenvoudige opbouw van de programmaas. Beide programmaas zijn voorzien van de mogelijkheid tot dynamische geheugen allokatie. Voor IBM komputers is deze mogelijkheid standaard ingebouwd. Voor beide programmaas zijn uitvoerige gebruikershandleidingen beschikbaar (Van Rijckevorsel en De Leeuw, 1978,1979).

5.4. Enkele voorbeelden

5.4.1 Abortus

De al eerder besproken abortus gegevens bieden een fraaie gelegenheid PRINCALS met al zijn toeters en bellen toe te passen. We werken met een selektie van variabelen gekozen op grond van hun eendimensionale HOMALS diskriminatiematen, zoals beschreven in hoofdstuk 2. Het is naar aanleiding van het voorbeeld in hoofdstuk 3 interessant na te gaan in welke mate de tweede PRINCALS dimensie verschilt van de tweede HOMALS dimensie. Maar dit is niet de reden waarom we dit voorbeeld hier ten tonele voeren. In deze gegevens zijn meerdere variabelen die men graag afwijkend van de overigen zou willen schalen. Typische achtergrondsgegevens zoals religie en politieke keuze vragen a.h.w. om meervoudige schaling terwijl

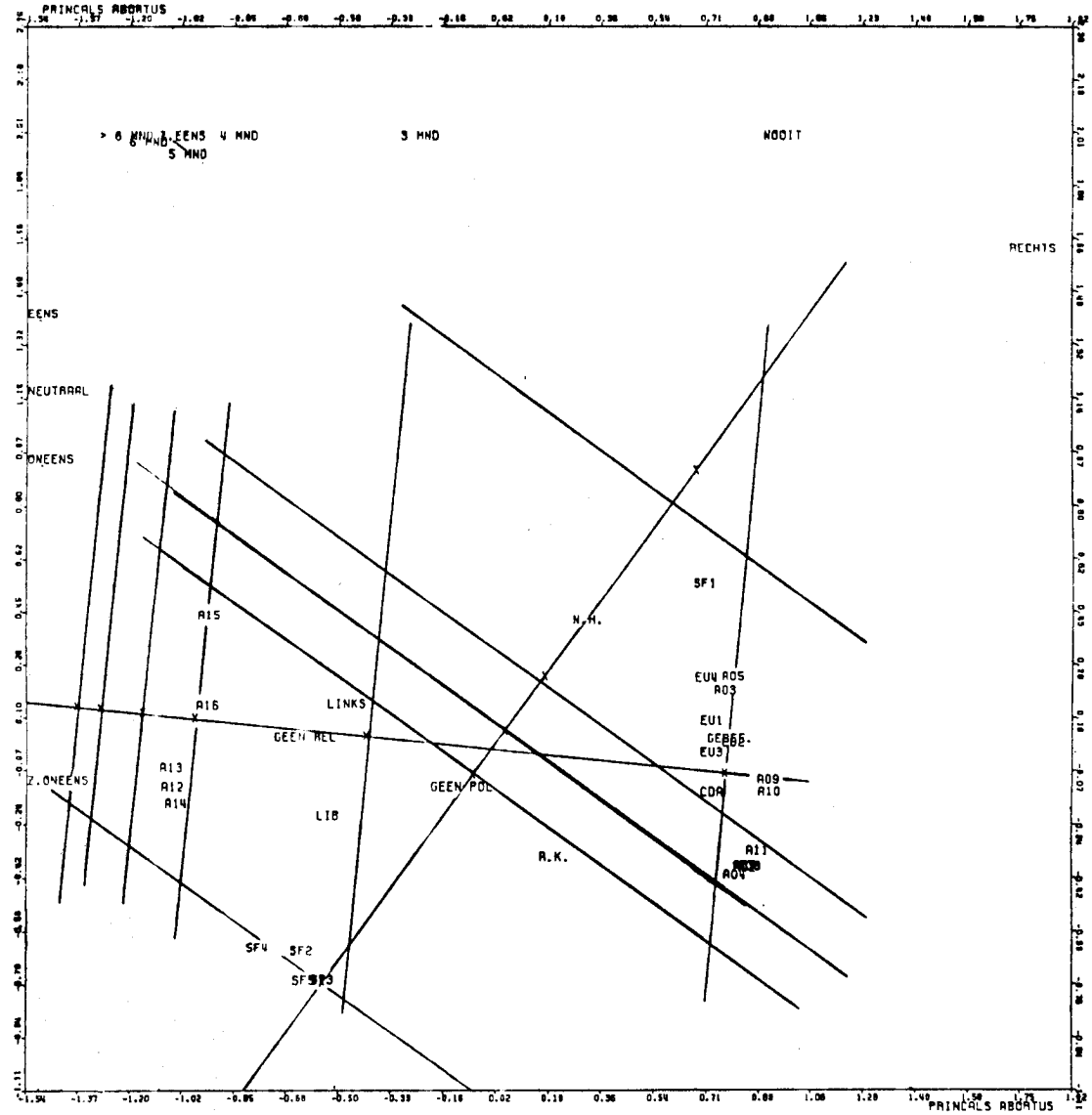


FIG:5.8 PRINCALS TWEE DIMENSIONALE OPLOSSING VAN ABORTUS DATA MET ORDINALE,NOMINALE EN MEERVOUDIG NOMINALE VARIABELEN. RICHTINGSCOSINUSSEN EN MEERVOUDIGE KATEGORIESCORES.

men likert-type items zoals A11 t/m A16 en SF01 t/m SF05 enkelvoudig ordinaal zou willen schalen. Het meetnivo is natuurlijk niet relevant voor binaire data zoals A02 t/m A08 en EU2 t/m EU4 omdat deze in alle meetnivoos op dezelfde manier geschaald worden. Dit betekent natuurlijk niet dat deze variabelen irrelevant zijn. De variabelen A09 en A10 nemen een ietwat bijzondere positie in. De betekenis van de kategorieën van deze variabelen verleidt ons tot ordinale spekulaties maar de kategorie nummers doen dit jammergenoeg niet. In hoofdstuk 2 is dit opgelost door deze variabelen te herkoderen. Het is leuk dat PRINCALS met enkelvoudige nominale schaling zonder herkoderen tot de "goede" volgorde komt. In figuur 5.8 zijn de gewichten (korrelaties) van de variabelen geplot met ter illustratie de bij variabele A09 en SF5 behorende hypervlakken. Let op de positie van de kategorie " nooit " van variabele A09. De gewichten van de variabelen op de eerste as zijn

aardig hoog. Dit is te verwachten want de variabelen zijn er met HOMALS speciaal op gekozen. Meer afwijkend en daarom opvallend is positie van de SF (**sexuele tolerantie**) items. Bij het konstrueren van de Likert-schaal in hoofdstuk 2 was al enigzins te zien dat weliswaar voor een subset van deze variabelen er twee groepen te onderscheiden waren, abortus items en SF items. Wij gaan hier niet verder in op PRINCALS analyses van deze subset ofschoon er wel interessante resultaten zijn. In onze plot liggen alle SF items ongeveer op een rechte lijn onder een hoek van 45^0 met de X-as. Oordelen over SF domineren de tweede dimensie samen met de meervoudige kategorieën politiek rechts en nederlands hervormd. Dit in tegenstelling tot de twee-dimensionale HOMALS oplossing in hoofdstuk 3 waar het hoefijzer deze items geheel opslurpt. De meervoudige schaling van politieke keuze en religie is duidelijk te herkennen. We zullen nu de inhoudelijke betekenis van het plaatje bespreken.

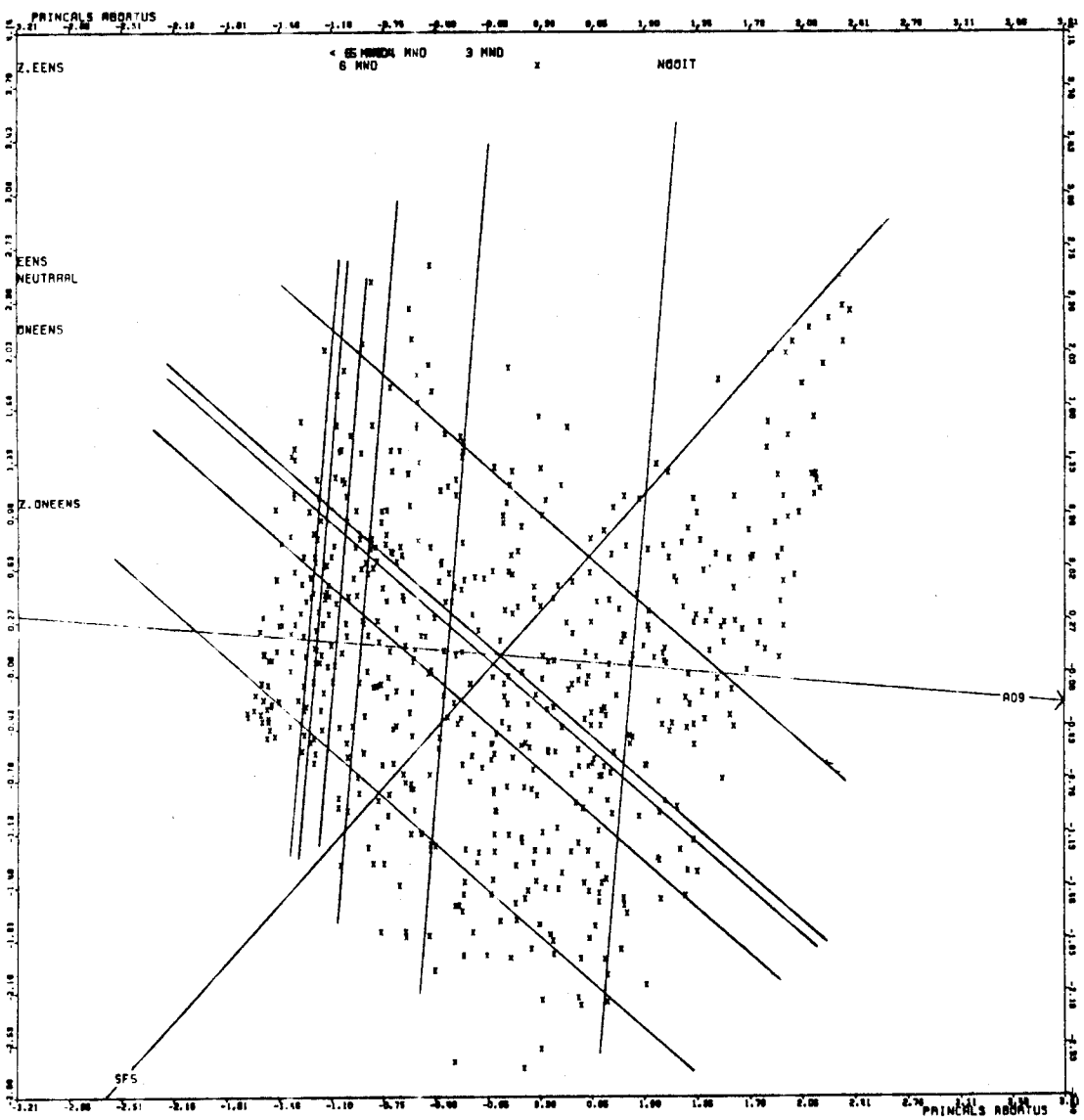


FIG.5.9 PRINCALS TWEE DIMENSIONALE OPLOSSING VAN ABORTUSDATA MET ORDINALE,NOMINALE EN MEERVOUDIG NOMINALE VARIABELEN. INDIVIDU SKORES.

Het lijkt wel of er twee kampen zijn. Individuen in de ene sektor zijn tegen abortus, tegen euthanasie, seksueel intolerant, rechts of CDA en gelovig, een vertrouwde kombinatie dus en het andere kamp is ongelovig, links of liberaal, denkt genuanceerd over abortus en is redelijk tolerant in seksuele zaken. Extreme intolerantie inzake seks is kenmerkend voor politiek rechts en ook maar in mindere mate voor nederlands hervormd. Individuen zonder politieke voorkeur liggen in gelijke mate verspreid over beide sektoren.

In de plot van de individu-skores, zie figuur 5.9, kunnen wij zien dat de "voorstanders" van abortus nogal dicht bij elkaar liggen, terwijl de tegenstanders naar rechts uit waaieren (letterlijk en figuurlijk). Met een beetje goede wil en wat fantasie heeft de puntenwolk in figuur 5.9 de vorm van een dikke V. De rechterpoot wordt gevormd door confessionele, politiek rechtse, tegenstanders van abortus. De linkerpoot door linkse niet-gelovige voorstanders van abortus en euthanasie. De seksueel tolerante leden van beide groeperingen komen elkaar tegemoet onder in de punt van de V, ofschoon links over de gehele linie toleranter in seksuele zaken is.

5.4.2. Voorkeur in de tweede kamer

In HOMALS is het zo dat als er evenveel individuen als kategorieën zijn, de individu-skores en de kategorie-kwantifikaties op normalisatie na volkomen vrij zijn. Een genormaliseerde random punten wolk voldoet in zo'n geval aan de restriktie dat de kategorie-kwantifikatie en de skore van een individu dat in die kategorie valt samenvallen, perfekte fit dus. Dit is eigenlijk alleen maar interessant om te weten als je beseft dat dit in PRINCALS helemaal niet zo is. Sterker nog, een populaire toepassing van PRINCALS bestaat uit het analyseren van matriksen met juist deze eigenschappen. Om te voldoen aan de PRINCALS restrikties is een willekeurige puntenwolk niet goed genoeg. Een individu-skore moet op een hypervlak liggen van de kategorie waarin dat individu valt, hetgeen als hij of zij de enige is, te vergelijken is met de HOMALS restriktie, maar ook moeten kategorie-kwantifikaties geordend op een vektor liggen. Deze laatste restriktie verhinderd nu dat de random konfiguratie voldoet. De individu-skores moeten immers geordend op die vektor projekteren.

Dit betoog wordt wat duidelijker als wij uitleggen welk soort matriksen wij hier bedoelen. We willen met PRINCALS voorkeur-rangordes analyseren. De rangnummers die worden toegekend aan de objekten van voorkeur zijn de kategorieën. Van elk individu is een vektor met rangnummers beschikbaar. De achtergrond van deze handelswijze en de relaties met HOMALS en andere technieken zoals MDPREF (Carrol,1972) worden in hoofdstuk 10 behandeld.
Stel we hebben een matriks van n rangordes van m objekten. We willen de individuen afbeelden als vektoren in de ruimte en de m kategorieën als hypervlakken loodrecht op deze vektoren. De m objekten moeten ook in deze ruimte liggen. Let op, er zijn evenveel kategorieën als variabelen. De objekten moeten zodanig in deze ruimte liggen dat hun projekties op een individu-vektor de voorkeursrangorde van dat individu weergeeft. Een dergelijke weergave kunnen wij alleen bereiken door de n x m rangorde matriks te kantelen en aldus met PRINCALS te analyseren. De individuen zijn de kolommen van de matriks en de objekten de rijen. Elke kolom heeft nu evenveel kategorieën als er rijen zijn en elke kategorie komt maar één keer voor. Als wij deze matriks in deze vorm met HOMALS zouden analyseren komen we niet verder dan de random start konfiguratie en zijn we in een iteratie klaar. In PRINCALS echter niet.
Ter illustratie gebruiken we de tweede kamer gegevens uit 1968 (vgl. Daalder en Rusk 1972, De Leeuw 1973, Daalder en Van de Geer 1977). Het zijn 141 voorkeursrangordes van 141 tweede kamerleden voor 12 politieke partijen in 1968. In de plaatjes staat met een letterkode vermeld van welke partij een individu lid is (zie tabel 5.1).

PARTIJ	AANTAL	KODE
CPN	-	-
PSP	4	P
PvdA	37	L
D'66	7	6
PPR	-	-
KVP	42	K
ARP	15	A
CHU	12	U
VVD	17	V
BP	4	B
SGP	2	S
GPV	1	G

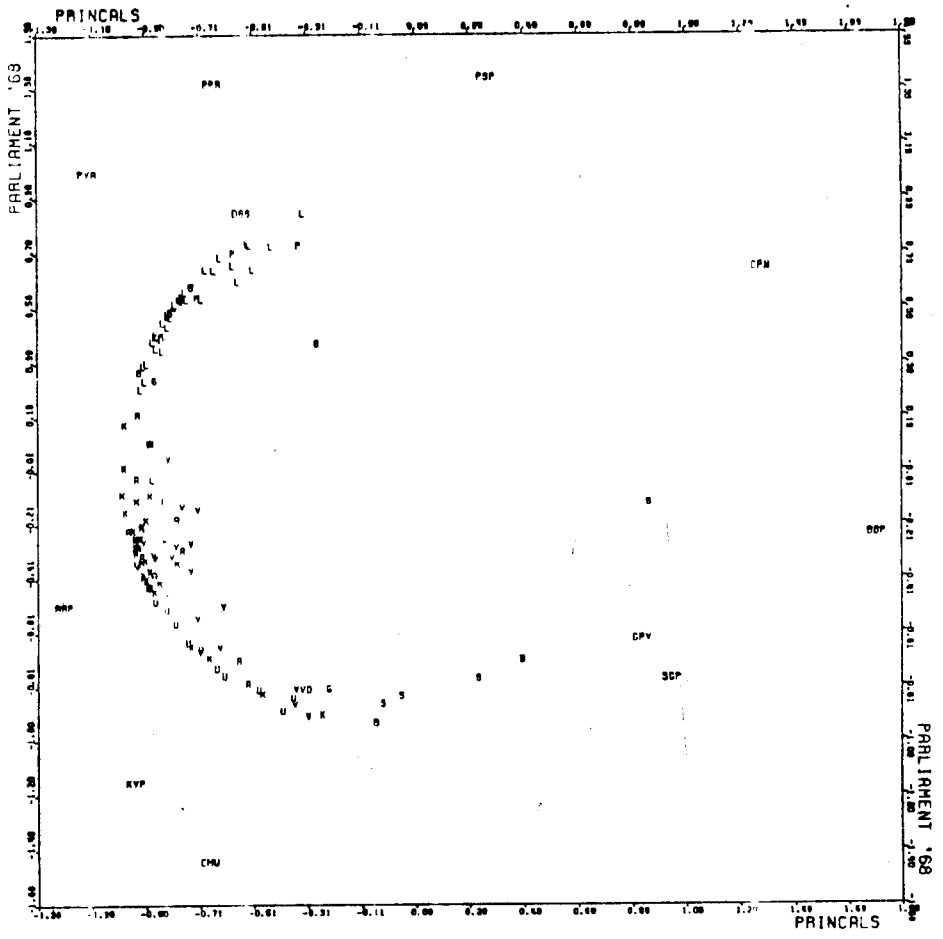
- = geen antwoord

Tabel 5.1

We geven twee oplossingen. De eerste is de initiele metrische konfiguratie in figuur 5.10. Deze hebben wij als aanvangschatting gebruikt voor de ordinale analyse, waarvan de uiteindelijke oplossing in figuur 5.11 staat. De twee plaatjes lijken veel op elkaar; ze staan alleen t.o.v. elkaar gespiegeld. Ook de bijbehorende eigenwaarden verschillen weinig van elkaar (zie tabel 5.2).

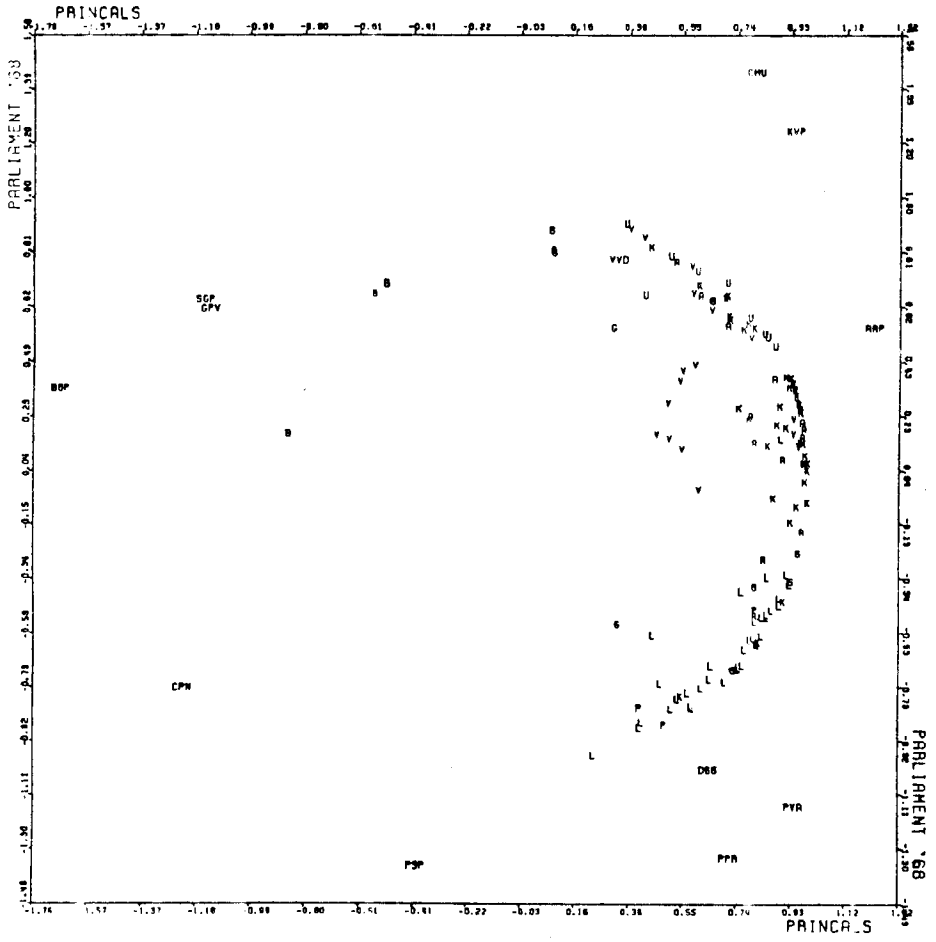
dim.	num.	ord
1	.56	.64
2	.25	.28

Tabel 5.2



figuur 5.10 de voorkeurs rangordes van 141 kamerleden met 12 partijpunten. De initiele metrische konfiguratie.

In het algemeen is het duidelijk dat kamerleden hun eigen partij bovenaan in hun voorkeursrangorde zetten. Daarom vinden we in de buurt van de PvdA, D'66 en PPR de meeste L, P en 6 punten,en in de buurt van de CHU, KVP en VVD de meeste U, V, en K punten met hier en daar een verdwaalde A (=AR) etc..
De positie van de CPN is wat instabiel zoals we in hoofdstuk 9 met behulp van de zogenaamde "bootstraps" zullen zien. Dit komt doordat deze partij zowel door het gematigd linkse blok als door het rechts-confessionele blok als de minst geprefereerde partij wordt gezien, terwijl de genoemde blokken in onze oplossing nog aardig van elkaar verwijderd liggen.



figuur 5.11 de voorkeurs-rangordes van 141 kamerleden met 12 partijpunten. De uiteindelijke ordinale oplossing.

5.4.3. Van Jaar tot Jaar

De gegevens gebruikt in dit voorbeeld zijn de zg. "Van Jaar tot Jaar " data, waarvan een uitvoerige beschrijving is opgenomen in de appendix.

Stel u hebt de overtuiging dat de genoemde 25 variabelen veel te maken hebben met sukses op school. U bent dan in het gezelschap van pakweg 6 miljoen andere nederlanders die dat ook denken.

Variabele kode	PRINCALS numeriek	PRINCALS ordinaal	HOMALS meervoudig
BVA	.56	.63	.63
OPV	.63	.65	.65
OPM	.49	.51	.51
AKG	.23	.24	.24
URB	.17	.18	.18
ASO	.19	.22	.22
ASL	.14	.25	.27
BIL	.20	.25	.24
BIM	.40	.44	.43
INT	.39	.35	.35
OOA	.22	.22	.22
DWO	.09	.09	.09
BMB	.44	.42	.41
INS	.07	.05	.05
KLS	.14	.15	.15
DLO	.42	.39	.38
LL6	.05	.10	.12
ADV	.77	.81	.81
KGS	.65	.65	.65
PRE	.80	.81	.80
TON	.86	.89	.89
AOS	.42	.45	.46
LLS	.32	.35	.37
EXT	*.30*	*.42*	*.59*
EIN	.84	.86	.86

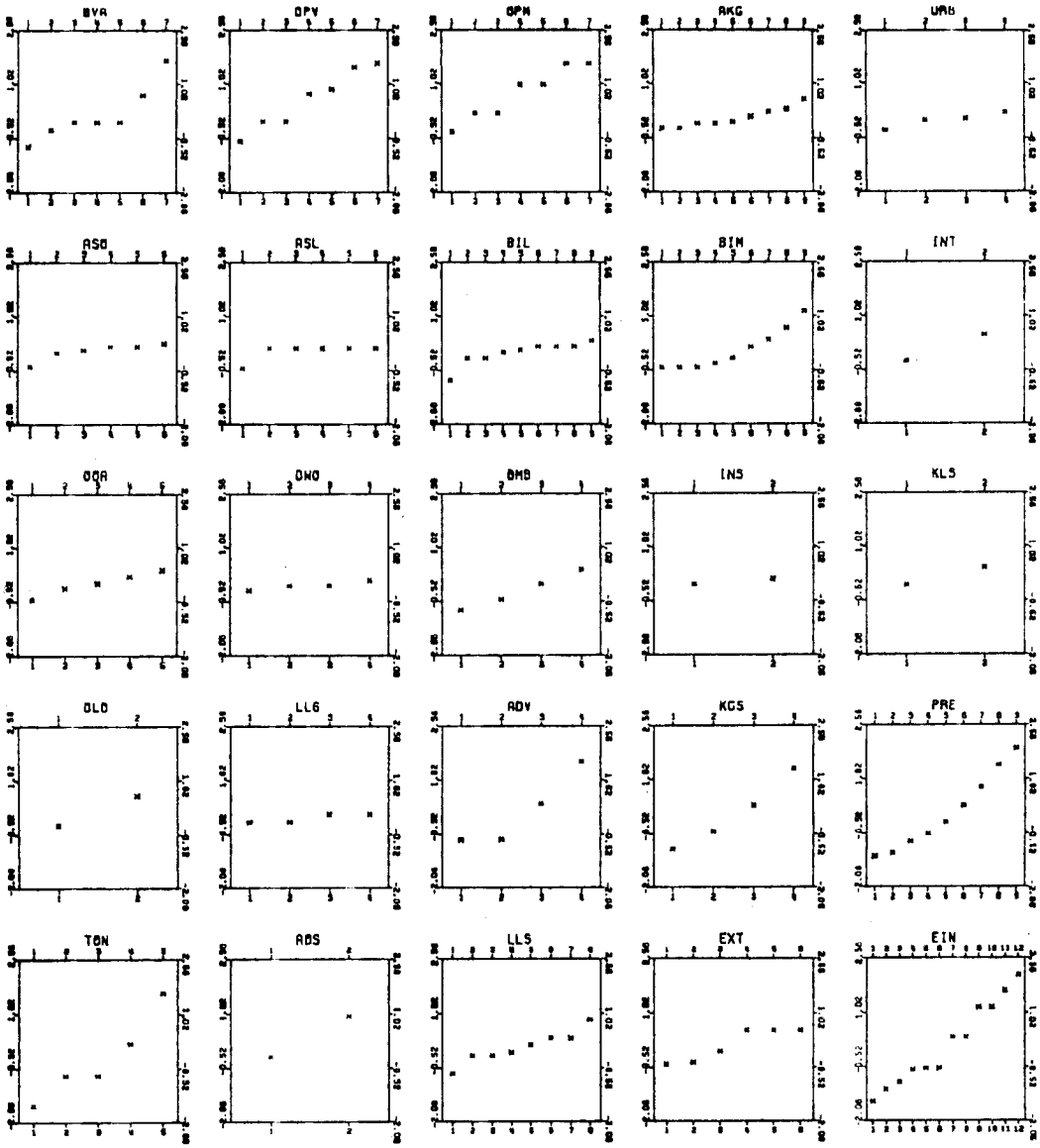
Tabel 5.3 Korrelaties met de eerste principale komponent.

En afhankelijk van uw mening over aanleg en omgeving vindt niet alle 25 variabelen even belangrijk. Ook hierin staat u niet alleen. Het is daarom aantrekkelijk en noodzakelijk na te gaan bij een steekproef van schoolverlaters welke van deze variabelen een rol spelen bij sukses op school.
Zoals al eerder in hoofdstuk 2 is aangetoond is de eerste HOMALS dimensie een benadering van "sukses op school". We laten zien dat de eerste PRINCALS dimensie, zowel ordinaal als numeriek slechtere benaderingen zijn en bij twee variabelen laten wij zien waarom.
De eerste principale komponent van deze 25 variabelen is een benadering van sukses op school en de hoogte van de korrelaties van de variabelen met deze komponent geeft de belangrijkheid t.a.v. dit kriterium weer. Alle variabelen willen we in PRINCALS zo schalen dat ze maximaal korreleren met de eerste komponent (zie tabel 5.3). Hoe goed de benadering in zijn geheel is kunnen we aflezen aan de eigenwaarde die hoort bij de eerste principale komponent. Deze eigenwaarde is in dit geval gelijk aan het gemiddelde van de wortels van de korrelaties in tabel 5.3. De korrelaties vervullen dezelfde rol als de diskriminatie maten uit HOMALS in hoofdstuk 2.

PRINCALS	.2143	num.
PRINCALS	.2340	ord.
HOMALS	.2420	nom.

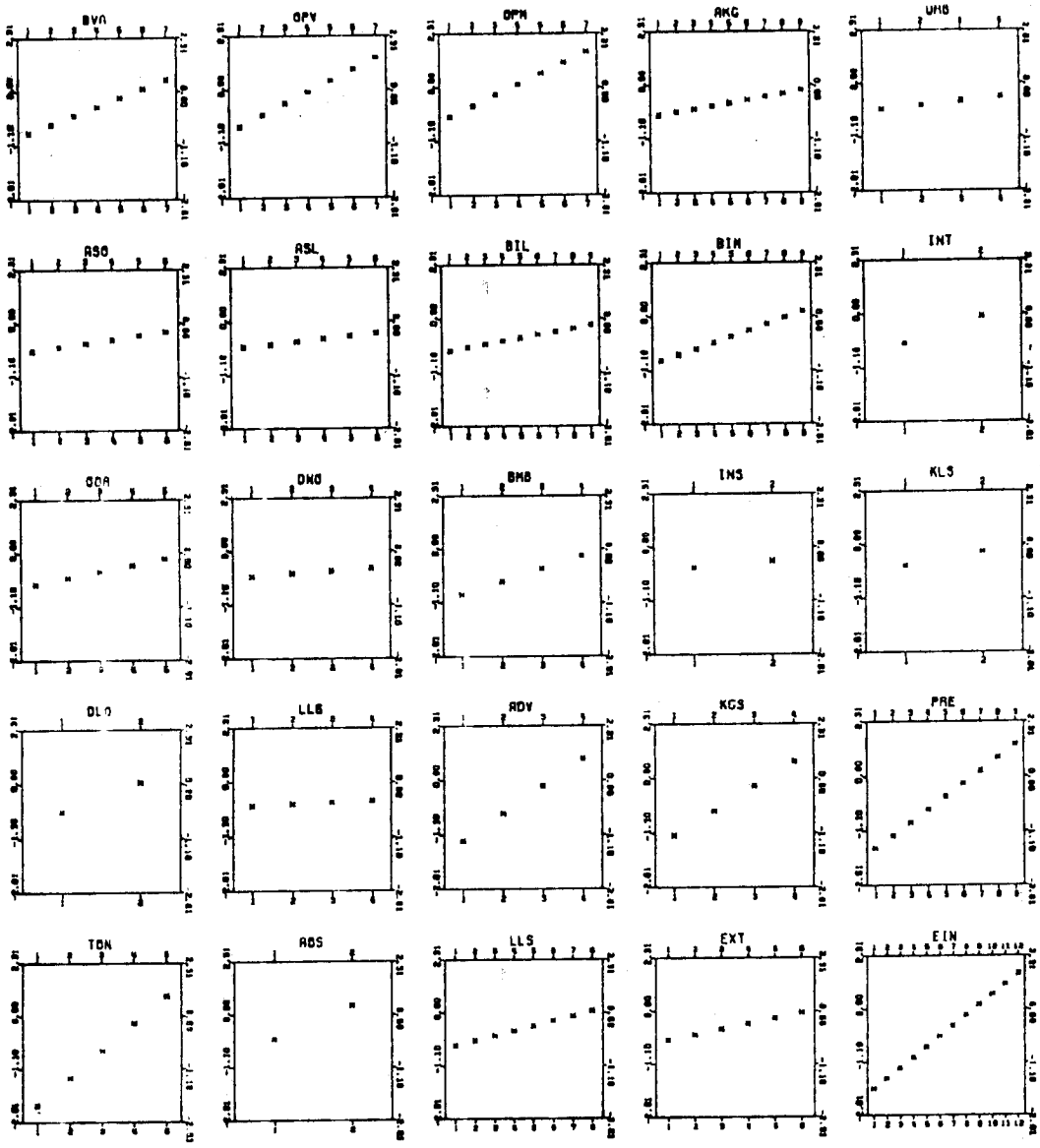
Tabel 5.4 eerste eigenwaarde.

We mogen de eigenwaarden van HOMALS en PRINCALS met elkaar vergelijken omdat in het een-dimensionale geval beide technieken hetzelfde kriterium maximaliseren. (zie tabel 5.4)



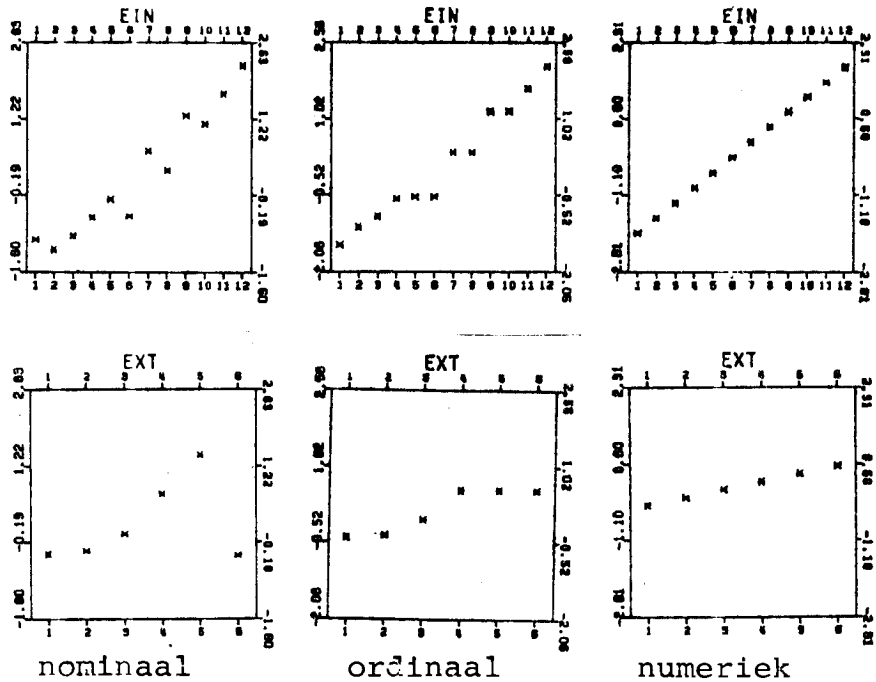
figuur 5.12 een-dimensionale ordinale transformaties uitgezet tegen de kategorie nummers.

Alleen de transformaties die HOMALS en PRINCALS gebruiken zijn verschillend. Deze transformaties uitgezet tegen de kategorie nummers staan in figuur 5.12 in het ordinale geval en in figuur 5.13 in het numerieke geval. (vgl. de HOMALS transformaties in figuur 2.11)
In het een-dimensionale geval zijn enkelvoudig en meervoudig nominaal schalen aan elkaar gelijk. Het meest spektakulaire van deze toepassing zijn de verschillen in transformaties tussen de diverse analyses. De hoek die de transformaties maken met de X-as is pro-



figuur 5.13 een-dimensionale numerieke transformaties uitgezet tegen de kategorie nummers

portioneel met de korrelatie van de desbetreffende variabele met de principale komponent. Voor de niet-numerieke transformaties kunnen wij beter over het schaalbereik van de transformaties in plaats van de hoek praten. Hoe kleiner de hoek, of hoe kleiner het schaalbereik des te lager is de korrelatie.



figuur 5.14 drie verschillende transformaties voor twee variabelen

Ter illustratie van de "kwaliteit" van de transformaties gaan we hier wat dieper in op de manier waarop de variabelen EIN en EXT geschaald zijn. EXT is het aantal extra-curriculaire aktiviteiten in het gevolgde sekundaire onderwijs en EIN is het behaalde eindnivo na de lagere school. In figuur 5.14 staan de diverse schalingen naast elkaar. Bij numeriek schalen geeft de korrelatie aan hoe goed de transformatie past op de variabele. Het is echter niet te zien welke kategorieën eventueel moeilijk te schalen zijn en op die manier de numerieke hypotese geweld aan doen. Bij ordinaal schalen is dit wel te zien. Bij EIN worden bijvoorbeeld de kategorieën 4,5,6 aan elkaar gelijk gemaakt. Dit gebeurt ook bij 7,8 en 9,10. De reden is dat niet afgemaakte hogere opleidingen, de kategorieën 6,8 en 10, in termen van kategorie nummers hoger aangeslagen worden dan lagere wel afgemaakte opleidingen (4,5,7 en 9). Meervoudige schaling vat de kategorie-nummers als labels op en is ongevoelig voor rangorde door onderzoekers aangebracht. Het feit dat de meervoudige transformaties zo lineair zijn is in hoofdstuk 2 uitvoerig aan bod gekomen.

Bij de variabele EXT is de fit bij enkelvoudig numeriek slecht bij enkelvoudig ordinaal redelijk en bij meervoudig nominaal goed te noemen (zie tabel 5.3). Dit komt door kategorie 6 (=nooit). Enkelvoudige schaling heeft geen verweer tegen de situatie waarin alle rangnummers behalve een toenemen met het aantal aktiviteiten. De toename in eigenwaarde tussen ordinaal en meervoudig schalen is grotendeels veroorzaakt door deze kategorie (zie tabel 5.3 en 5.4).

5.5 HOMALS, PRINCALS, en nog veel meer

Zoals we gezien hebben in hoofdstuk 2.1 kunnen we één-dimensionale HOMALS interpreteren als een techniek voor het vinden van transformaties $\phi_j(\underline{h}_j)$ zodat de grootste eigenwaarde van de matriks R met elementen $r_{j\ell} = R(\phi_j(\underline{h}_j), \phi_\ell(\underline{h}_\ell))$ zo groot mogelijk wordt. We geven nu een soortgelijke interpretatie van PRINCALS.

De verliesfunktie in PRINCALS is, in de notatie die stochastische veranderlijken gebruikt,

$$\sigma(\underline{x};\Phi;A) = \frac{1}{m}\sum_{j=1}^{m} V(\underline{x} - \phi_j(\underline{h}_j)a_j).$$

Hierbij zijn de a_j vektoren met p elementen, en is $\underline{x}$ een stochastische p-vektor. We veronderstellen als normalisatie dat $V(\underline{x}) = I$, en dat $V(\phi(\underline{h}_j)) = 1$. We kunnen het verlies ook schrijven als

$\sigma(\underline{x};\Phi;A) = I + A'A - A'B - B'A = I - B'B + (A - B)'(A - B)$.

Hierbij zijn zowel A als B matriksen van de afmetingen m × p, A heeft als elementen a_{js} en B heeft als elementen $b_{js} = R(\underline{x}_s, \phi_j(\underline{h}_j))$. Merk op dat, evenals in 2.1.5, het verlies gelijk is aan een matriks. Hier gelden weer dezelfde opmerkingen, onder de meeste 'redelijke' interpretaties van $\succsim$ geldt $\sigma(\underline{x};\Phi;A) \succsim I - B'B$, en dus $\sigma(\underline{x};\Phi;*) = I - B'B$. Minimaliseren over $\underline{x}$ geeft vervolgens, evenals in 2.1.5, het resultaat $\sigma(*;\Phi;*) = I - \Lambda_p$, waarbij Λ_p de diagonale matriks met de p grootste eigenwaarden van de matriks R is. Nogmaals, dit is een beetje anders dan PRINCALS. Bij PRINCALS gebruiken we, evenals bij HOMALS, het spoor van de verlies-matriks, en daardoor maksimaliseren we de som van de p grootste eigenwaarden van R. Onze verlies-matriks suggereert een meer algemene theorie. We hebben daarvoor eerst wat wiskunde nodig.

Stel $||.||$ is een unitair invariante matriks norm, gedefinieerd op de ruimte van alle m × m matriksen (vergelijk Von Neumann, 1937, Mirsky, 1958). Dus

a: $||A|| > 0$ als $A \neq 0$.

b: $||cA|| = |c|.||A||$.

c: $||A + B|| \leq ||A|| + ||B||$.

d: $||AK|| = ||KA|| = ||A||$ voor alle K zdd $KK' = K'K = I$.

We definieren bovendien symmetrische ijkfunkties over $\mathbb{R}^m$ als

a: $\omega(x) > 0$ als $x \neq 0$.

b: $\omega(cx) = |c|.\omega(x)$.

c: $\omega(x + y) \leq \omega(x) + \omega(y)$.

d: $\omega(Px) = \omega(x)$ voor iedere permutatiematriks P.

e: $\omega(Sx) = \omega(x)$ voor iedere tekenmatriks S (dwz S diagonaal met

diagonaalelementen gelijk aan +1 of -1).

Stel nu dat $A = K \Lambda L'$ de singuliere waarde dekompositie van A is, dan geldt dus $||A|| = ||\Lambda||$, en uit de eigenschappen van $||.||$ volgt dat $||\Lambda||$ een symmetrische ijkfunktie van de vektor $\lambda = diag(\Lambda)$ is. De unitair invariante norm van een matriks is dus een symmetrische ijkfunktie van de singuliere waarden van de matriks (ook andersom is deze stelling juist). We bespreken nu de stelling van Ky Fan over symmetrische ijkfunkties. Stel x en y zijn vektoren met $x_1 \geq \ldots \geq x_m \geq 0$ en $y_1 \geq \ldots \geq y_m \geq 0$. Dan definieren we $x \underset{\sim}{\geq} y$ als en alleen al $x_1 \geq y_1$, $x_1 + x_2 \geq y_1 + y_2$, $x_1 + x_2 + x_3 \geq y_1 + y_2 + y_3$, enzovoorts. De stelling van Fan (1951) zegt nu dat $x \underset{\sim}{\geq} y$ als en alleen als $\omega(x) \geq \omega(y)$ voor <u>iedere</u> symmetrische ijkfunktie ω. Als $x_1 \geq y_1$, $x_2 \geq y_2$, enzovoorts, dan natuurlijk $x \underset{\sim}{\geq} y$ en dus $\omega(x) \geq \omega(y)$. Als dus de geordende singuliere waarden van A groter zijn dan die van B, dan geldt dat $||A|| \geq ||B||$ voor iedere unitair invariante norm.

Een algemeen niet-lineair PCA probleem kan nu gedefinieerd worden als het maksimaliseren van $||R||$, waarbij $||.||$ een bepaalde unitair invariante norm is. HOMALS is het speciale geval $||R|| = \lambda_+(R)$, en PRINCALS in p dimensies is het speciale geval $||R|| = \Lambda_p(R)$, de som van de p grootste eigenwaarden. Er zijn echter andere interessante gevallen, zoals de gemiddelde kwadratische eigenwaarde van R (of, wat op hetzelfde neerkomt, de kwadratensom van alle elementen van R). Volgens de stelling van Ky Fan speelt het PRINCALS kriterium een speciale rol: een schaling is optimaal voor <u>alle</u> unitair invariante normen als en alleen als hij optimaal is voor PRINCALS in 1,2,...,m dimensies.

We kunnen wat verder komen met de algemene theorie als we aannemen dat er voor R een representatie bestaat van de vorm

$$R = \sum_{s=1}^{\infty} A_s C_{ss} A_s,$$

met A_s diagonaal, en $\sum_{s=1}^{\infty} A_s^2 = I$. In hoofdstuk 2 hebben we gezien dat een dergelijke representatie bestaat voor een groot aantal theoretische verdelingen, met de multinormaalverdeling als het meest prominente voorbeeld. We gebruiken nu de ongelijkheden

$$||R|| \leq \sum_{s=1}^{\infty} ||A_s C_{ss} A_s|| \leq \sum_{s=1}^{\infty} tr(A_s^2)||C_{ss}|| \leq \max_{s=1}^{\infty} ||C_{ss}||.$$

We zien dus dat het in dit geval voldoende is om ieder van de C_{ss} apart te bekijken, en degene met de grootste norm uit te kiezen.

Als dit C_{tt} is, dan maken we $A_t = I$, en $A_s = 0$ voor alle $s \neq t$. Dit impliceert dat $\phi_j(\underline{h}_j) = g_{jt}(\underline{h}_j)$, en als de g_j dus orthogonale polynomen zijn, dan zijn de $\phi_j(\underline{h}_j)$ allemaal polynomen van dezelfde graad.

Wat impliceert dit voor de normaalverdeling? We weten dat daarvoor $C_{ss} = R^{(s)}$, de matriks met korrelaties tot de macht s. Over het algemeen lijkt het zo te zijn dat voor alle unitair invariante normen geldt dat $||R^{(s+1)}|| \leq ||R^{(s)}||$. Zoals we weten is het hiervoor voldoende dat de relatie geldt voor de som van de p grootste eigenwaarden, voor alle p. Zolang we nog geen tegenvoorbeeld gevonden hebben nemen we maar aan dat dit inderdaad opgaat. Dan geldt dus $t = 1$ en de optimale transformatie is $\phi_j(\underline{h}_j) = g_{j1}(\underline{h}_j)$. De optimale transformatie is dus lineair voor alle unitair invariantie matriks normen waarvoor $||R|| > ||R^{(s)}||$ voor alle s.

Als we HOMALS meerdimensionaal toepassen op normaalverdeelde gegevens, dan berekent HOMALS als het ware alle eigenwaarden van alle $R^{(s)}$, en vervolgens selekteert hij de p grootste. Daarbij is het dus in ieder geval zo dat de grootste eigenwaarde altijd de grootste eigenwaarde van R is, maar de tweede grootste kan de grootste van $R^{(2)}$, maar ook de tweede grootste van R zijn. Voor de derde grootste zijn er nog meer mogelijkheden. In het algemeen zullen verschillende dimensies dikwijls transformaties van een andere graad geven, hoewel het zo blijft dat alle transformaties in dezelfde dimensie van dezelfde graad zijn. PRINCALS daarentegen selekteert zijn p-dimensionale oplossing altijd uit dezelfde $R^{(s)}$, dus alle p transformaties zijn lineair, of kwadratisch. We vermoeden dat de optimale transformatie lineair is in het geval van multinormaal verdeelde gegevens voor alle unitair invariante normen, maar bij de vermenging van twee binormaal verdelingen met korrelaties +ρ en -ρ hebben we gezien dat de optimale transformatie, zelfs in één dimensie best kwadratisch kan zijn.

De hier besproken theorie is redelijk algemeen, maar omvat toch niet alle mogelijkheden. Chang en Bargmann (1974) minimaliseren de determinant van de korrelatiematriks. De determinant is geen unitair invariante norm. Gebruik van de determinant is interessant vanwege de bekende identiteit $\det(R) = \det(R_j)(1 - r_j' R_j^{-1} r_j)$, waarbij R_j de deelmatriks van alle korrelaties behalve die met j is, en waarbij r_j de korrelatie van alle variabelen met j is. Als een skoring dus de determinant minimaliseert, dan geeft de skoring voor variabele de maskimale multipele korrelatie tussen j en de rest (voor gegeven

skoringen van de overige variabelen). De ongelijkheid van Oppenheim (Styan, 1973, stelling 3.6, pag 226) impliceert dat $\det(R^{(s+1)}) \geq \det(R^{(s)})$. In sommige gevallen kan het ook nuttig zijn zo te skoren dat de norm zo klein mogelijk wordt of de determinant zo groot mogelijk wordt. We willen dan R zoveel mogelijk laten lijken op de eenheidsmatriks, met name als we de skores later in multipele regressie willen gebruiken kan dit nuttig zijn. Als we minimaliseren in plaats van maksimaliseren is het echter niet steeds mogelijk alleen naar de individuele C_{ss} te kijken.

We geven nu een klein voorbeeld van een multinormaalverdeling met vier variabelen. De korrelatiematriks is van de vorm

1 A B C
A 1 C B
B C 1 A
C B A 1

De eigenvektoren van R zijn (kolomsgewijs)

+1 +1 +1 +1
+1 +1 -1 -1
+1 -1 +1 -1
+1 -1 -1 +1

en de bijbehorende eigenwaarden zijn

$\lambda_1 = 1 + A + B + C,$
$\lambda_2 = 1 + A - B - C,$
$\lambda_3 = 1 - A + B - C,$
$\lambda_4 = 1 - A - B + C.$

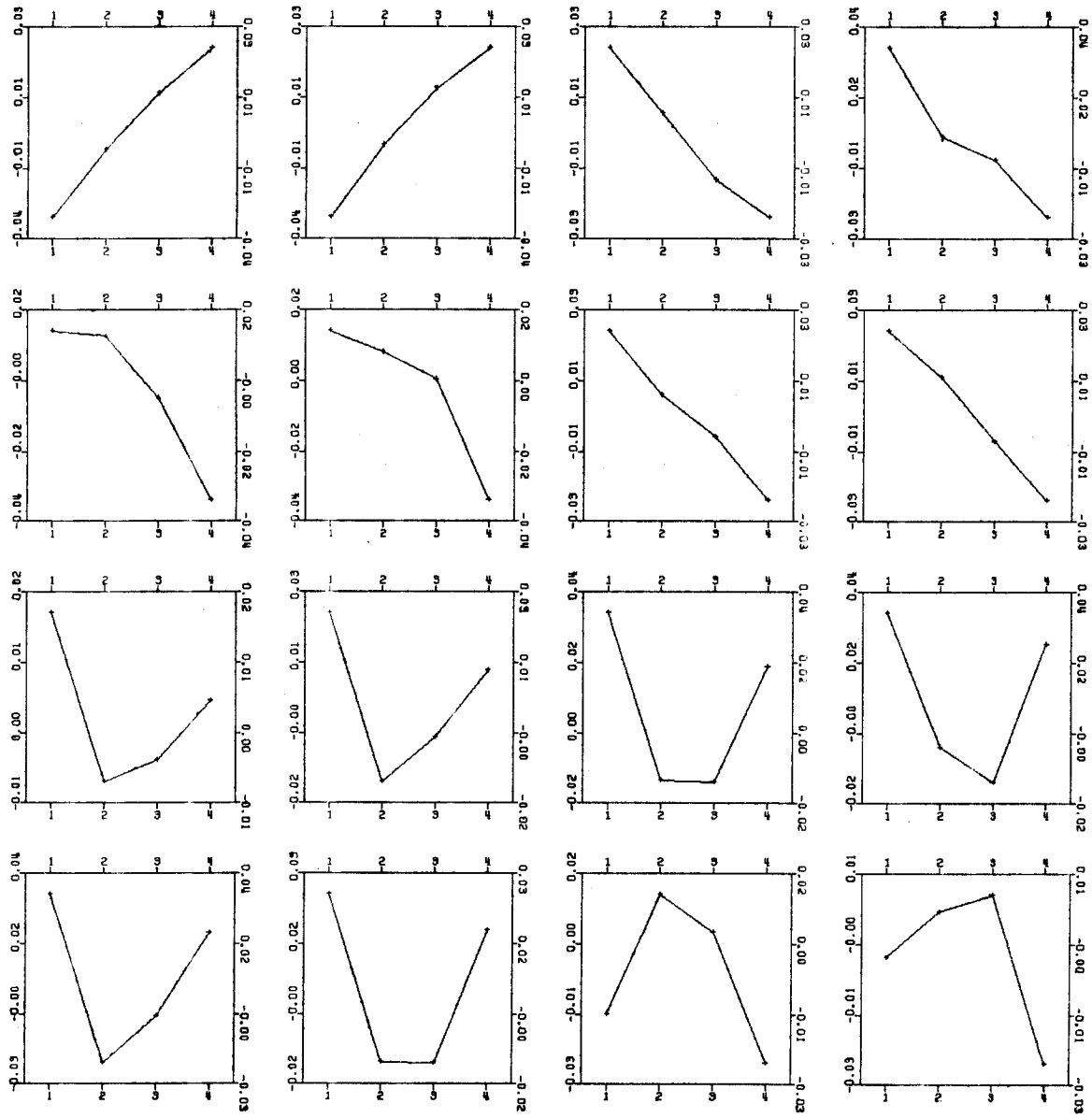
Het voorbeeld is opmerkelijk omdat $R^{(s)}$, de matriks met korrelaties tot de macht s, dezelfde vorm heeft, en daardoor ook dezelfde eigenvektoren. De volgorde van de eigenwaarden ligt echter niet vast, die ligt aan de relatieve grootte van A, B, en C. En als we de eigenwaarden van de $R^{(s)}$ gaan bekijken, dan wordt het volgorde probleem aardig ingewikkeld, zelfs voor dit kleine voorbeeld. We weten dat altijd $\lambda_{(1)}^{(1)}$, de grootste eigenwaarde van $R^{(1)}$, de grootste eigenwaarde van alle eigenwaarden van alle $R^{(s)}$ is. We willen nu A, B, en C kiezen zodat $\lambda_{(2)}^{(1)} - \lambda_{(1)}^{(2)}$ zo groot mogelijk is. Dat is dus de tweede eigenwaarde van R min de grootste eigenwaarde van $R^{(2)}$. Na wat gepuzzel blijkt de oplossing te zijn $A = \frac{1}{2}$ en $B = C = 0$, waarvoor $\lambda_{(1)}^{(s)} = \lambda_{(2)}^{(s)} = 1 + (\frac{1}{2})^s$ en $\lambda_{(3)}^{(s)} = \lambda_{(4)}^{(s)} = 1 - (\frac{1}{2})^s$. Dus $\lambda_{(2)}^{(1)} - \lambda_{(1)}^{(2)} = \frac{1}{4}$.

We gebruiken deze korrelatiematriks om een vierdimensionale normaalverdeling te definieren, waaruit we een steekproef van 1000 individuen trekken. We weten dat HOMALS in twee dimensies over het algemeen hoefijzers geeft, omdat $\lambda_{(1)}^{(2)} > \lambda_{(2)}^{(1)}$, en dat PRINCALS geen hoefijzers geeft omdat PRINCALS beide dimensies uit dezelfde $R^{(s)}$ moet kiezen. In dit voorbeeld verwachten we echter geen hoefijzer, maar voor de eerste twee dimensies verwachten we lineaire transformaties, voor de

volgende twee kwadratische transformaties, enzovoorts. Als ψ_1 en ψ_2 de Hermite Chebyshev polynomen van de eerste en tweede graad zijn, dan verwachten we voor de eerste vier dimensies de volgende eigenwaarden en transformaties.

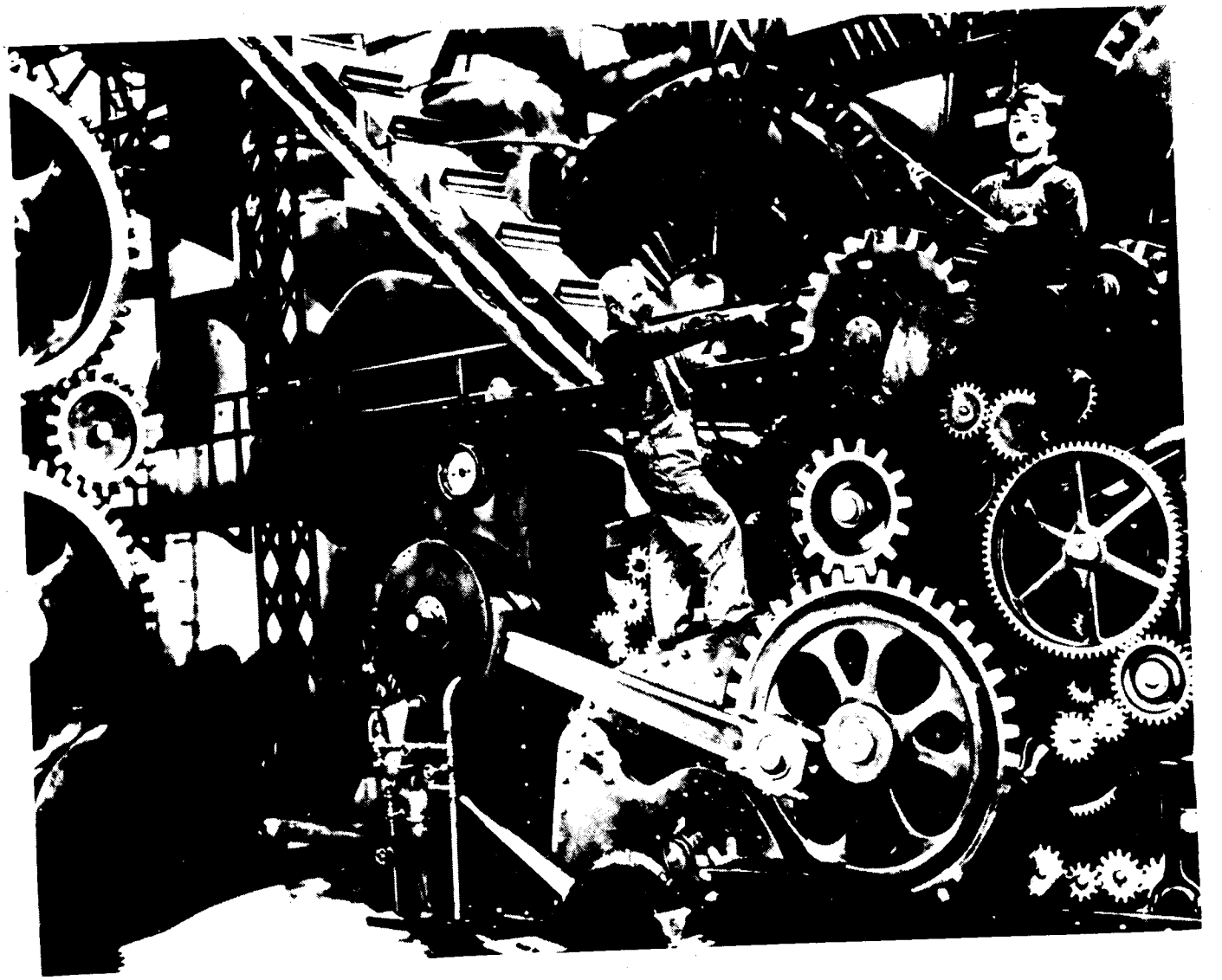
eigenwaarde 1.50	transformatie	$(+\psi_1 +\psi_1 +\psi_1 +\psi_1)$	gevonden 1.5311
1.50		$(+\psi_1 +\psi_1 -\psi_1 -\psi_1)$	1.4306
1.25		$(+\psi_2 +\psi_2 +\psi_2 +\psi_2)$	1.1807
1.25		$(+\psi_2 +\psi_2 -\psi_2 -\psi_2)$	1.1523

In de laatste kolom van dit informele tabelletje staan de eigenwaarden die we vinden in onze steekproef. We moeten daarbij bedenken dat we bovendien de kontinue normaalverdeelde variabelen in vier kategorieen gediskretiseerd hebben. Niettemin komen onze voorspellingen aardig uit, gezien ook figuur 5.



Figuur 5.15
transformaties. Rijen dimensies, kolommen variabelen.

6. Aansturing van programma's.



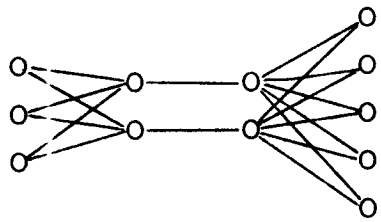
7. CANALS en HOMALS

7.1. Niet-lineaire kanonische korrelatie analyse

7.1.1. Inleiding

In de klassieke multivariate analyse kennen we het kanonische korrelatiemodel, waarbij twee sets van variabelen met elkaar in verband gebracht worden. Hotelling(1936) was de eerste die dit model introduceerde. Een voorbeeld dat we met het kanonische korrelatiemodel zouden kunnen aanpakken is een data set die bestaat uit konsumptiecijfers van een aantal goederen met daarnaast de prijs van de goederen. Is er verband tussen de konsumptiegegevens en de prijs van de goederen en welke variabelen dragen het meest bij tot dit verband? We kunnen bijvoorbeeld ook denken aan een verzameling achtergrondgegevens , zoals milieu, opleiding ouders, inkomen e.d. van een groep schoolkinderen. Daarnaast kunnen we ons een verzameling schoolresultaten van de kinderen voorstellen. De vraag rijst welk verband er bestaat tussen de achtergrondgegevens en de schoolresultaten van de kinderen. Een derde voorbeeld is dat men zich af kan vragen, of de meningen van mensen over een aantal onderwerpen zoals abortus en euthanasie in verband te brengen zijn met iemands stemgedrag.

Schematisch ziet het kanonische korrelatiemodel er als volgt uit (Van de Geer,1971):



eerste set tweede set

figuur 7.1. Schematische voorstelling van het kanonische korrelatiemodel.

Beide sets van variabelen omspannen een lineaire deelruimte. We zoeken in deze twee lineaire deelruimtes een gemeenschappelijke richting. Als deze richting inderdaad te vinden is , is er sprake van perfekte fit. Is er geen richting te vinden die de lineaire deelruimtes

voortgebracht door de eerste en de tweede set gemeenschappelijk hebben, dan zoeken we in beide lineaire deelruimtes een richting die zo goed mogelijk met de andere richting overeenstemt. In deze situatie is er sprake van imperfekte fit en is er een maat nodig voor de gelijkenis tussen de richtingen. We nemen hiervoor de korrelatie tussen de richtingen of kanonische assen en noemen dit de kanonische korrelatiekoëfficient.

Symboliseren we de eerste set van variabelen met Q_1 (nxm_1) en de tweede set van variabelen met Q_2 (nxm_2), dan zegt het kanonische korrelatiemodel dat we een deelruimte van de ruimte omspannen door de Q_1-vektoren zoeken, die zo veel mogelijk lijkt op een deelruimte van de ruimte omspannen door de Q_2-vektoren. Een deelruimte voortgebracht door de Q_1- of de Q_2-vektoren wordt uitgedrukt door lineaire kombinaties van deze vektoren. We zoeken dus gewichtsmatriksen A_1 (m_1xp) en A_2 (m_2xp) zodanig dat de kolommen van Q_1A_1 zoveel mogelijk lijken op de kolommen van Q_2A_2 en zodanig dat de kolommen van Q_1A_1 en Q_2A_2 ortogonale bases zijn. We willen dat het verschil tussen de overeenkomstige kolommen van Q_1A_1 en Q_2A_2 zo klein mogelijke kwadratensommen heeft, dwz $SSQ(Q_1A_1-Q_2A_2)$ moet minimaal zijn.De gelijkenis tussen de deelruimtes wordt uitgedrukt in de diagonaal van de matriks $A_1'Q_1'Q_2A_2/n$, de kanonisccne korrelatie koëfficienten. De ortogonale bases Q_1A_1 en Q_2A_2 noemen we de kanonische assen van de eerste en de tweede set. We noemen $(SSQ(Q_1A_1-Q_2A_2)/np)$ de verliesfunktie, de stress of $\sigma(Q,A)$.

CANALS is een model voor niet-lineaire kanonische korrelatie analyse (De Leeuw,1973)(Van der Burg en De Leeuw,1978). Behalve dat we bij CANALS gewichtsmatriksen zoeken zoals hiervoor beschreven staat, zoeken we bij CANALS ook een betere representatie van de variabelen. De enige eis die we aan de nieuwe variabelen stellen is dat ze in een ruimte liggen, die bepaald is door de toegestane transformaties, die afhangen van het meetnivo van de variabelen. Daarom kiezen we de nieuwe variabelen zodanig dat ze enerzijds zo goed mogelijk in het kanonische korrelatiemodel passen,anderzijds aan de schalingseisen van hun meetnivo voldoen. de verschillende meetnivo's zijn nominaal (enkelvoudig en meervoudig), ordinaal en numeriek. Daarnaast kunnen alle meetnivo's nog onderverdeeld worden in diskreet en kontinu (zie hfdst 8).In praktijk zijn nog

niet alle meetnivo's m.b.v. het programma CANALS te verwerken. De kontinue optie is nog niet klaar en de optie diskreet nominaal meervoudig is provisorisch opgelost door de desbetreffende variabelen in evenveel binaire variabelen te splitsen als ze kategorieën hebben. De eisen van het nominaal enkelvoudige nivo houden in dat alle herschalingen van een variabele van dit type voldoen aan het feit dat objekten die in dezelfde kategorie liggen in dezelfde kategorie blijven. De nivoeisen voor diskreet ordinale variabelen zijn identiek aan die van de nominale variabelen, plus nog een eis over de volgorde van de kategorieën. De eisen van het diskreet numerieke nivo laten alleen lineiare transformaties toe van de oorspronkelijke variabelen. De ruimte waarin iedere variabele j moet liggen om aan de eisen van het meetnivo te voldoen, noemen we de optimale schalingsruimte, deze wordt gesymboliseerd door C_j.

7.1.2. Het CANALS algoritme

Het CANALS algoritme kan als volgt geformuleerd worden:

Minimaliseer de verliesfunktie:
$(SSQ(Q_1A_1-Q_2A_2))/np = \sigma(Q,A)$ over Q_1, Q_2, A_1 en A_2
onder voorwaarde dat:
de kolommen van Q_1A_1 en Q_2A_2 ortogonaal en gestandaardiseerd zijn, en
de kolommen van matriks $Q=(Q_1|Q_2)$ overeenstemmen in meetnivo met de oorspronkelijke variabelen en gestandaardiseerd zijn.
De voorwaarden zien er mathematisch als volgt uit:
$A_1'Q_1'Q_1A_1=nI$, $A_2'Q_2'Q_2A_2=nI$, $q_j'q_j=n$ en $q_j \varepsilon C_j$ met $j=1,\ldots,m$.
De kolommen van Q kunnen herschreven worden als G_jz_j, waarbij G_j de indikatormatriks van variabele j is en z_j de vektor van kategorieskores van variabele j. De schaling van vektor q_j kan ook beschreven worden in termen van z_j en d_j de marginale frekwenties (kolomsommen van G_j). Aangezien we bij CANALS geen gebruik maken van deze herschrijving zullen we de notatie Q handhaven. Deze herschrijving is alleen van belang voor de relatie CANALS en HOMALS.

We willen het kanonische korrelatieprobleem oplossen met een alternerende kleinste kwadraten metode, vandaar de naam CANALS. Dit betekent dat we, wat de variabelen betreft alternerend de

modelparamaters en de optimale schalingsparameters willen oplossen met een kleinste kwadraten metode. Behalve de variabelen zijn er in het model ook nog parameters die van de dimensionaliteit van het kanonische korrelatieprobleem afhangen, nl de gewichtsmatriksen. We alterneren de berekening van de gewichten met de berekening van de variabelen. Voor de berekening van de gewichten willen we ook een kleinste kwadraten metode gebruiken.

Het oplossen van de <u>modelparameters</u> voor de variabelen q_j zouden we kunnen doen door de verliesfunktie naar q_j te differentiëren en het resultaat gelijk aan nul te stellen. Dit geeft:

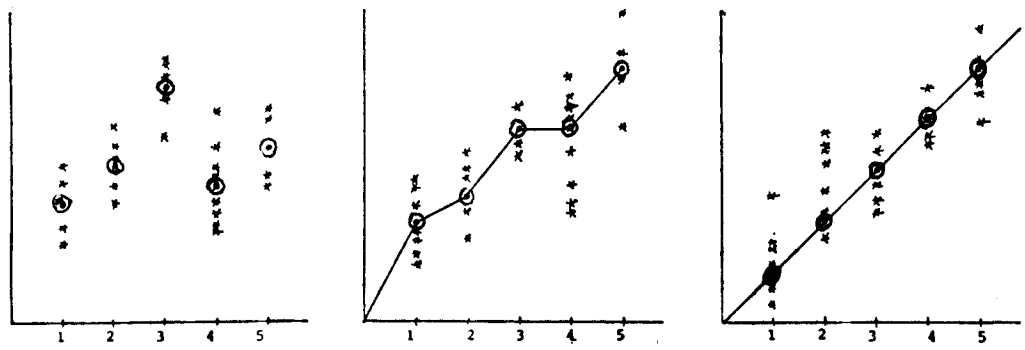
$q_j=(Q_2A_2-Q_1A_1)'a_j+a_j'a_jq_j$ voor $j=1,\ldots,m_1$ en

$q_j=(Q_1A_1-Q_2A_2)'a_j+a_j'a_jq_j$ voor $j=m_1+1,\ldots,m$

a_j' is de j-de rij van matriks $A\triangleq\left(\frac{A_1}{A_2}\right)$

We zijn nu even vergeten dat de kondities $A_1'Q_1'Q_1A_1=nI$ en $A_2'Q_2'Q_2A_2=nI$ van kracht waren. Sterkers zelfs, door een variabele van de eerste set te berekenen wordt de eerste konditie geschonden en door een variabele van de tweede set te berekenen wordt de tweede konditie geschonden.

De <u>optimale schalingsparameters</u> berekenen we door de kleinste kwadraten benadering van de optimale schalingsruimte te berekenen (vergelijk hfdst 8). In het diskreet enkelvoudig nominale geval bereiken we dit door voor alle objekten die in één kategorie liggen de <u>gemiddelde skore</u> te nemen i.p.v. de berekende skores die de verliesfunktie minimaliseren. Voor ordinale variabelen berekenen we de kleinste kwadraten benadering door een <u>monotone regressie</u> uit te voeren op de kategoriegemiddelden (Kruskal, secondary approach) en in het lineiare geval krijgen we een kleinste kwadraten benadering door een <u>lineaire regressie</u> uit te voeren van de berekende variabelen op de oorspronkelijke variabelen. Figuur 7.2. geeft de verschillende benaderingen weer. De vertikale as bevat de berekende waarden van een variabele, die gevonden worden door de verliesfunktie te minimaliseren. De horizontale as bevat de oorspronkelijke waarden van de variabele, in dit geval 1,....,5.



nominale schaling ordinale schaling numerieke schaling

Figuur 7.2. Drie typen regressie

Nadat de variabelen herschaald zijn, moeten ze ook nog gestandaardiseerd worden. Een uitgebreide bespreking van optimale schaling komt in hoofdstuk 8 aan de orde.

De modelparameters A_1 en A_2, die van de dimensionaliteit van de oplossing afhangen, kunnen we berekenen door $\sigma(Q,A)$ te minimaliseren onder voorwaarde dat $A_1'Q_1'Q_1A_1=nI$ en dat $A_2'Q_2'Q_2A_2=nI$. Minimalisatie van de verliesfunktie onder deze kondities, komt neer op maksimalisatie van spoor$(A_1'Q_1'Q_2A_2)$ onder dezelfde voorwaarden. Herschrijf spoor$(A_1'Q_1'Q_2A_2)$ als:

$$\text{spoor } (A_1(Q_1'Q_1)^{\frac{1}{2}}(Q_1'Q_1)^{-\frac{1}{2}}Q_1'Q_2(Q_2'Q_2)^{-\frac{1}{2}}(Q_2'Q_2)^{\frac{1}{2}}A_2) \qquad (7.1)$$

Substitutie van:

$B_1=(1/\sqrt{n})A_1(Q_1'Q_1)^{\frac{1}{2}}$,

$B_2=(1/\sqrt{n})A_2(Q_2'Q_2)^{\frac{1}{2}}$ en

$T=(Q_1'Q_1)^{-\frac{1}{2}}Q_2'Q_2(Q_2'Q_2)^{-\frac{1}{2}}$ in formule 7.1 levert:

maksimaliseer spoor$(B_1'TB_2)$ onder voorwaarde dat $B_1'B_1=I$ en $B_2'B_2=I$.

Dit wordt opgelost door B_1 gelijk aan de p eigenvektoren van TT' te maken die bij de p grootste eigenwaarden behoren, en door B_2 gelijk aan de p eigenvektoren van $T'T$ te maken die bij de p grootste eigenwaarden behoren. Schrijven we $T=Z\Lambda W'$ (SVD) dan geldt:

$A_1=\sqrt{n}(Q_1'Q_1)^{-\frac{1}{2}}Z$ en $A_2=\sqrt{n}(Q_2'Q_2)^{-\frac{1}{2}}W$.

Substitutie van A_1 en A_2 in de verliesfunktie geeft:

$\sigma(Q,*)=2-(2/p)\Sigma_{s=1}^{p}\lambda_s$, waarbij λ_s de p grootste eigenwaarden van TT' of $T'T$ zijn (vergelijk appendix A.1).

Na de berekening van de schaling en de standaardisatie van de variabelen en de berekening van de gewichten,beginnen we weer opnieuw bij de berekening van de variabelen. Het iteratieproces stopt als de stress niet meer genoeg afneemt.
Zoals het CANALS probleem nu opgelost is,vinden we het bij Young, De Leeuw en Takane (1976). We hebben hiervoor echter al vermeld dat bij de berekening van de variabelen altijd één der kondities m.b.t. de gewogen variabelen geschonden wordt. Daarom is deze oplossing niet goed, tenzij we te maken hebben met slechts één variabele in een van de sets.Dan is deze oplossing wel juist en is er sprake van konvergentie van het algoritme, omdat het gewicht van de set met maar één variabele in alle gevallen gelijk aan één blijft.

Aangezien het oplossen van de variabelen van de ene set niet gehinderd wordt door de konditie van de gewogen variabelen van de andere set, gaan we het CANALS algoritme enigszins herzien. We berekenen de modelparameters van de ene set onder voorwaarde dat de gewogen variabelen van de andere set ortogonaal zijn.Het CANALS algoritme ziet er nu als volgt uit:

1 minimaliseer $\sigma(Q,A)$ over A_1 onder voorwaarde dat
$A_2'Q_2'Q_2A_2=nI$, $q_j\varepsilon C_j$ en $q_j'q_j=n$ voor $j=1,\ldots,m$

2 voor $l=1,\ldots m_1$

2^a minimaliseer $\sigma(Q,A)$ over q_l onder voorwaarde dat
$A_2'Q_2'Q_2A_2=nI$, $q_j\varepsilon C_j$ en $q_j'q_j=n$ voor $j=1,\ldots,m$ en $j\neq l$

2^b zoek de kleinste kwadraten benadering van q_l die in C_l ligt en standaardiseer.

3 minimaliseer $\sigma(Q,A)$ over A_2 onder voorwaarde dat
$A_1'Q_1'Q_1A_1=nI$, $q_j\varepsilon C_j$ en $q_j'q_j=n$ voor $j=1,\ldots,m$

4 voor $l=m_1+1,\ldots,m$

4^a minimaliseer $\sigma(Q,A)$ over q_l onder voorwaarde dat
$A_1'Q_1'Q_1A_1=nI$, $q_j\varepsilon C_j$ en $q_j'q_j=n$ voor $j=1,\ldots,m$ en $j\neq l$

4^b Zoek de kleinste kwadraten benadering van q_l die in C_l ligt en standaardiseer

We zullen bewijzen dat het minimaliseren van de verliesfunktie onder één of twee kondities mbt de gewogen variabelen tot ekwivalente oplossingen van A_1 en A_2 leidt. Ekwivalent wil zeggen identiek op een ortogonale en/of diagonale rotatie na. Minimalisatie van $\sigma(Q,A)$ over A_1 en A_2 onder voorwaarde dat $A_1'Q_1'Q_1A_1=nI$ levert op dezelfde manier als hiervoor beschreven is:

$A_2=\sqrt{n}(Q_1'Q_1)^{-\frac{1}{2}}W$

A_1 lossen we op door de verliesfunktie te differentiëren:

$\partial\sigma/\partial A_1=0 \rightarrow A_1=\sqrt{n}(Q_1'Q_1)^{-1}Q_1'Q_2A_2=\sqrt{n}(Q_1'Q_1)^{-\frac{1}{2}}T(Q_2'Q_2)^{-\frac{1}{2}}A_2$, dus

$A_1=\sqrt{n}(Q_2'Q_2)^{-\frac{1}{2}}Z\Lambda$.

We zien dat A_1 en A_2 inderdaad ekwivalent zijn aan de A_1 en A_2 van het oorspronkelijke CANALS probleem. De minimum waarde van de verliesfunktie verkrijgen we door A_1 en A_2 in te vullen. Dit levert $\sigma(Q,*)=1-(1/p)\Sigma_{s=1}^{p}\lambda_s$, waarbij Λ de p grootste singuliere waarden van T bevat (zie appendix A.1).
Op analoge manier wordt minimalisatie van $\sigma(Q,A)$ over A_1 en A_2 onder voorwaarde dat $A_2'Q_2'Q_2A_2=nI$ opgelost. Dit levert:

$A_1=\sqrt{n}(Q_1'Q_1)^{-\frac{1}{2}}Z$

$A_2=\sqrt{n}(Q_2'Q_2)^{-\frac{1}{2}}W\Lambda$

$\sigma(Q,*)=1-(1/p)\Sigma_{s=1}^{p}\lambda_s$

Z en W zijn weer de eigenvektoren van TT' en $T'T$ behorende bij de p grootste eigenwaarden, Λ bevat deze eigenwaarden.
Alle oplossingen van A_1 en A_2 zijn dus ekwivalent. De minimale waarde van de verliesfunktie verschilt voor één of twee kondities. Als we één konditie nemen maakt het niet uit voor de waarde van σ wélke konditie we nemen.

De herziene benadering van het CANALS model is nog niet helemaal volledig. De eerste twee deelproblemen worden opgelost met de konditie $A_2'Q_2'Q_2A_2=nI$ en de laatste twee deelproblemen met de konditie $A_1'Q_1'Q_1A_1=nI$. We moeten nog transformaties van A_1 en A_2 vinden, zodanig dat we van de ene konditie op de andere overschakelen zonder dat we de stress veranderen. Geldt nu de eerste konditie dan kunnen we de symmetrische matriks $A_2'Q_2'Q_2A_2$ herschrijven als nSS', waarbij S een reguliere matriks is (wij hebben een Choleski dekompositie gekozen, S is dan een driehoeksmatriks (Stewart,1973)).

De transformaties S en $(S^{-1})'$ voor A_1 en A_2 voldoen aan de gestelde eis, nl

$SSQ(Q_1A_1-Q_2A_2) = tr(nI-2A_1'Q_1'Q_2A_2+nSS')$ en

$SSQ(Q_1A_1S-Q_2A_2(S^{-1})') = tr(nSS'-2A_1'Q_1'Q_2A_2+nI)$

Dit betekent dus, dat als we A_1 in A_1S en A_2 in $A_2(S^{-1})'$ veranderen, we van het ene deelprobleem op het andere deelprobleem overgestapt zijn zonder dat de oplossingen wezenlijk veranderd zijn. Voor het geval dat de tweede konditie geldig is, kunnen we de matriks $A_1'Q_1'Q_1A_1$ herschrijven als het produkt van twee reguliere matriksen. De rest gaat analoog.

De nieuwe benadering van het CANALS model is nu bijna volledig. We lossen probleem 1 en 2 op, dan schakelen we over op de andere konditie door A_1 en A_2 te transformeren zoals hiervoor beschreven is. Vervolgens lossen we probleem 3 en 4 op. Daarna transformeren we A_1 en A_2 zodanig, dat we weer terug zijn bij het eerste deelprobleem met de oude konditie en we beginnen de cyclus weer van voren af aan. Als de afname van de stress klein genoeg is, houden we op na het vierde deelprobleem. We zoeken dan nog nieuwe <u>transformaties</u> van A_1 en A_2 zodanig, dat we een stress krijgen die gelijk is aan de stress van het oorspronkelijke CANALS probleem en zodanig dat aan beide kondities voldaan is. Dit bereiken we door matriks A_1 met U en matriks A_2 met $U\Lambda^{-1}$ te vermenigvuldigen, waarbij $A_2'Q_2'Q_2A_2=U\Lambda^2U'$ (eigenvektordekompositie, appendix A.1).
De matriksen A_1U en $A_2U\Lambda^{-1}$ zijn de gezochte gewichtsmatriksen van het oorspronkelijke CANALS probleem. De matriks Λ bevat de p kanonische korrelatie koëfficienten.

De <u>oplossing</u> van de verschillende <u>deelproblemen</u> verkrijgen we door de verliesfunktie te differentiëren en het resultaat gelijk aan nul te stellen.
Het <u>eerste deelprobleem</u> kunnen we oplossen door te stellen:
$A_1 = n(Q_1'Q_1)^{-1}Q_1'Q_2A_2$, zoals eerder besproken is. Maar we willen liever de berekening van een inverse matriks vermijden. Daarom <u>benaderen</u> A_1 door te stellen:

$A_1 \triangleq A_1+\theta E_{js}$, $E_{js}=1$ voor element (j,s) en $E_{js}=0$ voor alle andere elementen, j=1,...,m en s=1,....,p.
Differentiatie van de verliesfunktie naar θ levert de kleinste

kwadraten benadering van θ op, nl

$\theta=q_j'(Q_1A_1-Q_2A_2)_s/n$,

$(Q_1A_1-Q_2A_2)_s$ is de s-de kolom van matriks $(Q_1A_1-Q_2A_2)$.
We benaderen de elementen van matriks A_1 niet één keer, maar verschillende keren, totdat er aan een iteratiekriterium voldaan is.

De oplossing van het tweede deelprobleem is reeds besproken bij de behandeling van het oorspronkelijke CANALS probleem.
$\partial\sigma/\partial q_1=0 \rightarrow q_1=(Q_2A_2-Q_1A_1)'a_1+a_1'a_1q_1$
In het rechter lid staat de oude waarde van q_1 en in het linker lid de nieuwe waarde van q_1.
Ook de schaling van de variabelen hebben we reeds besproken bij de behandeling van het oorspronkelijke CANALS probleem. Dit was voor nominale variabelen gelijk aan het berekenen van de kategoriegemiddelden, in het ordinale geval het doen van een monotone regressie op deze gemiddelden en in het numerieke geval een lineaire regressie op de variabelen.

Probleem 3 en 4 worden analoog aan probleem 1 en 2 opgelost:

$A_2 \triangleq A_2+\theta E_{js}$, $E_{js}=1$ voor element (j,s) en $E_{js}=0$ voor alle andere elementen, $j=m_1+1,....,m$ en $s=1,...,p$.

$\partial\sigma/\partial\theta=0 \rightarrow \theta=q_j'(Q_2A_2-Q_1A_1)_s$

$\partial\sigma/\partial q_1=0 \rightarrow q_1=(Q_1A_1-Q_2A_2)'a_1+a_1'a_1q_1$ en $l=m_1+1,...,m$

De schaling van de variabelen is identiek.

We hebben nu voor alle deelproblemen en alle tussentijdse transformaties een oplossing gevonden die in het programma CANALS gehanteerd worden.

7.1.3. Het CANALS programma

Zoals bij het CANALS algoritme reeds besproken is, zijn we geinteresseerd in herschalingen van de variabelen, gewichten en in de kanonische korrelaties. Hoe hoger de kanonische korrelatie, hoe beter de fit. De eerste kanonische korrelatie koëfficient is altijd groter dan de tweede, enz. I.h.a. moet een kanonische korrelatiekoëfficient hoog zijn, willen de bijbehorende assen bruikbaar zijn. Hoe hoog heeft ook met de korrelaties van de variabelen met de verschillende assen te maken (zie verder op). De herschalingen van de variabelen

zijn ook van belang om te weten met wat voor soort variabelen we te doen hebben. Blijkt het dat de transformaties niet lineair zijn dan betekent dit dat de kanonische korrelatie koëfficient verhoogd is door deze transformaties vergeleken met de kanonische korrelatiekoëfficient van een lineaire analyse. Onze ervaring is dat de schalingen van een variabele in het geval dat er sprake is van missing skores op die variabele, er nogal eens 'gedegenereerd' uit kunnen zien. Bv. alle nieuwe kategorieskores negatief en dichtbij nul. Dit komt omdat het CANALS model eerst de unieke patronen in de lineaire deelruimtes omspannen door de variabelen zoekt. De missing skores laten zich perfekt fitten, omdat aan deze skores geen schalingseisen gesteld worden (elke missing skore is een aparte kategorie). Maar ook andere unieke patronen, niet veroorzaakt door missing data, domineren snel de eerste kanonische assen. Mocht men niet geïnteresseerd zijn in deze uitbijters in de data set, dan moet men meer dimensies berekenen. Men kan ook een aantal individuen verwijderen of op een andere wijze kategoriseren.

De gewichten die berekend worden in het CANALS model geven de bijdrage van iedere herschaalde variabele tot de kanonische assen. De gewichten kunnen nogal eens verwarring wekken, aangezien de gewichten van de variabelen die onderling hoog korreleren (binnen een set) erg verschillend kunnen zijn. Daarom is het van belang de korrelaties van de herschaalde variabelen met de kanonische assen te interpreteren samen met de gewichten. Deze korrelaties geven een soort betrouwbaarheid van de gewichten weer.De kwadratensom van de korrelaties per as geeft de hoeveelheid verklaarde variantie door de desbetreffende as. En elke individuele korrelatie geeft in het kwadraat de verklaarde variantie van de desbetreffende variabele en de desbetreffende as (Thorndike,1977). Is er nu sprake van hoge interkorrelaties binnen sets en tussen sets van herschaalde variabelen, dan kunnen we hoge korrelaties met de kanonische assen verwachten, terwijl niet alle gewichten hoog hoeven te zijn. Als de interkorrelaties klein zijn binnen sets, dan zullen de gewichten en de korrelaties meer met elkaar overeen stemmen. Dit blijkt ook uit de formule van de korrelaties,nl

$Q_1'Q_1A_1/n$, $Q_1'Q_2A_2/n$, $Q_2'Q_1A_1/n$ en $Q_2'Q_2A_2/n$.

De korrelaties zijn gewogen sommen van de interkorrelaties. Daarom moeten we dus altijd de gewichten interpreteren aan de hand van de korrelaties met de kanonische assen (zie 7.2.3). Stewart en Love (1968) stellen zelfs voor alleen de korrelaties en niet de gewichten te interpreteren. Er is veel geschreven over instabiliteit van regressiegewichten (bv Cooley en Lones, 1971). Instabiliteit van gewichten geldt bij het kanonische korrelatie model dubbel zo hard als bij het multiple regressie model, aangezien kanonische korrelatie analyse als een dubbele regressie analyse opgevat kan worden. In het algemeen gaat men er van uit dat de korrelaties stabieler zijn dan de gewichten (Thorndike en Weiss,1973). We hebben dit willen aantonen in een bootstrapstudie (zie hfdst 9) van CANALS, maar dit konden we niet zondermeer bevestigen. Wel hebben we gevonden dat de kanonische korrelatie koëfficienten veel stabieler zijn dan de gewichten. We moeten deze bootstrapstudie nog verder uitwerken.
De korrelaties van de variabelen met de kanonische assen worden tesamen wel de kanonische strukturen genoemd en de individuele korrelaties kanonische ladingen. Dit laatste naar analogie met principale komponenten analyse. De kanonische assen kunnen nl als latente variabelen opgevat worden, kanonische korrelatie analyse is dan een dubbele principale komponenten analyse. Het verschil met principale komponenten analyse is dat het kriterium op grond waarvan de assen gekozen worden bij kanonische korrelatie analyse anders is dan bij principale komponenten analyse.

7.1.4. Toepassingen van CANALS

7.1.4.1. Ekonomische ongelijkheid en politieke stabiliteit.

De data voor dit voorbeeld komen uit een artikel van de politicoloog B.M. Russett, gepubliceerd in de reader Quantitative History (1969).
Russett verzamelde gegevens over 47 landen met betrekking tot verschillende indikatoren van ekonomische aard, die hij in verband wilde brengen met wat we hier politieke variabelen zullen noemen. We gaan niet in op de manier waarop Russett de diverse variabelen gekwantificeerd heeft.
Russett gaat er van uit dat de wijze waarop het bouwland verdeeld is, een maatstaf is voor de mate van ongelijkheid en gebruikt hiervoor twee indikatoren: de GINI-indeks en het percentage boeren dat tesamen de helft van de grond in bezit heeft (in dit voorbeeld aangeduid met VERD). De GINI-indeks geeft het verschil tussen de feitelijke en de ideale verdeling van het land. Ideaal is dat alle boeren evenveel land ter beschikking hebben. Hoe hoger de GINI-indeks, des te ongelijker is het land verdeeld onder de boeren. Bij de variabele VERD geldt eveneens "hoe hoger, hoe ongelijker".
Als derde indikator wordt het percentage boerenbedrijven opgevoerd, dat de grond in pacht heeft (PACH).
Als ekstra ekonomische variabelen neemt Russett het bruto nationaal produkt (BRNP) per hoofd van de bevolking in 1955 (in $1055). en het percentage van de beroepsbevolking dat werkzaam is in de landbouw (LARB); Deze variabele zegt tegelijk iets over de mate van industrialisatie.
De politieke stabiliteit van een land wordt geoperationalisserd door vier variabelen. De eerste, de instabiliteit van het leiderschap, is een funktie van het aantal politieke leiders en het aantal jaren dat een land onafhankelijk geweest is in de periode 1945-1961. Een lage indeks betekent een stabiel leiderschap. Deze variabele wordt aangegeven met LEID.
De tweede politieke variabele is het aantal gewelddadigheden in dezelfde periode, variërend van samenzweringen tot langdurige guerilla-aktiviteiten (GEWE).
De derde variabele is het het aantal doden ten gevolge van burgeroorlog, revolutie en rellen van 1950 tot 1962, per miljoen inwoners (BURG).

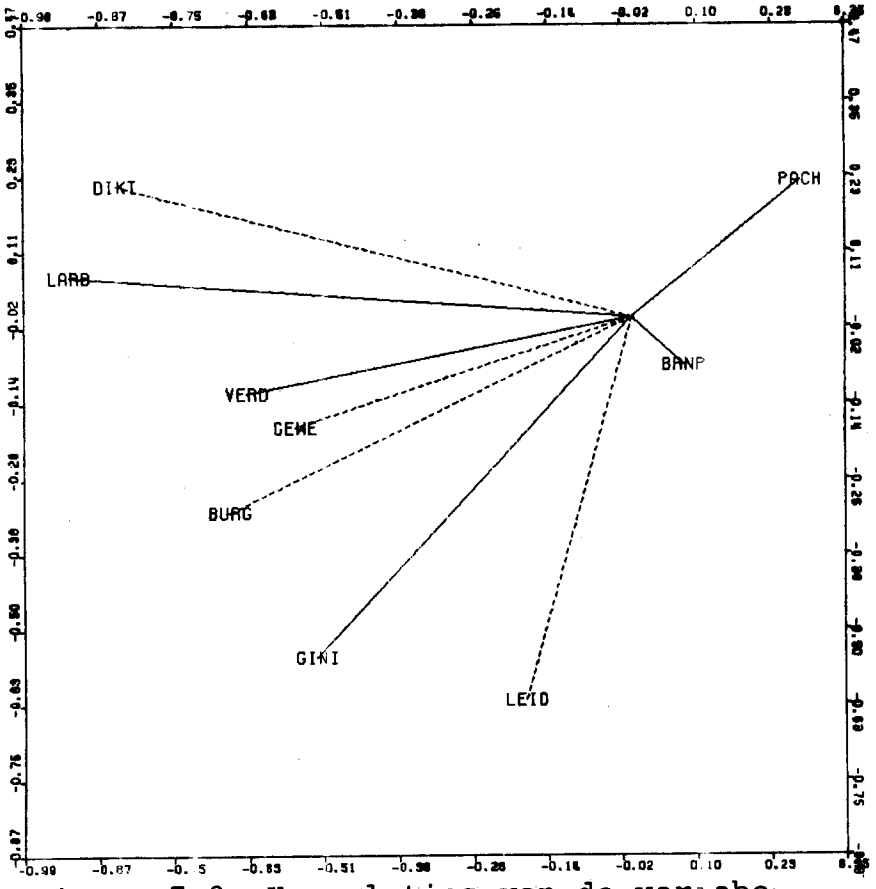
Als vierde politieke variabele oppert Russett de stabiliteit van de demokratie (DIKT); stabiel wil zeggen een ononderbroken voortduren van demokratie sinds de 1e wereldoorlog en het ontbreken van een totalitaire partij die bij verkiezingen meer dan 20 procent van de stemmen kreeg (over de laatste 30 jaar). Onstabiele demokratieën voldoen niet aan deze kriteria, maar hebben wel sinds de 1e wereldoorlog min of meer vrije verkiezingen gehad. In diktaturen zijn vrije verkiezingen in het algemeen niet voor gekomen. Russett merkt op, dat hij aan de hand van deze kriteria de landen niet kontinu kan rangordenen (zoals hij bij de andere variabelen doet), maar genoegen moet nemen met deze drie-deling en zo geen korrelaties uit kan rekenen.

	GINI	VERD	PACH	BRUP	LAND	LEID	GEWE	BUR	DIK	EKON.VAR	POL.VAR
ARGENTINIE	86.3	98.2	32.9	374	25	13.6	57	217	2	5 5 4 4 2	5 3 4 2
AUSTRALIE	92.9	99.6	----	1215	14	11.3	0	0	1	6 5 9 7 1	4 1 1 1
BELGIE	58.7	85.8	62.3	1015	10	15.5	8	1	1	2 3 5 6 1	6 2 2 1
BOLIVIA	93.8	97.7	20.0	66	72	15.3	53	663	3	6 5 3 1 4	6 3 5 3
BRAZILIE	83.7	98.5	9.1	262	61	15.5	49	1	2	5 5 2 4 4	6 3 2 2
CANADA	49.7	82.9	7.2	1667	12	11.3	22	0	1	1 2 2 7 1	4 2 1 1
CHILI	93.8	99.7	13.4	180	30	14.2	21	2	2	6 5 3 3 2	5 2 2 2
COLUMBIA	84.9	98.1	12.1	330	55	14.6	47	316	2	5 5 3 4 3	5 3 4 2
COSTARICA	88.1	99.1	5.4	307	55	14.6	19	24	2	5 5 2 4 3	5 2 3 2
CUBA	79.2	97.8	53.8	361	42	13.6	100	2900	3	4 5 5 4 3	5 3 5 3
DENEMARKEN	45.8	79.3	3.5	913	23	14.6	0	0	1	1 1 2 6 2	5 1 1 1
DOMIN.REP	79.5	98.5	20.8	205	56	11.3	6	31	3	4 5 3 3 3	4 2 3 3
ECUADOR	86.4	99.3	14.6	204	53	15.1	41	18	3	5 5 3 3 3	6 3 2 3
EGYPTE	74.0	98.1	11.6	133	64	15.8	45	2	3	4 5 3 2 4	6 3 2 3
EL SALVADOR	82.8	98.8	15.1	144	63	15.1	9	2	3	5 5 3 3 4	6 2 2 3
FILIPPIJNEN	56.4	88.2	37.3	101	59	14.0	15	292	3	2 3 4 3 5	5 2 4 1
FINLAND	59.9	86.3	2.4	941	46	15.6	4	0	2	2 3 2 6 3	6 2 1 2
FRANKRIJK	58.3	86.1	26.0	1046	26	16.3	46	1	2	2 3 4 6 2	6 3 2 2
GUATAMALA	86.0	99.7	17.0	179	68	14.9	45	57	3	5 5 3 3 4	5 3 3 3
GRIEKENLAND	74.7	99.4	17.7	139	48	15.8	9	2	2	4 5 3 3 3	6 2 2 2
GROOTBRITTAN.	71.0	93.4	44.5	998	5	13.6	12	0	1	4 4 5 6 1	5 2 1 1
HONDURAS	75.7	97.4	16.7	137	66	13.6	45	111	3	4 5 3 2 4	5 3 4 3
IERLAND	59.8	85.9	2.5	509	40	14.2	9	0	1	2 3 2 5 2	5 2 1 1
INDIA	52.2	86.9	53.0	72	71	3.0	83	14	1	2 3 5 1 4	2 3 2 1
IRAK	88.1	99.3	75.0	195	81	16.2	24	344	3	5 5 5 3 5	6 2 4 3
ITALIE	80.3	98.0	23.8	442	29	15.5	51	1	2	5 5 4 5 2	6 3 2 2
JAPAN	47.0	81.5	2.9	240	40	15.7	22	1	2	1 2 2 3 2	6 2 2 2
JOEGOSLAVIE	43.7	79.8	0.0	297	67	0.0	9	0	3	1 1 1 4 4	1 2 1 3
LUXEMBURG	63.8	87.7	18.8	1194	23	12.8	0	0	1	3 3 3 7 2	5 1 1 1
LYBIE	70.0	93.0	8.5	90	75	14.8	8	0	3	3 4 2 1 4	5 2 1 3
NEDERLAND	60.5	86.2	53.3	708	11	13.6	2	0	1	3 3 5 6 1	5 2 1 1
NICARAGUA	75.7	96.4	----	254	68	12.8	16	16	3	4 5 9 4 4	5 2 2 3
NWZEELAND	77.3	95.5	22.3	1259	16	12.8	0	0	1	4 5 4 7 1	5 1 1 1
NOORWEGEN	66.9	87.5	7.5	969	26	12.8	1	0	1	3 3 2 6 2	5 2 1 1
OOSTENRIJK	74.0	97.4	10.7	532	32	12.8	4	0	2	4 5 2 5 2	5 2 1 2
PANAMA	73.7	95.0	12.3	350	54	15.6	29	25	3	4 4 3 4 3	6 2 3 3
PERU	87.5	96.9	----	140	60	14.6	23	26	3	5 5 9 2 3	5 2 3 3
POLEN	45.0	77.7	0.0	468	57	8.5	19	5	3	1 1 1 5 3	3 2 2 3
SPANJE	78.0	99.5	43.7	254	50	0.0	22	1	3	4 5 5 4 3	1 2 2 3
TAIWAN	65.2	94.1	40.0	102	50	0.0	3	0	3	3 4 4 2 3	1 2 1 3
URUQUAY	81.7	96.6	34.7	569	37	14.6	1	1	1	5 5 4 5 2	5 2 2 1
VENEZUELA	90.9	99.3	20.6	762	42	14.9	36	111	3	6 5 3 6 3	5 2 4 3
VER.STATEN	70.5	95.4	20.4	2343	10	12.8	22	0	1	4 5 3 8 1	5 2 1 1
WSTDUITSLAND	67.4	93.0	5.7	762	14	3.0	4	0	2	3 4 2 6 1	2 2 1 2
ZUIDVIETNAM	67.1	94.6	20.0	133	65	10.0	50	1000	3	3 4 3 2 4	4 3 5 3
ZWEDEN	57.7	87.2	18.9	1165	13	8.5	0	0	1	2 3 3 7 1	3 1 1 1
ZWITSERLAND	49.8	81.5	18.9	1229	10	8.5	0	0	1	1 2 3 7 1	3 1 1 1

Tabel 7.1. Datamatrix, oorspronkelijk (links) en gecategoriseerd (rechts)

Het specifieke van CANALS is nu juist hierin gelegen dat zo'n kategorische variabele geen enkel probleem vormt. Om dit te benadrukken hebben we ook de andere variabelen gekategoriseerd, zoals weergegeven in tabel 7.1.

We hebben de gekategoriseerde data met CANALS geanalyseerd, eerst in twee, later ook in drie dimensies, beide met de ordinale optie. De eerste set bestaat uit de 5 ekonomische variabelen en de tweede uit de 4 politieke variabelen.
We bespreken eerst een aantal aspekten van de twee-dimensionale oplossing.

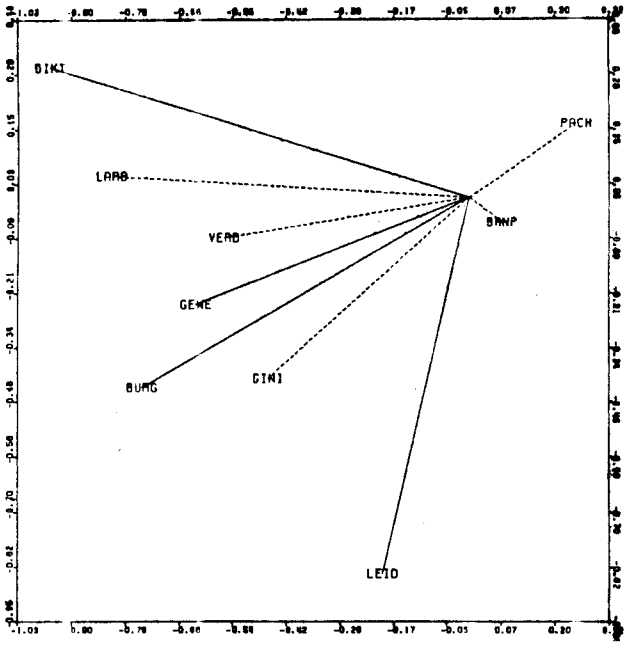


Figuur 7.3. Korrelaties van de variabelen met de kanonische assen van de 1e set

Figuur 7.3. geeft de korrelaties van de twee groepen variabelen weer in de ruimte van de ekonomische variabelen, de politieke variabelen zijn in deze ruimte geprojekteerd en weergegeven d.m.v. stippellijnen. De vektoren wijzen in de volgende richtingen:

DIKT - diktatuur
LARB - hoog perct. landarbeiders
GEWE - veel gewelddadigheden
BURG - veel doden
PACH - hoog perct. gepacht land
VERD - ongelijke verdeling
GINI - idem
LEID - instabiel leiderschap
BRNP - hoog bruto nat. produkt

We zien hoe de diverse variabelen met elkaar samenhangen, de lengte van een vektor geeft aan hoe belangrijk die variabele is.

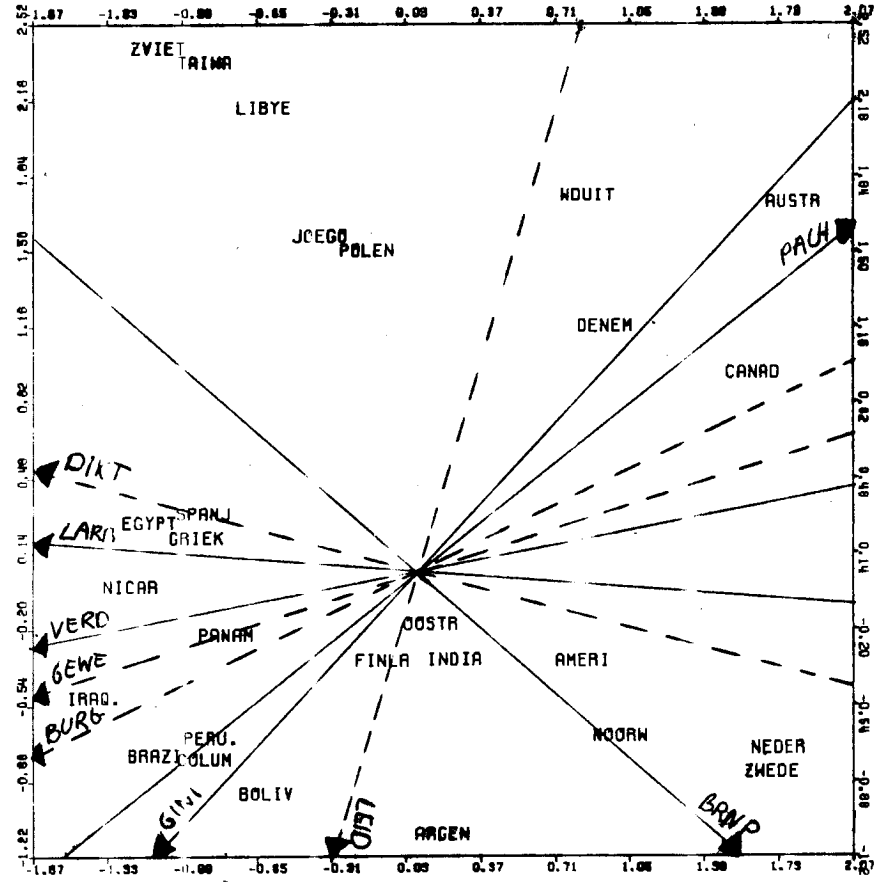


Figuur 7.4. Korrelaties van de variabelen met de kanonische assen van de 2e set

Figuur 7.4. geeft de variabelen weer in de ruimte van de politieke variabelen. De relaties tussen de vektoren zijn praktisch hetzelfde als die in figuur 7.3.; de kanonische korrelaties zijn dan ook .89 en .74. Daarom bespreken we beide figuren tegelijk.

We zien dat van de ekonomische variabelen het percentage landarbeiders (LARB) en de GINI-indeks de belangrijkste rol spelen en dat bruto nationaal produkt nauwelijks meetelt. Van de politieke variabelen valt ons op, dat DIKT en LEID belangrijk zijn, maar dat deze variabelen helemaal niets over elkaar zeggen. Dit komt ons niet vreemd voor, aangezien diktaturen ofwel zeer regelmatig van politiek leider wisselen (militaire staatsgrepen) of juist jarenlang dezelfde politikus aan het hoofd hebben (Joegoslavië). Demokratieën echter worden bijna vanzelfsprekend gekenmerkt door een tamelijk wisselend leiderschap.
Diktatuur hangt nauw samen met een hoog percentage landarbeiders, instabiel leiderschap nog het meest met een ongelijke verdeling volgens de GINI-indeks. GINI samen met VERD heeft een relatie met het voorkomen van geweld in een land. De variabele PACH speelt niet zo'n grote rol, maar hangt wel nauw samen met een eerlijke landverdeling.

Het is bij kanonische korrelatie analyse niet de gewoonte om naar de individu-skores (d.w.z. individuen geprojekteerd in de ruimte van de kanonische assen) te kijken. Wij waren hier echter wel in geïnteresseerd, hebben daarom de individuskores berekend en geplot in figuur 7.5. en 7.6.

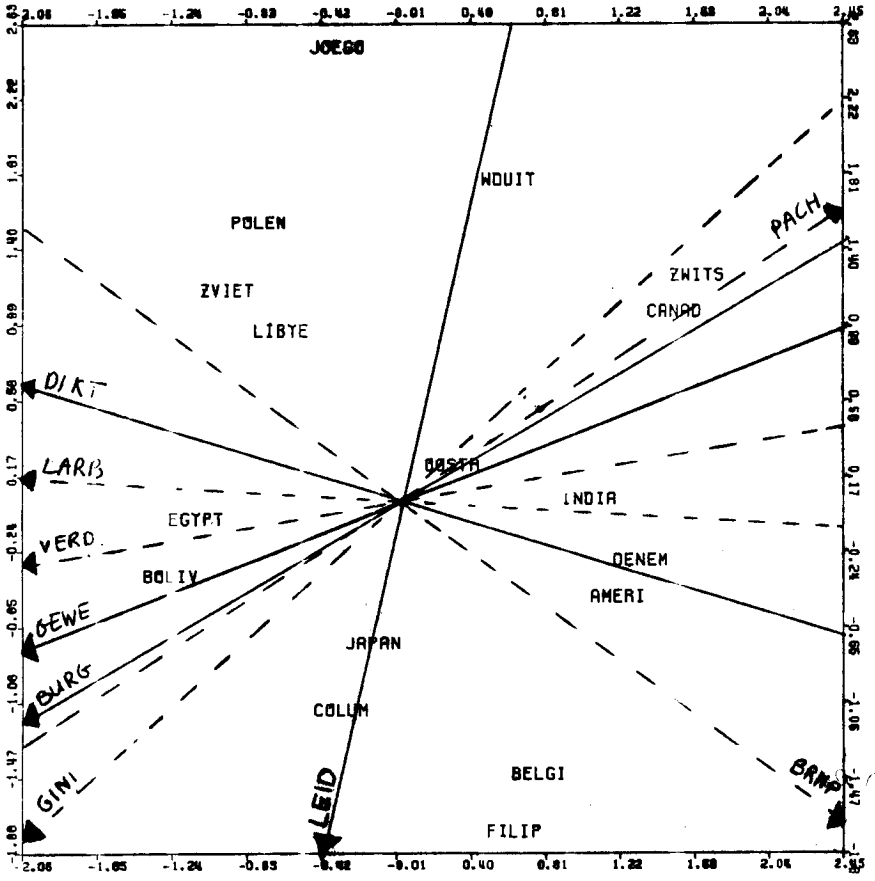


Figuur 7 5. Individu-skores in de kanonische ruimte van de le set

De variabelen worden weer aangegeven met vektoren. De lengte van de vektoren hebben we geen rol laten spelen, van belang is nu de projektie van de landen op deze vektoren.
De volgende landen vormen een kluster en zijn in figuur 7.5. met één label aangegeven.

ARGENTINIE	BOLIVIA	BRAZILIE	CANADA	COLUMBIA	DENEMARKEN	EGYPTE
Chili	Costa Rica	Guatamala	Zwitserl	Ecuador	Japan	Honduras
Italië		El Salvad		Venezuela		
FINLAND	GRIEKENL	NEDERLAND	N O O R W E G E N		VER.STATEN	SPANJE
Filipp	Domin.Rep	België	Frankr	Gr.Britt	Nw.Zeeland	Cuba
			Luxemb	Ierland		

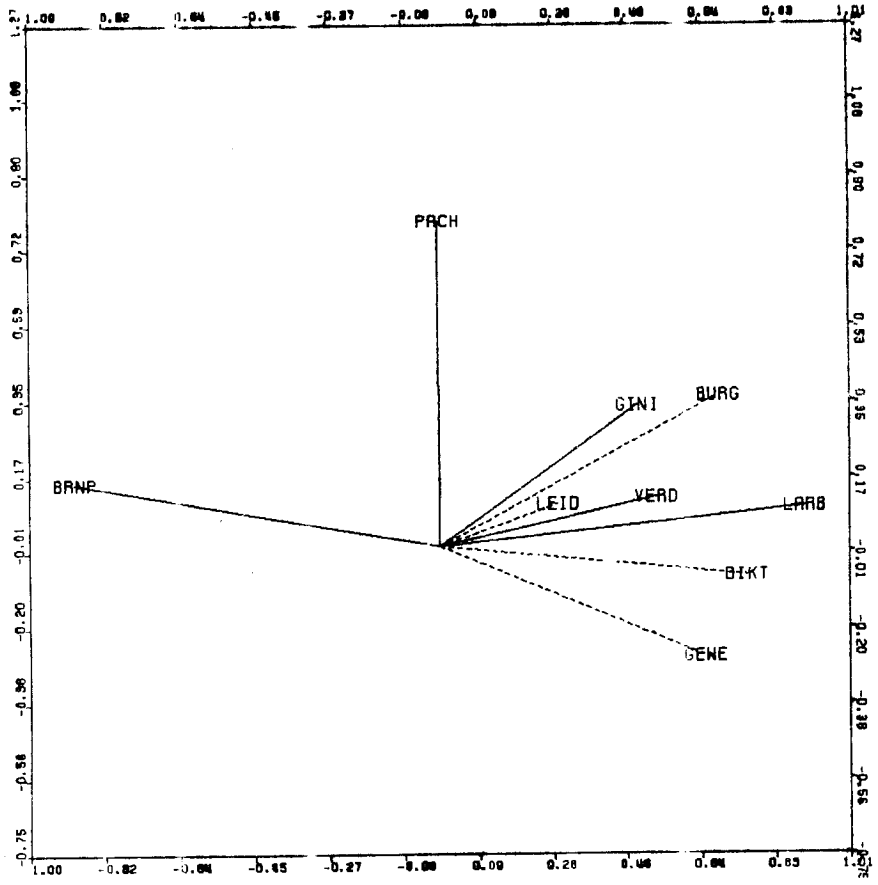
We kijken naar de belangrijkste trends in de ruimte van de ekonomische variabelen.
De landen met een hoog percentage landarbeiders (b.v. Irak, Brazilië, Zuid-Vietnam, Joegoslavië) projekteren hoog op LARB en zitten dus links in de plot. De GINI-indeks echter verdeelt deze landen: links onder de landen met een zeer ongelijke verdeling (Irak), links boven de landen waar het "eerlijker" toegaat. Rechts in de plot liggen de geïndustrialiseerde landen (deze projekteren laag op LARB) met weer onder de "oneerlijken" en boven in de plot de "eerlijken". Australië en WestDuitsland horen eigenlijk niet bij de landen met een gelijke verdeling te liggen. Ze zijn er terecht gekomen door de politieke variabelen DIKT en LEID. West-Duitsland is een West-Europees land, dat gekenmerkt wordt door een instabiele demokratie en een stabiel leiderschap, hetgeen uniek is. De variabele instabiliteit van het leiderschap verdeelt de plot in tweeën: boven in de plot liggen de landen met een stabiel leiderschap, onderin de "instabielen".
De variabele diktatuur geeft een perfekte drie-deling, van links in de plot de diktaturen, in het midden de instabiele en rechts de stabiele demokratieën. Voor Australië geldt iets dergelijks als West-Duitsland (ook dit land komt in de richting van stabiel leiderschap), waar nog bijkomt dat dit land een missing skore heeft op de variabele PACH.
We vinden deze drie-deling door de variabele diktatuur nog overzichtelijker terug in de plot van de individu-skores in de ruimte van de politieke variabelen (figuur 7.6.)
De variabele BURG echter zorgt voor enige verschuiving. Aangezien BURG de landen ongeveer indeelt in géén doden en wel doden (het maakt niet uit hoeveel), komen Joegoslavië en Taiwan samen te vallen, verhuizen Oostenrijk en Finland in de richting van het stabiele leiderschap en verlaat België zijn buurland Nederland en de andere Westerse landen en komt samen te liggen met Uruguay. Dat figuur 7.6. een nog sterkere klustering laat zien dan figuur 7.5. komt in de eerste plaats door de "ties" in de politieke variabelen, maar vooral ook door de transformaties van die data. We hebben de transformaties van de variabelen afgezet tegen de originele gegevens (dus niet tegen de gekategoriseerde data) ; deze zijn terug te vinden in figuur 7.9.



Figuur 7.6. Individu-skores in de kanonische ruimte van de 2e set

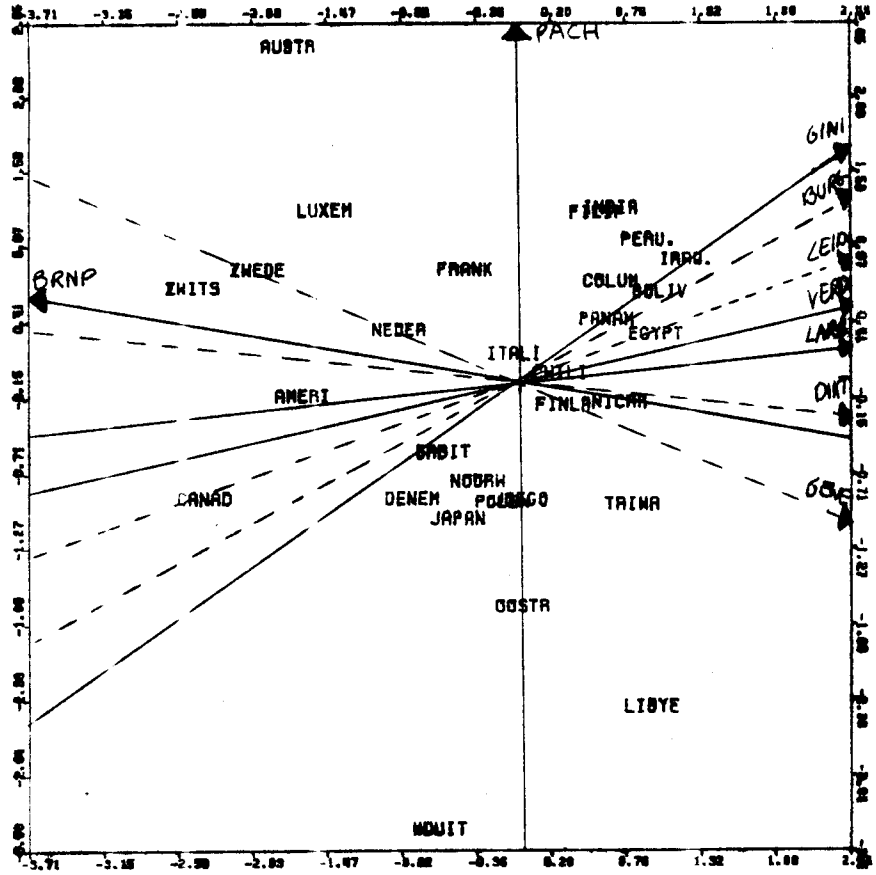
KLUSTERS						
BELGIE	BOLIVIA	COLUMBIA	DENEMARKEN	EGYPTE	OOSTENR	POLEN
Uruguay	Panama	Costa R.	Luxemburg	Nicarag	Finland	Spanje
	Honduras	Argentin	Nw.Zeeland	El Salv		
	Cuba			Ecuador		
V E R. S T A T E N		J A P A N		ZWITSERLAND	JOEGOSL	ZDVIETNAM
Ierland	Gr.Britt.	Frankr	Griekenl	Australië	Taiwan	Dom.Rep.
Noorweg	Nederland	Italië	Brazilië	Zweden		
		Chili				

We zijn enigszins verbaasd dat de variabele bruto nationaal produkt geen enkele rol speelt, terwijl we vermoeden dat deze variabele een hoge negatieve korrelatie heeft met het percentage landarbeiders. Daarom hebben we de analyse ook in drie dimensies gedaan. Ons vermoeden wordt bevestigd: bruto nationaal produkt korreleert hoog negatief met de 1e kanonische as en het percentage landarbeiders hoog positief (resp. -.893 en .899). De eerste 2 dimensies van deze oplossing geven



Figuur 7.7. Korrelaties van de variabelen met de 1e en 3e kanonische as van de 1e set

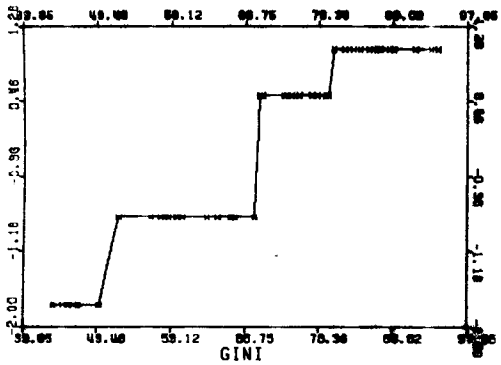
verder hetzelfde beeld van de andere variabelen: instabiel leiderschap korreleert hoog met de 2e kanonische as, diktatuur met de 1e (de vektoren staan weer loodrecht op elkaar), GINI en VERD spelen allebei een grote rol en GEWE is nu iets belangrijker dan BURG. Daarom laten we hier de eerste versus de derde dimensie zien, waar de variabele PACH een grote rol gaat spelen, maar weinig te maken heeft met BRNP en LARB.
De kanonische korrelaties zijn .876 .774 .646; we geven alleen de korrelaties en individu-skores in de kanonische ruimte van de ekonomische variabelen in een figuur weer. De landen klusteren in de ruimte van de 1e set minder dan bij onze twee-dimensionale oplossing; in de ruimte van de politieke variabelen vertonen zij exakt dezelfde klustering als die we gezien hebben in figuur 7.6. van de twee-dimensionale oplossing.

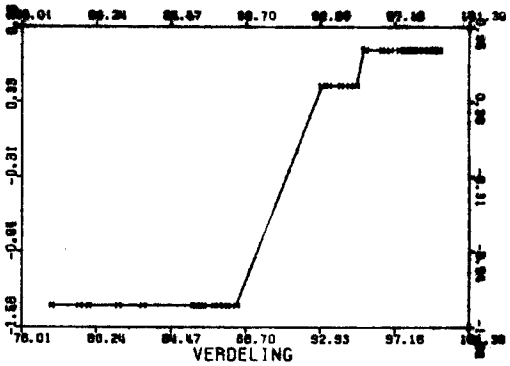


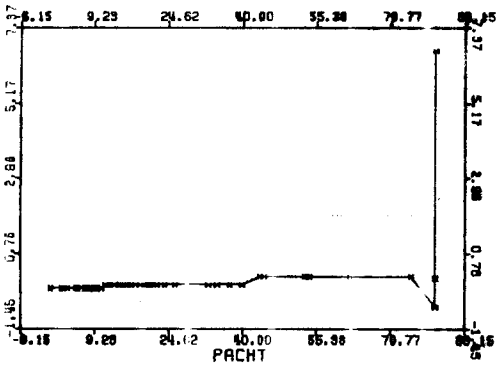
Figuur 7.8. Individu-skores in de kanonische ruimte van de ekonomische variabelen

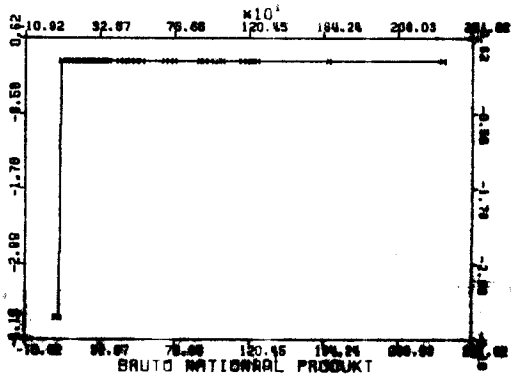
KLUSTERS				
BRAZILIE	BOLIVIA	PANAMA	EGYPTE	NOORWEGEN
ZdVietnam	Ecuador	Spanje	Dom.Rep.	Ierland
Costa R.	El Salv	Cuba	Griekenl	
Taiwan	Guatam.		Honduras	
COLUMBIA	ITALIE	NEDERLAND	VER.STATEN	
Venezuela	Uruguay	Belgie	NwZeeland	
	Argent.			

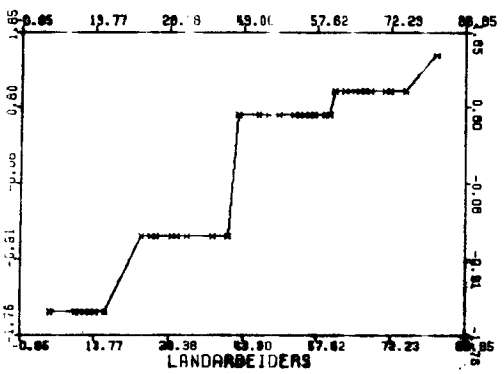
Figuur 7.8. laat zien dat BRNP de rijke van de arme landen scheidt, ongeveer op dezelfde wijze als LARB dat doet. De tegenstelling gaat dus ook tussen wél en niet-geïndustrialiseerde landen.
In deze laatst genoemde landen is meestal sprake van een diktatuur (de vektoren van BRNP en DIKT liggen bijna in elkaars verlengde).
De variabele PACH trekt zich van deze indeling weinig aan. Er zijn arme en rijke landen waar veel (of weinig) gepacht wordt, in diktaturen wordt òf heel veel gepacht van rijke landeigenaren (Zuid-Amerikaanse keuterboeren) of er wordt helemaal niet gepacht (Joegoslavië en Polen).



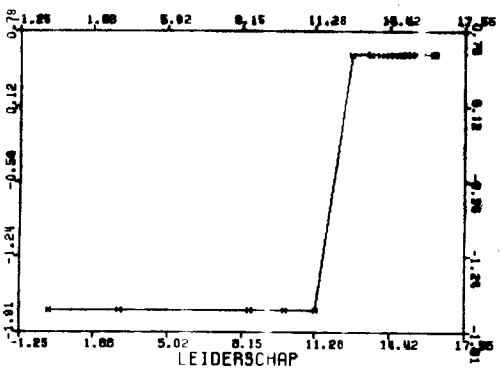


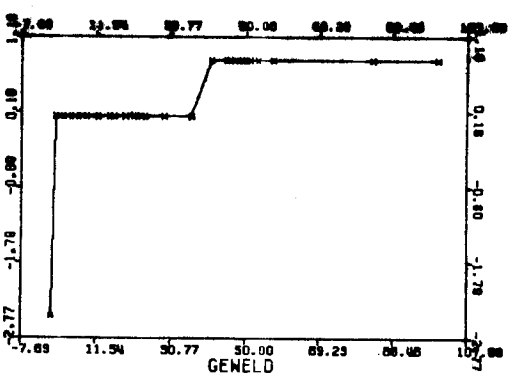


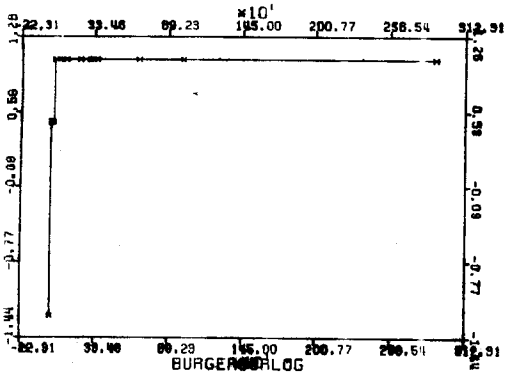


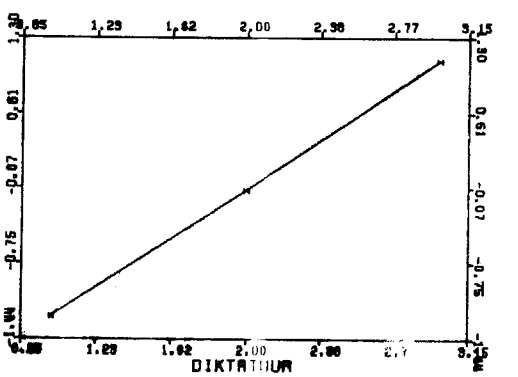


De onderstaande transformaties van de politieke variabelen laten zien dat deze variabelen bijna gedichotomiseerd zijn. Hetzelfde geldt voor BRNP. De belangrijke variabelen LARB en GINI geven onze kategorisering nog aardig te zien. De deuk in PACHT wordt veroorzaakt door de missing data.









Figuur 7.9. Data vóór kategorisering (horizontaal) versus getransformeerde skores (vertikaal) van de twee-dimensionale CANALS oplossing

7.1.4.2. Partijvoorkeur en standpunten in de Tweede Kamer

In 1972 is er een enquete gehouden onder de leden van de Tweede Kamer (zie o.a. Daalder en van de Geer, 1977). Van de 150 kamerleden hebben er 141 meegewerkt. Er is onder andere gevraagd naar de mening van de kamerleden over een aantal issues, te weten,ontwikkelingshulp, abortus, orde, inkomensverschillen, medezeggenschap, belastingen en defensie. Deze meningen zijn weergegeven op een negenpuntsschaal. Voor elk onderwerp was de laagste en hoogste kategorie gedefinieerd (zie tabel 7.2.). Verder hebben de kamerleden de partijen moeten ordenen naar verwantschap. Dit leverde 14 partijvoorkeuren.De skores op deze partijvoorkeuren geven aan welk rangnummer ieder kamerlid aan de desbetreffende partij heeft gegeven. Hoewel deze rangnummers rij-kondi-tioneel zijn verzameld, behandelen we ze hier zonder meer kolom-konditioneel. Bovendien hebben we alleen de voorkeuren voor de vier grootste partijen bekeken (te weten, voor PvdA, ARP, KVP en VVD). Behalve issues en voorkeuren zijn ook de partijlidmaatschappen van de kamerleden bekend. De issues gekombineerd met de

	1. ONTWIKK	
de overheid moet veel meer geld aan ontwikkelingshulp uitgeven	(1).....(9)	de overheid moet veel minder geld aan ontwikkelingshulp uitgeven
	2. ABORTUS	
de overheid moet abortus volledig verbieden	(1).....(9)	de vrouw moet zelf over abortus beslissen
	3. ORDE	
de overheid treedt te hard op tegen ordeverstoringen	(1).....(9)	de overheid moet harder optreden tegen ordeverstoringen
	4. INKOMVS	
de verschillen in inkomen moeten zo blijven	(1).....(9)	de verschillen in inkomen moeten kleiner worden
	5. MEDEZEG	
in de bedrijven moet alleen de direktie beslissen	(1).....(9)	de werknemers moeten medezeggenschap hebben
	6. BELASTI	
de belasting moet verhoogd worden t.b.v. algemene voorzieningen	(1).....(9)	de belasting moet verlaagd worden, zodat iedereen zelf kan beslissen wat hij met zijn geld doet
	7. LEGERS	
De regering moet aandringen op inkrimping van legers	(1).....(9)	de regering moet aandringen op handhaving van sterke legers

Tabel 7.2. De zeven issues met de betekenis van hun polen

voorkeuren hebben we gebruikt in een niet-lineaire kanonische korrelatie analyse en de issues gekombineerd met het partijlidmaatschap in een niet-lineaire kanonische diskriminant analyse (zie 7.4.2.).

Voor zover ons bekend is, zijn de meningen van de kamerleden over de issues en de partijvoorkeuren altijd apart geanalyseerd. Wij willen hier nagaan of de kamerleden die een voorkeur hebben voor b.v. de VVD er nu echt andere meningen op na houden dan kamerleden met een voorkeur voor de PvdA. Daartoe hebben we een CANALS analyse gedaan op de meningen over de issues enerzijds en de voorkeur voor de 4 grootste partijen anderzijds. Van de genoemde 141 kamerleden heeft er één geen mening gegeven over de issues, en heeft een ander zich niet over zijn partijvoorkeuren uitgesproken. Drie kamerleden antwoordden niet op 2 of 3 issues, en één heeft zijn verwantschap met slechts 1 van de 4 door ons te analyseren partijen uitgesproken. Deze 6 kamerleden hebben we niet in de analyse opgenomen.

Een twee-dimensionale analyse met ordinale opties leverde een kanonische korrelatie van 1.00 in de eerste dimensie. Dit wijst op perfekte fit (zie 7.1.3.). Om erachter te komen waardoor dit veroorzaakt werd, hebben we de ruwe data goed bekeken; met name van die kamerleden, die een missing skore op de variabele "orde" hadden. Deze variabele had nl. een korrelatie van -1.00 met de eerste kanonische as. Er bleek 'n kamerlid te zijn die geen mening over orde had uitgesproken. Omdat dit kamerlid zich tamelijk extreem uitliet over de KVP, korreleerde verwantschap met de KVP ook -1.00 met de eerste kanonische as. Dit kamerlid veroorzaakte zo het beeld dat de mening "de overheid moet harder optreden" perfekt samenhangt met een grote afkeer van de KVP. Toen we de data zonder het desbetreffende kamerlid analyseerden verdween de bovengeschetste relatie. De hierna gegeven resultaten zijn derhalve gebaseerd op de uitspraken van 134 kamerleden.

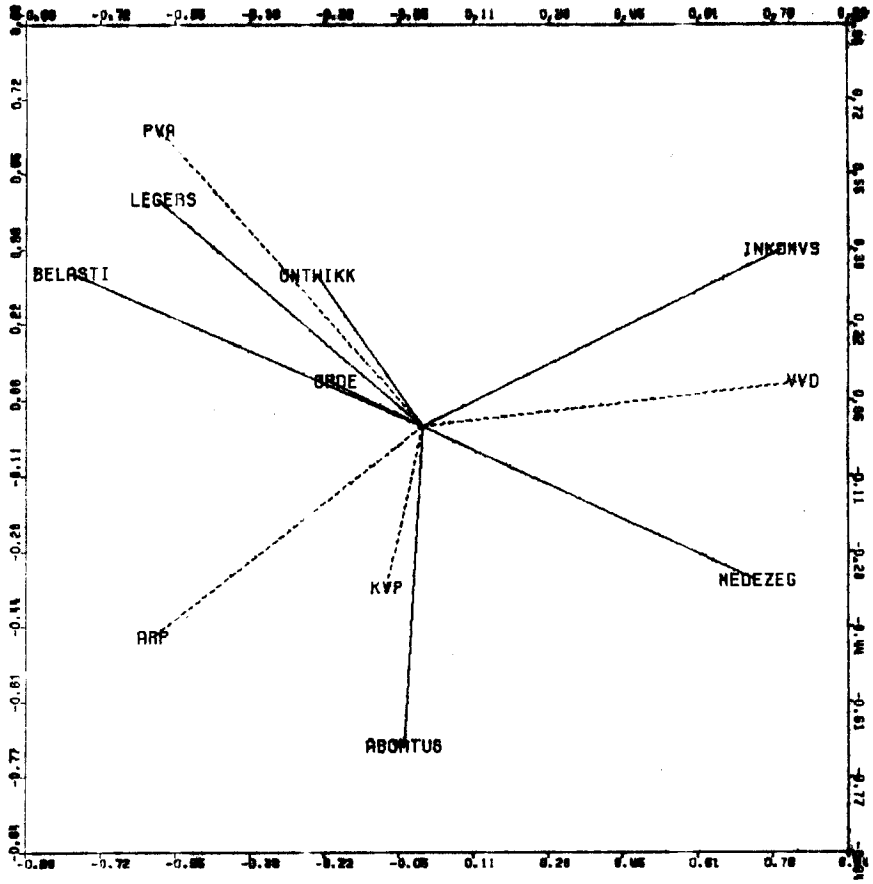
De twee-dimensionale oplossing heeft nu kanonische korrelaties van .92 en .87. Omdat deze korrelaties zo hoog zijn hebben de vektoren in de ene ruimte (zie figuur 7.10) ongeveer dezelfde onderlinge relaties als die in de andere ruimte (zie figuur 7.11).

We zien, dat de variabelen inkomensverschillen, medezeggenschap en belastingen de grootste rol spelen, op de voet gevolgd door "legers" en "abortus". De vektoren van de issues staan in de richting van een hoge skore op de genoemde negenpuntsschaal, zodat we naar de beschrijving van de variabelen moeten kijken om te zien wat ze betekenen. In ons voorbeeld staan ze voor:

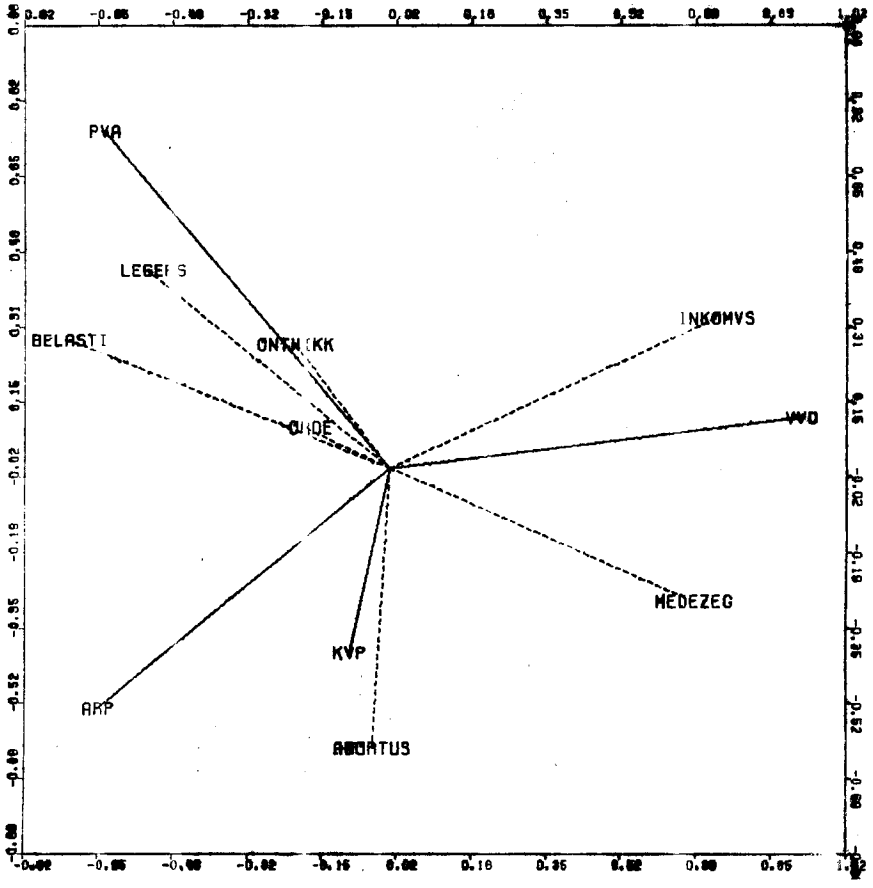
ontwikkelingshulp	- veel minder
(sterke) legers	- handhaven
orde	- de overheid moet harder optreden
belasting	- moet verlaagd worden
abortus	- de vrouw beslist
medezeggenschap	- voor werknemers
inkomensverschillen	- moeten kleiner worden

De vektoren van de partijen wijzen in de richting van "afkeer" (een hoge skore geeft totaal geen affiniteit aan). De verwantschap voor de PvdA, de VVD en de ARP zegt het meest over de samenhang met standpunten over belangrijke issues; de sympatie/antipatie voor de KVP doet dit in mindere mate, maar deze verwantschap zegt dan wel het meeste over standpunten m.b.t.abortus. We moeten ons bij het kijken naar de vektoren natuurlijk realiseren, dat zij doorgetrokken kunnen worden in tegengestelde richting. Hierdoor wordt het wellicht nog evidenter dat b.v. sympatie voor de ARP nauw samenhangt met de opinie dat de inkomensverschillen kleiner moeten worden en dat als je weet dat iemand het standpunt "abortus vrij" inneemt, je moeilijk uit kunt maken of hij of zij veel sympatie voor de PvdA dan wel voor de VVD zal hebben. Uit voorkeur voor de ARP valt geen eenduidig standpunt inzake het issue "legers" af te leiden. De ARP geniet blijkbaar sympatie van 'n groep kamerleden die zich uitspreekt vóór inkrimping en van een andere, die zich sterk maakt voor "handhaving". Deze veronderstelling wordt gestaafd als we kijken naar de relatie tussen de verwantschap met de ARP en die met de PvdA: de ARP zal hoog op het verlanglijstje van kamerleden te "linker" en te "rechter" zijde voorkomen.

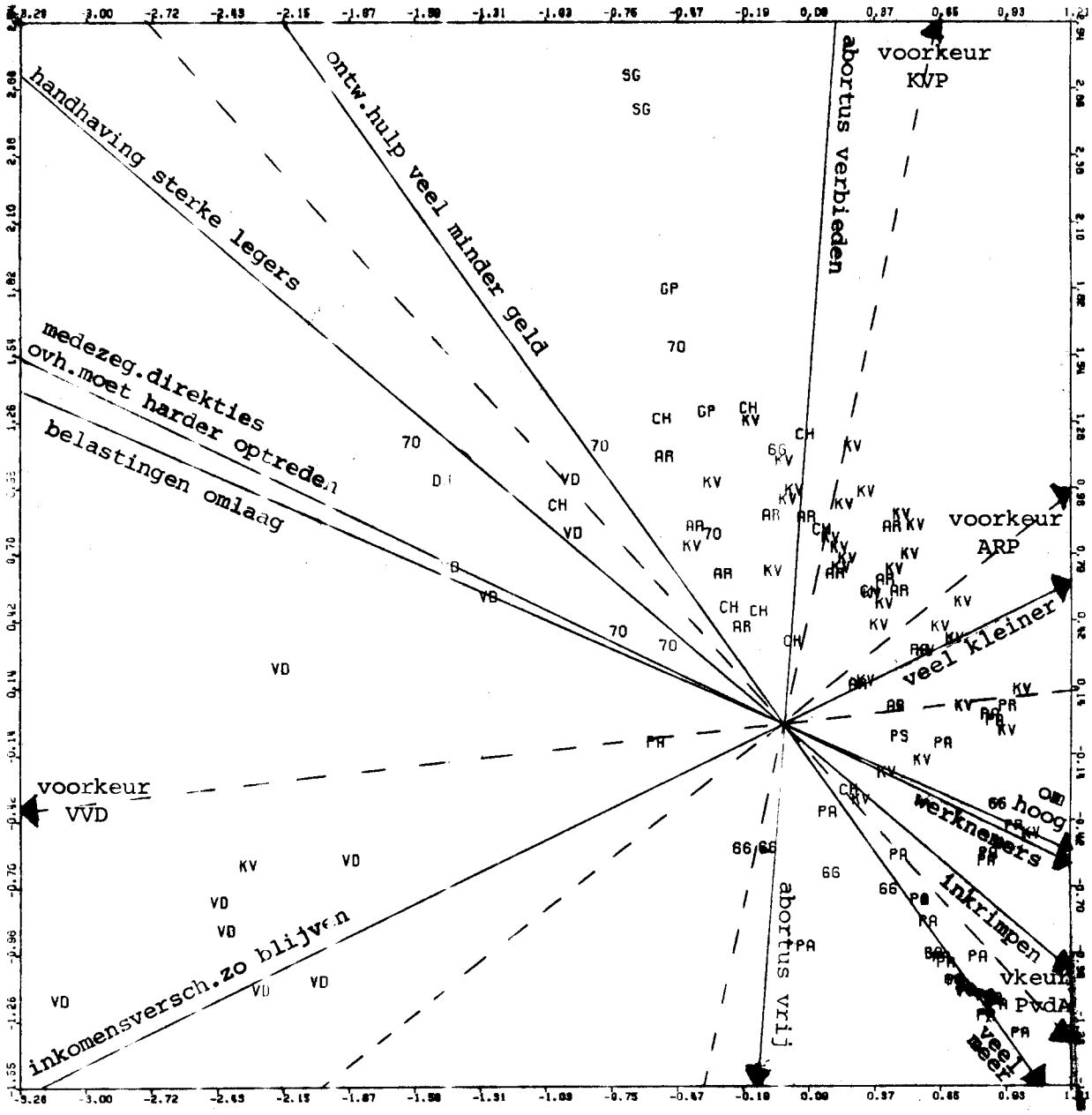
Figuur 7.12 en 7.13 zijn op dezelfde manier gemaakt als beschreven bij het eerste voorbeeld (7.1.4.1). De projekties op de vektoren geven nu de meningen en voorkeuren weer van de individuele kamerleden.



Figuur 7.10. Korrelaties van de variabelen met de kanonische assen van de eerste set (issues)



Figuur 7.11. Korrelaties van de variabelen met de kanonische assen van de tweede set (voorkeuren)

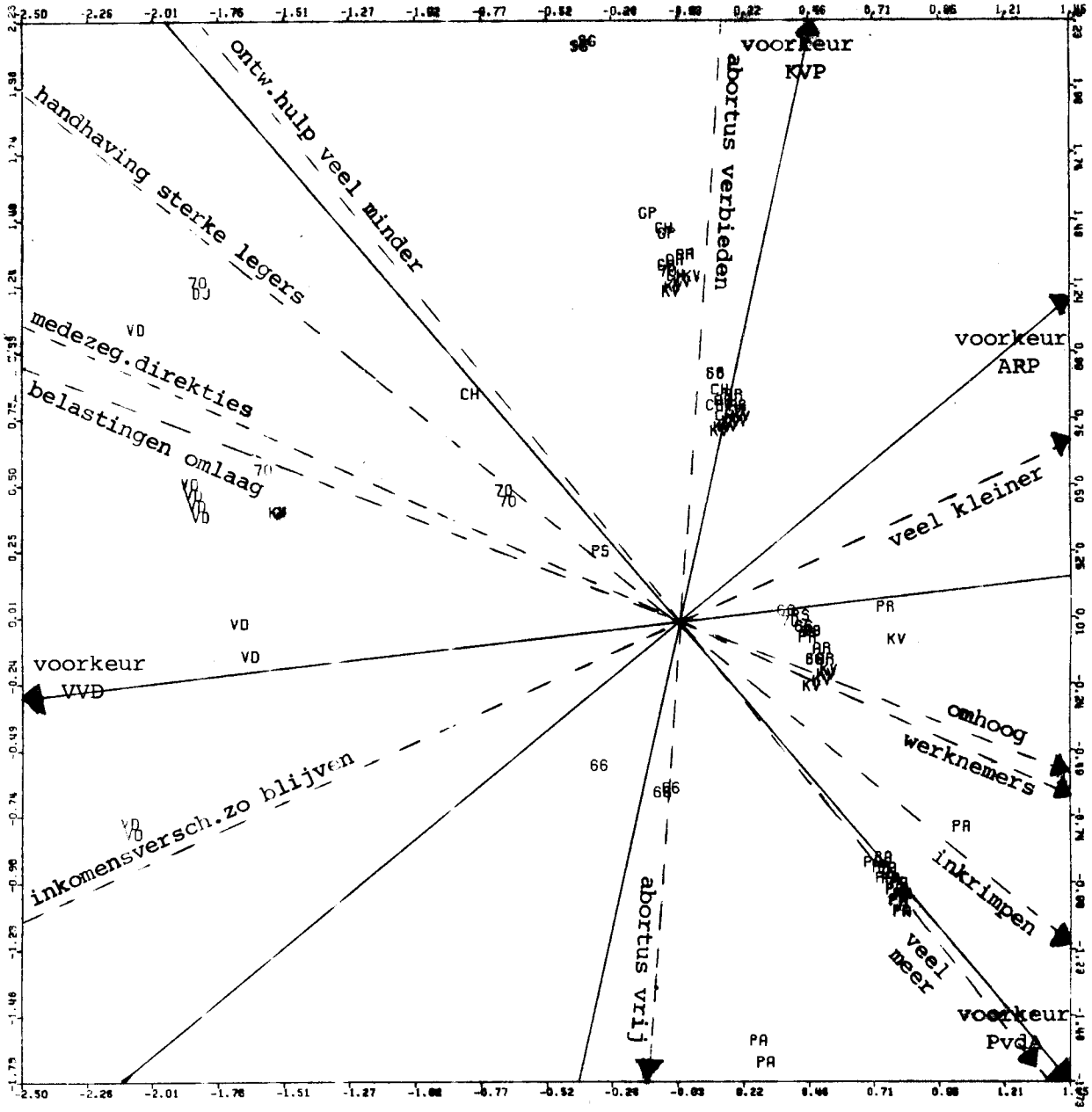


Figuur 7.12. Individu-skores in de kanonische ruimte van de politieke issues

PA - PvdA 66 - D'66 PR - PPR PS - PSP AR - ARP KV -KVP CH - CHU
70 -DS'70 VD - VVD GP - GPV SG - SGP DJ - eenmansfraktie

In figuur 7.12 zijn de vektoren van de issues getekend in de richting van de ons sympatieke stellingnamen. Veel in deze figuur spreekt voor zichzelf; daarom een paar korte opmerkingen.

Het issue inkomensverschillen blijkt in grote lijnen de VVD- en DS'70-kamerleden van de rest te scheiden; abortus kontrasteert



Figuur 7.13. Individu-skores in de kanonische ruimte van de partijvoorkeuren

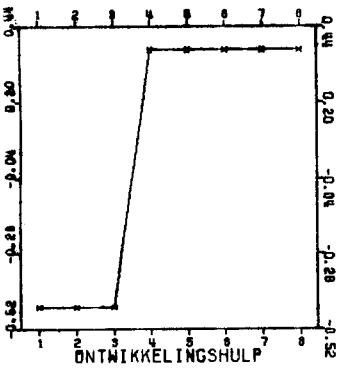
de confessionelen met liberalen en links. Het is opmerkelijk dat de variabele "legers" de confessionelen verdeelt; "belasting" doet dit in nog iets sterkere mate, evanals "medezeggenschap". Opvallend is ook de grote spreiding van de VVD-ers in de andere richting: zij delen blijkbaar niet hetzelfde standpunt over abortus.

Kijken we in de ruimte van de voorkeuren dan blijkt Nederland als drie-stromen-land in beeld gebracht. De klustering is opvallend: de PvdA-ers zitten op één lijn, de confessionelen vormen drie

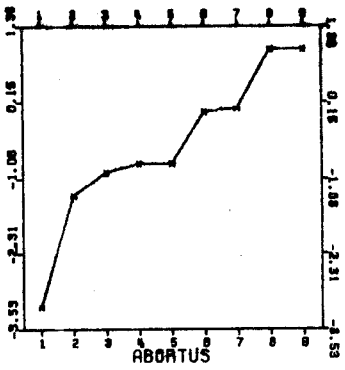
klusters, verdeeld door hun voorkeur voor de PvdA en de VVD en de stellingnamen over abortus, leger, belastingen en medezeggenschap. Wat de inkomensverschillen betreft zitten de confessionelen op één lijn met de PvdA. Voor de confessionelen met een meer gematigd standpunt over abortus geldt, dat zij hun meningen over de issues, die wijzen op affiniteit met de PvdA, ook vertalen in een grotere voorkeur voor deze partij. De VVD is verdeeld in conservatieve en meer liberale kamerleden.

De positie van de PSP'ers kan misschien verbazing wekken. Beiden zijn rond het centrum terecht gekomen. Vooral voor diegene, die links boven het centrum van de plot ligt, geldt dat de grote partijen blijkbaar "lood om oud ijzer" zijn. Voor de volledigheid moet over beide plots met individu-skores worden opgemerkt dat 2 resp. 3 VVD-kamerleden niet zijn opgenomen. Zij namen zo'n extreem standpunt in over inkomensverschillen, belasting en/of abortus, dat zij geheel geïsoleerd links onder in de plot kwamen te liggen, ver van hun mede VVD-ers af. Als we ze in de plot hadden opgenomen, zouden andere kamerleden in het centrum,met kleine onderlinge afstand,door de schaling erg geklonterd zijn en zouden veel labels onleesbaar zijn geworden.

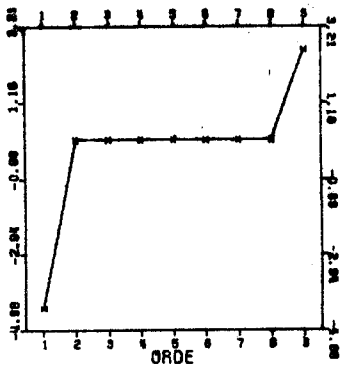
Tot slot van dit voorbeeld een korte blik op de <u>transformaties</u> van de variabelen (zie figuur 7.14). De meningen over ontwikkelingshulp worden niet erg genüanceerd getransformeerd: je bent als het ware óf voor veel <u>meer</u> geld óf voor veel <u>minder</u> geld. We hebben echter al gezien in figuur 7.10. en 7.11. dat ontwikkelingshulp niet zo'n belangrijke rol speelt. Hetzelfde kan opgemerkt worden over de variabele "orde". De variabele medezeggenschap zegt voornamelijk iets over <u>hoeveel</u> medezeggenschap de werknemers zouden moeten hebben;slechts een enkeling in de Kamer vond dat alleen de direktie het voor het zeggen moest hebben. Uit de transformaties van de mate van verwantschap met de ARP kunnen we aflezen dat eigenlijk voornamelijk "afkeer" van de ARP een duidelijk standpunt vertegenwoordigt (zie ook figuur 7.11.) ; een grote voorkeur voor de ARP krijgt dezelfde kwantifikatie als een lichte voorkeur. Ook de transformaties voor inkomensverschillen maken het de moeite waard om naar figuur 7.11. terug te kijken. De transformaties van de vier partijvoorkeuren vertonen ieder een geheel ander beeld. Hetzelfde geldt voor de transformaties van verscheidene issues. Deze verschillen zouden inhoudelijk geïnterpreteerd kunnen worden als wel of geen <u>politiseringseffekten</u>.



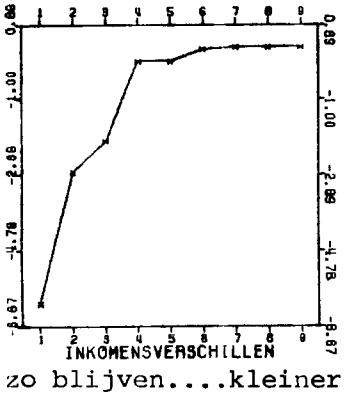
meer.....minder



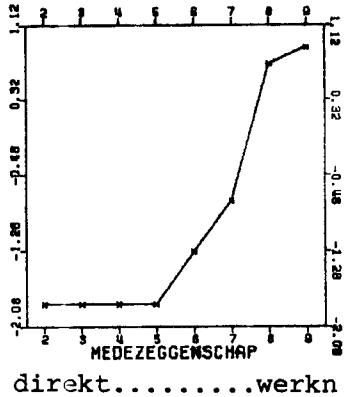
verbieden.....vrij



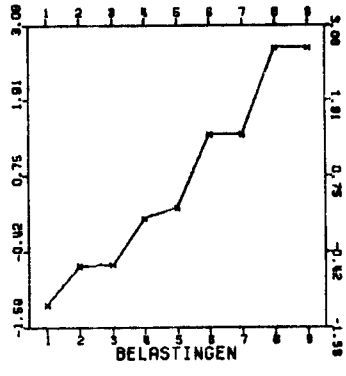
te hard.....harder



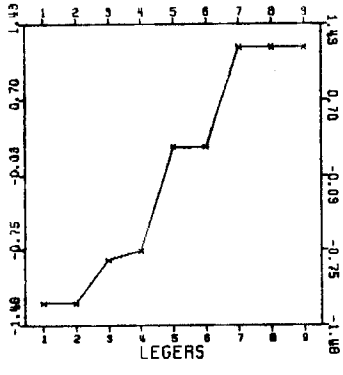
zo blijven....kleiner



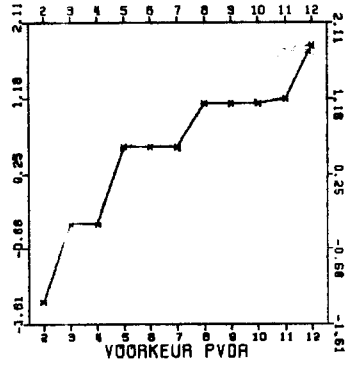
direkt.....werkn



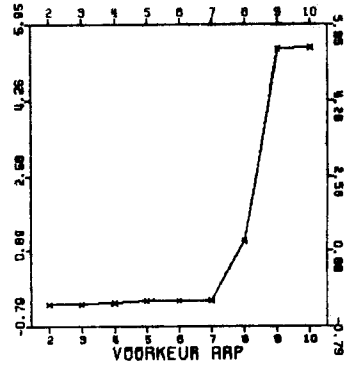
hoger.....lager



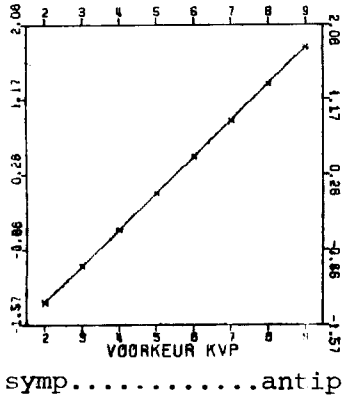
inkrimp.....handh



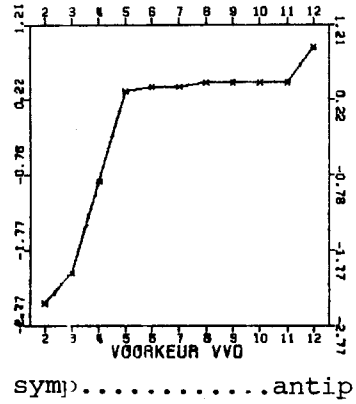
symp.....antip



symp.....antip



symp.....antip



symp.....antip

Figuur 7.14. Oorspronkelijke data (horizontaal) versus getransformeerde scores (vertikaal)

7.2. Niet lineaire multiple regressie

7.2.1. MORALS

In het algemeen maakt men in de multivariate analyse onderscheid tussen multiple regressie en kanonische korrelatie. Het verschil hierbij is het aantal variabelen in één der sets, nl één of meer. De kanonische korrelatiekoëfficient wordt in het multiple regressie geval multiple korrelatiekoëfficient genoemd en wordt meestal in het kwadraat gepresenteerd. Maken we dit onderscheid ook in de niet-lineaire multivariate analyse , dan is de generalisatie van multiple regressie analyse MORALS en van kanonische korrelatie analyse CORALS. Door Young, De Leeuw en Takane (1976) is er een algoritme besproken zowel voor MORALS als CORALS. Zoals eerder vermeld (7.1.2) is hun procedure slechts geschikt voor niet-lineaire multiple regressie analyse en niet voor niet-lineaire kanonische korrelatie analyse, aangezien één der kondities geschonden wordt bij de berekening van de variabelen als beide sets meer dan één variabele bevatten. Het CANALS algoritme is geschikt voor MORALS en CORALS, maar omdat de rekenprocedure voor MORALS er veel eenvoudiger kan uitzien dan in het programma CANALS het geval is ,willen we op den duur een apart programma voor MORALS maken

Het MORALS algoritme ziet er als volgt uit:

Minimaliseer $(SSQ(Q_1 a - q_m))/n \triangleq \sigma(Q,a)$ over $q_1,..q_m$ en a

onder voorwaarde dat $q_j' q_j = n$ en $q_j \varepsilon C_j$ voor $j=1,...,m$

waarbij $Q_1(nxm_1)$, $a(m_1)$, $m_1=m-1$ en $Q=(q_1|q_2|...|q_m)$

Dit probleem is in de volgende deelproblemen te splitsen:

1 minimaliseer $\sigma(Q,a)$ over a onder voorwaarde dat:

$q_j' q_j = n$ en $q_j \varepsilon C_j$ voor $j=1,..,m$

Dit geeft $a=(Q_1' Q_1)^{-1} Q_1' q_m$

Om de berekening van de inverse matriks te vermijden kan a ook benaderd worden:

$a \triangleq a+\theta e_j$, waarbij $e_j=1$ voor element j en $e_j=o$ voor alle andere

elementen, j=1,...,m-1 en

$\theta=q_j'(Q_1a-q_m)/n$

2 voor l=1,...,m-1

2^a minimaliseer $\sigma(Q,a)$ over q_1 onder voorwaarde dat $q_j'q_j=n$ en $q_j \varepsilon C_j$ voor j=1,....,m-1 en $j\neq 1$

Dit geeft $q_1=(q_m-Q_1a)a_1+a_1^2q_1$.

2^b Zoek de kleinste kwadraten benadering van q_1 die tot C_1 behoort (kategoriegemiddelden, monotone regressie op de gemiddelden of lineaire regressie, zoals bij CANALS) en standaardiseer.

3^a minimaliseer $\sigma(Q,a)$ over q_m onder voorwaarde dat $q_j'q_j=n$ en $q_j \varepsilon C_j$ voor j=1,...,m-1.

Dit geeft $q_m=Q_1a$

3^b Zoek de kleinste kwadraten benadering van q_m die tot C_m behoort en standaardiseer.

4 Bereken de multiple korrelatiekoëfficient = de korrelatie tussen Q_1a en q_m.

Als het verschil met de vorige multiple korrelatiekoëfficient te groot is ga dan terug naar 1.

5 Standaardiseer Q_1a op de multiple korrelatie koëfficient (en niet op n zoals bij CANALS gebruikelijk is).

Hierna volgen een aantal voorbeelden van een niet-lineaire multiple regressie analyse.

7.2.2. Toepassingen van MORALS

7.2.2.1 Van Jaar tot Jaar

Een voorbeeld van MORALS vinden we in de sekundaire analyse van het 'Van Jaar Tot Jaar' onderzoek (De leeuw en Stoop,1979). Hierin zijn een aantal achtergrondgegevens mbt de vooropleiding van ouders, hun wensen t.o.v. hun kinderen en de situatie op school vergeleken met het eindnivo van het kind. De lijst van variabelen staat in appendix B.1. Er zijn een aantal geneste niet-lineaire regressie analyses uitgevoerd, zowel met eindnivo als geslacht als de te voorspellen variabele. De kolommen in tabel 7.3 en 7.4 geven de korrelaties en de regressiegewichten van de achtereenvolgende analyses. Eerst zijn de variabelen BVA tm URB gebruikt om eindnivo of geslacht te voorspellen.Daarna BVA tmURB en ASO tm BIM, enz. Er zijn acht regressieanalyses gedaan met een toenemend aantal variabelen. De volgorde van de variabelen is ruw weg door de faktor tijd bepaald.De multiple korrelatiekoëfficient staat onderaan de tabellen. Het grote verschil tussen deze twee analyses is dat de gewichten zich bij eindnivo instabiel gedragen en bij sexe stabiel (gewichten en korrelaties zijn beide positief genomen). Bij de zesde regressie analyse van eindnivo vindt er bij de gewichten een verschuiving plaats van BVA,BIM en DLO naar ADV. Bij de zevende analyse verdwijnt ADV ten gunste van TON en in de achtste analyse is TON verdwenen ten gunste van AOS,LLS en EXT. De korrelatie van TON is overigens hoog gebleven. De korrelaties en de gewichten bij sexe komen heel aardig overeen en gedragen zich ook zeer stabiel. Mbt deze data set kan men zich dus beter een uitspraak over een relatie van deze analysevariabelen met sexe dan met eindnivo veroorloven. Voor sexe is het aantal leerlingen op de school nogal dominerend. Het aantal leerlingen hangt erg samen met het type school. Het is dus niet te verwonderen dat dit sexe redelijk kan voorspellen. Ook het aspiratienivo van de ouders doet wel wat in de regressie. De beroepsinteresse middelbaaronderwijs heeft weer te maken met het type school.
Bij eindnivo zijn het voornamelijk de eerste keuze vervolgonderwijs en de kenmerken van de school die bepalend zijn.Het advies van de onderwijzer korreleert ook nog redelijk met de te voorspellen variabele. We hebben eindnivo ook nog apart voor meisjes en jongens bekeken om te ontdekken of er verschillen zijn tussen meisjes en

EINDNIVO

	Korrelaties								regressiegewichten							
BVA	.468	.455	.450	.450	.436	.411	.216	.195	.285	.236	.228	.229	.222	.122	.107	.069
OPV	.451	.439	.416	.416	.376	.376	.201	.088	.235	.218	.163	.161	.135	.082	.038	.015
OPM	.329	.326	.307	.304	.297	.306	.168	.070	.152	.133	.112	.110	.102	.049	.033	.019
AKG	.177	.178	.175	.175	.174	.165	.014	.064	.127	.123	.099	.100	.087	.072	.038	.013
URB	.065	.056	.009	.018	.001	.054	.004	.001	.021	.028	.027	.031	.028	.020	.009	.009
ASO		.187	.064	.073	.159	.145	.060	.194		.178	.036	.035	.035	.030	.041	.048
ASL		.172	.170	.168	.151	.149	.121	.062		.092	.079	.078	.076	.034	.028	.017
BIL		.083	.088	.092	.091	.091	.098	.226		.053	.052	.056	.056	.061	.043	.063
BIM		.345	.330	.330	.316	.276	.216	.222		.219	.190	.190	.160	.066	.031	.077
INT			.335	.334	.327	.234	.163	.183			.210	.210	.201	.113	.105	.092
OOA			.210	.185	.248	.195	.034	.168			.049	.048	.055	.033	.041	.034
DWO			.178	.105	.212	.123	.023	.109			.067	.038	.115	.042	.031	.076
BMB			.350	.338	.347	.267	.023	.016			.097	.090	.084	.024	.013	.004
INS			.238	.220	.232	.161	.163	.237			.233	.203	.205	.166	.134	.220
KLS				.083	.080	.087	.090	.070				.044	.043	.015	.026	.003
DLO					.431	.398	.395	.272					.323	.132	.071	.021
LL6					.072	.094	.062	.033					.073	.023	.027	.020
ADV						.744	.485	.456						.419	.105	.371
KGS						.366	.207	.145						.086	.033	.026
PRE						.674	.395	.170						.189	.069	.018
TON							.880	.709							.726	.066
AOS								.750								.254
LLS								.806								.291
EXT								.850								.208
MC	.557	.630	.680	.682	.739	.853	.917	.983	.557	.630	.680	.682	.739	.853	.917	.983

Tabel 7.3. Acht multiple regressies van eindnivo op acht geneste groepen variabelen van 'Van Jaar tot Jaar'

GESLACHT

	correlaties									regressiegewichten								
BVA	.143	.134	.134	.134	.133	.133	.131	.114	.114	.141	.129	.130	.130	.132	.123	.118	.115	.126
OPV	.106	.108	.106	.106	.105	.104	.102	.102	.099	.108	.103	.104	.103	.102	.102	.103	.101	.102
OPM	.090	.085	.085	.085	.085	.081	.082	.080	.081	.089	.073	.070	.072	.065	.052	.055	.060	.060
AKG	.061	.055	.052	.052	.050	.050	.051	.054	.052	.061	.061	.062	.062	.067	.071	.066	.057	.058
URB	.042	.038	.031	.028	.037	.041	.041	.042	.042	.038	.042	.031	.031	.037	.042	.041	.054	.058
ASO		.256	.264	.264	.263	.263	.263	.263	.262		.258	.282	.283	.278	.242	.241	.226	.209
ASL		.056	.061	.060	.061	.060	.061	.062	.061		.085	.087	.089	.091	.084	.075	.087	.080
BIL		.218	.222	.222	.223	.223	.219	.200	.199		.174	.168	.169	.169	.161	.160	.145	.143
BIM		.211	.212	.212	.211	.212	.213	.210	.209		.160	.155	.157	.150	.148	.152	.129	.120
INT			.192	.193	.192	.190	.191	.183	.182			.193	.194	.191	.184	.182	.170	.168
OOA			.135	.134	.100	.123	.067	.080	.150			.083	.083	.063	.046	.056	.049	.054
DWO			.190	.190	.190	.186	.190	.189	.184			.275	.279	.272	.244	.284	.306	.282
BMB			.220	.220	.220	.215	.220	.216	.213			.242	.242	.226	.196	.252	.262	.231
INS			.161	.163	.151	.134	.150	.086	.072			.110	.112	.106	.088	.105	.065	.066
KLS				.008	.008	.008	.008	.008	.008				.021	.020	.031	.030	.043	.042
DLO					.088	.089	.089	.088	.088					.069	.079	.073	.084	.086
LL6					.075	.072	.074	.072	.070					.069	.046	.047	.039	.037
ADV						.144	.140	.114	.111						.131	.156	.109	.114
KGS						.135	.134	.134	.134						.138	.126	.118	.112
PRE						.035	.031	.010	.001						.052	.026	.008	.012
TON							.099	.056	.055							.137	.054	.042
AOS								.148	.151								.153	.185
LLS								.376	.375								.331	.326
EXT								.200	.202								.096	.086
EIN									.176									.189
m.c.	.213	.426	.483	.484	.492	.521	.532	.634	.650	.213	.426	.483	.484	.492	.521	.532	.634	.650

Tabel 7.4. Negen multiple regressies van geslacht op negen geneste groepen variabelen van 'Van Jaar tot Jaar'.

jongens (zie tabel 7.5). Voor meisjes zijn de variabelen BVA,OPV,INT,OOA,DLO en KGS een stuk belangrijker dan voor jongens. Voor jongens is het vooral AOS die een grote rol speelt mbt eindnivo. Voor meisjes betekent dit, dat ze meer afhankelijk zijn van de opleiding van de vader, de interesse van de ouders en van hun eigen prestaies dan jongens. Bij jongens is het eindnivo beter door het aantal andere opleidingen verbonden aan hun school (wat samenhangt met het type school) te voorspellen. Overigens zijn ook hier de variabelen TON,AOS, LLS en EXT weer de beste voorspellers van het eindnivo.

		Korrelaties		Gewichten	
1	BVA	.14	.02	.08	.08
2	OPV	.34	.24	.10	.03
3	OPM	.20	.26	.04	.04
4	AKG	.16	.10	.05	.05
5	URB	.01	.00	.03	.01
6	INT	.26	.08	.10	.07
7	ASO	.17	.06	.03	.04
8	OOA	.14	.02	.03	.02
9	DWO	.13	.21	.05	.06
10	BMB	.09	.10	.00	.01
11	KLS	.10	.09	.00	.01
12	DLO	.46	.37	.08	.04
13	LL6	.08	.02	.02	.03
14	ADV	.65	.62	.19	.25
15	KGS	.30	.13	.05	.05
16	BIL	.14	.17	.07	.05
17	BIM	.21	.28	.03	.08
18	PRE	.48	.40	.09	.09
19	TON	.89	.89	.38	.28
20	AOS	.57	.77	.02	.15
21	LLS	.68	.74	.25	.20
22	EXT	.77	.82	.10	.14
23	ASL	.09	.12	.05	.04
24	INS	.17	.25	.11	.16
M.C.		.96	.96		
		M	J	M	J

Tabel 7.5 Korrelaties en gewichten voor meisjes(M) en jongens(J) van de multiple regressies van eindnivo op de variabelen van 'Van Jaar tot Jaar' en mutiple korrelatiekoëfficienten(M.C.)

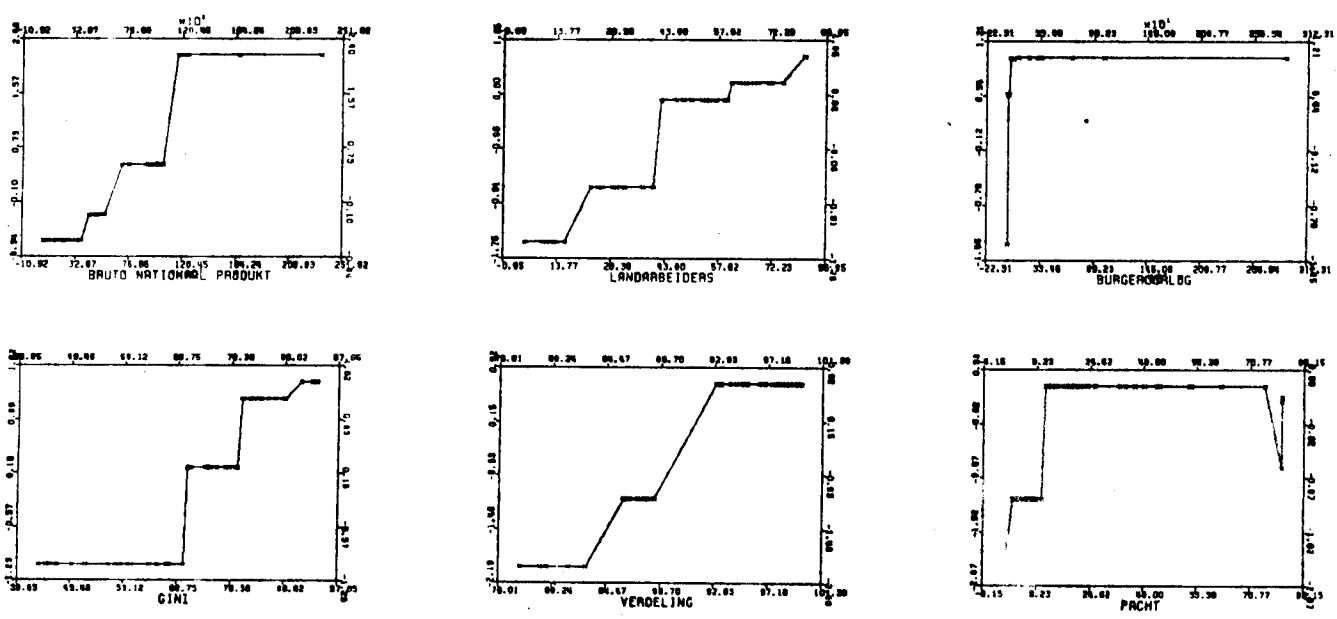
7.2.2.2 Burgeroorlog,revolutie of rellen

De ekonomische en politieke gegevens van Russett (1969) hebben we gebruikt voor een niet-lineaire multiple regressie analyse. Een beschrijving van de data set vinden we bij de toepassingen van CANALS (7.1.4.2). De ekonomische variabelen GINI,VERD,PACH,BRNP en LARB vormen de ene set en het aantal doden door burgeroorlog,revolutie of rellen,BURG, de andere set. De multiple korrelatie koëfficient is .858. De korrelaties en de gewichten van de ekonomische variabelen met de variabele BURG staan in tabel 7.6. Kijken we naar de korrelaties dan geldt: hoe hoger de GINI-indeks is (dwz hoe ongelijker het land verdeeld is) , hoe meer kleine boeren er zijn (VERD),hoe groter het aantal gepachte boerderijen is (PACH), hoe lager het bruto nationaal produkt (BRNP) en hoe hoger het percentage arbeiders dat in de

	gew.	korr.
GINI	.529	.580
VERD	.387	-.490
PACH	.410	.473
BRNP	-.667	-.589
LARB	.596	.261

Tabel 7.6 Regressie gewichten en korrelaties mbt BURG.

landbouw werkt (LARB).Dit gaat samen met een groter aantal doden ten gevolge van burgeroorlog,revolutie of rellen.In de gewichten zijn GINI en VERD tegengesteld aan elkaar. GINI en VERD zijn beide indeksen voor de verdeling van het land. Ze hebben een onderlinge korrelatie van .90 (berekend over de oorspronke-ijke data). Russett gaat ervan uit dat GINI een beter indeks is voor landverdeling dan VERD (de gebruikelijke indeks) en daarom ook een beter voorspeller is voor de politieke instabiliteit. Dit weerspiegelt zich in de hogere korrelaties van GINI met de politieke variabelen dan de korrelaties van VERD. Het geldt ook voor ons geval, ondanks het feit dat we de variabelen gekategoriseerd hebben. Omdat de gewichten van GINI en VERD tegengesteld zijn, heffen ze elkaars effekt voor een groot deel op. Dit kan alleen omdat GINI en VERD zo'n hoge onderlinge korrelatie hebben (multikolleariteit). Dit zelfde geldt ook enigszins voor bruto nationaal produkt en percentage landarbeiders. De korrelatie tussen BRNP en LARB is -.80 (berekend over de ongekategoriseerde variabelen). LARB komt er slecht van af met een gewicht van .261, omdat de meeste variantie al door BRNP verklaard is. GINI, BRNP en LARB zijn de variabelen die het beste het aantal doden door burgeroorlog kunnen voorspellen. De herschalingen van de variabelen staan in figuur 7.15. Bij 'pacht' zien we een daling, terwijl we



Figuur 7.15. Herschalingen van de variabelen van Russett. Horizontaal de oorspronkelijke skores en vertikaal de herschalingen berekend mbv niet-lineaire multiple regressie analyse.

alleen monotoon stijgende transformaties verwachten. Dit komt omdat de laatste drie landen in de figuur een missing skore hebben. De variabele BURG wordt zodanig getransformeerd, dat we bijna van een tegenstelling wel doden geen doden tgv burgeroorlog kunnen spreken. Het zijn voornamelijk de westerse landen waarin geen doden voorkomen, nl Australië,Canada,Finland,Groot Brittannië,Ierland,Joegoslavië,Luxemburg,Lybië,Nederland,Noorwegen,Nieuw Zeeland,Oostenrijk, Taiwan,West Duitsland,Zweden ,Verenigde Staten en Zwitserland.Deze landen hebben volgens tabel 7.1. in hfdst 7.1.4.1 geen doden tgv burgeroorlog,revolutie of rellen in de jaren 1950-1962. De landen India,Filippijnen,Panama,Griekenland,Honduras,Nicaragua,Spanje,Cuba, Dominicaanse Republiek,Brazilië,Columbia,Guatemala,Argentinië,Ecuador,Peru,Irak,Costa Rica,Venezuele,Bolivia en Zuid Vietnam hebben meer dan twee doden per 1.000.000 inwoners volgens tabel 7.1. Dit is duidelijk een tegenstelling rijk-arm en ook de tegenstelling ontwikkeld-minder ontwikkeld.

7.3. Niet-lineaire kanonische analyse met meer dan twee sets

In plaats van twee sets van variabelen kunnen we ook te maken hebben met meerder sets (Kettenring,1971). De generalisatie van het CANALS model naar M sets noemen we OVERALS. Het probleem dat we in OVERALS oplossen is:zoek een ortogonale basis die alle verzamelingen van variabelen zo veel mogelijk gemeenschappelijk hebben. Met andere woorden, zoek een ruimte X(nxp) en gewichtsmatriksen $A_1(m_1xp)$.... $A_M(m_Mxp)$ zodanig dat de lineaire deelruimten $Q_1A_1,\ldots,Q_MA_M$ zo veel mogelijk op X lijken in kleinste kwadraten zin, waarbij de matriks X kolomortogonaal moet zijn, dwz $X'X=nI$.

Ook hier kunnen we verder gaan dan de klassieke multivariate analyse door, zoals bij CANALS, een herschaling van de variabelen toe te staan die in overeenstemming is met het meetnivo van iedere variabele. Als we de ruimte gedefinieerd door de restrikties die het type variabele met zich meebrengt C noemen, kunnen we het OVERALS algoritme als volgt formuleren:

minimaliseer $\Sigma_{K=1}^{M} SSQ(X-Q_KA_K) \triangleq \sigma(Q,A,X)$ over Q, A en X

onder voorwaarde dat:

$X'X=nI$, $q_j'q_j=n$ en $q_j \varepsilon C_j$ voor $j=1,\ldots,m$

met $Q \triangleq (q_1, \ldots, q_M)$ en $A' \triangleq (A_1' | \ldots | A_M')$.

Ook het OVERALS probleem kan gesplitst worden in verschillende deelproblemen, maar dat zullen we hier niet verder bespreken, aangezien het OVERALS model voorlopig niet in de praktijk toegepast kan worden omdat het komputerprogramma er nog niet is. Wat we wel willen laten zien is dat het OVERALS algoritme voor twee sets van variabelen ekwivalent is met CANALS algoritme. De oplossingen van A_1 en A_2 zijn identiek op een ortogonale en/of diagonale transformatie na en de te minimaliseren funkties zijn identiek. Hiervoor bewijzen we eerst dat minimalisatie van de OVERALS verliesfunktie over X en A onder voorwaarde dat $X'X=nI$, overeenkomt met minimalisatie van dezelfde verliesfunktie onder voorwaarde dat de som van de kruisproduktenvan de lineaire kombinaties $Q_1A_1, \ldots, Q_MA_M$ kolomortogonaal is, dwz

$\Sigma_{L=1}^{M} \Sigma_{K=1}^{M} A_K' Q_K' Q_L A_L = nI$.

In het eerste geval minimaliseren we:

$F = tr(\Sigma_K (X - Q_K A_K)' (X - Q_K A_K) - (X'X - nI) L_1)$,

waarbij L_1 een symmetrische onbepaalde Lagrange vermenigvuldiger is. Differentiatie van F naar X, A_K en L_1 en gelijk stellen aan nul levert:

1) $\partial F/\partial X = 0 \rightarrow X = \Sigma_K Q_K A_K (MI - L_1)^{-1}$

2) $\partial F/\partial A_K = 0 \rightarrow A_K = (Q_K'Q_K)^{-1} Q_K'X$

3) $\partial F/\partial L_1 = 0 \rightarrow X'X = nI$

Substitutie van 2) in 1) geeft:

$X(MI - L_1) = \Sigma_K Q_K (Q_K'Q_K)^{-1} Q_K'X$,

definieer $P_K \triangleq Q_K (Q_K'Q_K)^{-1} Q_K'$ en $P \triangleq \Sigma_K P_K$, dan geldt:

4) $X(MI - L_1) = PX$

Stel $(MI - L_1) \triangleq R_1 \Delta R_1'$ met R_1 (pxp) ortogonaal (EVD, zie appendix A.1)

Kombinatie van 3) en 4) levert:

$XR_1/\sqrt{n} = U_1$; U_1 bevat p eigenvektoren van matriks P.

Dus $X = U_1 R_1' \sqrt{n}$ en $A_K = (Q_K'Q_K)^{-1} Q_K' U_1 R_1' \sqrt{n}$.

De verliesfunktie berekenen we door X en A_k te substitueren:

$\sigma = n(Mp - \Sigma_{s=1}^{p} \delta_s)$. Dit is minimaal als Δ de p grootste eigenwaarden van P bevat.

In het tweede geval minimaliseren we:

$F=tr(\Sigma_K(X-Q_KA_K)'(X-Q_KA_K)-(\Sigma_L\Sigma_KA_K'Q_K'Q_LA_L-nI)L_2)$,

waarbij L_2 een symmetrische onbepaalde Lagrange vermenigvuldiger is.

1) $\partial F/\partial X=0 \rightarrow X=(1/M)\Sigma_KQ_KA_K$

2) $\partial F/\partial A_K=0 \rightarrow A_K=(Q_K'Q_K)^{-1}Q_K(X+\Sigma_KQ_KA_KL_2)$

3) $\partial F/\partial L_2=0 \rightarrow \Sigma_K\Sigma_LA_K'Q_K'Q_LA_L=nI$

Substitueer $Q=(Q_1,..,Q_M)$ en $A'=(A_1',...,A_M')$

1) $X=(1/M)QA$

2) $A_K=(Q_K'Q_K)^{-1}Q_K'(X+QAL_2) \rightarrow \Sigma_KQ_KA_K=QA=P(X+QAL_2)$

3) $A'Q'QA=nI$

Substitutie van 1) in 2) geeft:

4) $QA=PQA(I/M+L_2)$

Stel $(I/M+L_2)^{-1}\overset{\Delta}{=}R_2\Delta R_2'$ met R_2(pxp) ortogonaal (EVD)

Kombinatie van 3) en 4) levert:

$QAR_2/\sqrt{n}=U_1$, waarbij U_1 p eigenvektoren van matriks P bevat.

Dus $X=(\sqrt{n}/M)U_1R_2'$ en $A_K=\sqrt{n}(Q_K'Q_K)^{-1}Q_KU_1\Delta^{-1}R_2'$.

De verliesfunktie berekenen we door X en A_K te substitueren.

$\sigma=n(\Sigma_{s=1}^{p}(1/\delta_s)-p/M)$. Dit is minimaal als Δ de p grootste eigenwaarden van matriks P bevat.

We zien dus dat de beide problemen de eerst p eigenwaarden van matriks P maksimaliseren en dat de oplossingen voor X en A_K inderdaad ekwivalent zijn. De waarde van het minimum van de beide verliesfunkties is echter verschillend.

We willen nu bewijzen dat het tweede OVERALS probleem identiek is aan het CANALS probleem. Dwz dat de te minimaliseren funkties op een konstante na identiek zijn en dat de oplossingen voor A_1 en A_2 van OVERALS en CANALS ekwivalent zijn. Substitutie van X=QA/2 in de verliesfunktie van OVERALS geeft:

$\Sigma_{K=1}^{2}SSQ(X-Q_KA_K)=(1/2)SSQ(Q_1A_1-Q_2A_2)$.

De verliesfunktie zijn dus op een konstante na identiek, maar de

kondities verschillen. De oplossing van A_K wordt in CANALS uitgedrukt in singuliere waarden en eigenvektoren van de matriks T,

$$T=(Q_1'Q_1)^{-\frac{1}{2}}Q_1'Q_2(Q_2'Q_2)^{-\frac{1}{2}}$$

De OVERALS oplossing van A_k wordt uitgedrukt in de eigenwaarden en eigenvektoren van de matriks P,

$$P=Q_1(Q_1'Q_1)^{-1}Q_1'+Q_2(Q_2'Q_2)^{-1}Q_2'$$

We kunnen echter ook de OVERALS oplossing in T uitdrukken. De singuliere waarden dekompositie van T is:

$T\triangleq Z\Lambda W'$

$Z(m_1 x m_1)$ en $W(m_2 x m_2)$ zijn ortogonaal en voor $\Lambda(m_1 x m_2)$ geldt $\lambda_{ij}=o$ voor $i\neq j$.

De konditie $\Sigma_{k=1}^{M}\Sigma_{l=1}^{M}A_K'Q_K'Q_LA_L=nI$ is te schrijven als:

$$(A_1'(Q_1'Q_1)^{\frac{1}{2}}|A_2'(Q_2'Q_2)^{\frac{1}{2}})\begin{bmatrix} I & Z\Lambda W' \\ W\Lambda Z & I \end{bmatrix}\begin{bmatrix}(Q_1Q_1)^{\frac{1}{2}}A_1 \\ (Q_2Q_2)^{\frac{1}{2}}A_2\end{bmatrix} = nI$$

Bovendien is $\begin{bmatrix} I & Z\Lambda W' \\ W\Lambda Z' & I \end{bmatrix}$ te splitsen in: $\frac{1}{2}\begin{bmatrix} Z & Z \\ W & -W\end{bmatrix}\begin{bmatrix} I+\Lambda & 0 \\ 0 & I-\Lambda\end{bmatrix}\begin{bmatrix} Z' & W' \\ Z' & -W'\end{bmatrix}$

dwz dat $\begin{bmatrix}(1/\sqrt{2n})(I+\Lambda)^{\frac{1}{2}}(Z'(Q_1'Q_1)^{\frac{1}{2}}A_1+W'(Q_2'Q_2)^{\frac{1}{2}}A_2) \\ (1/\sqrt{2n})(I-\Lambda)^{\frac{1}{2}}(W'(Q_1'Q_1)^{\frac{1}{2}}A_1-W'(Q_2'Q_2)^{\frac{1}{2}}A_2)\end{bmatrix} \triangleq \begin{bmatrix}C_1\\C_2\end{bmatrix}$ kolomortogonaal is.

Dit levert:

$$A_1=\sqrt{n/2}(Q_1'Q_1)^{-\frac{1}{2}}Z((I+\Lambda)^{-\frac{1}{2}}C_1+(I-\Lambda)^{-\frac{1}{2}}C_2)$$

$$A_2=\sqrt{n/2}(Q_1'Q_1)^{-\frac{1}{2}}W((I+\Lambda)^{-\frac{1}{2}}C_1+(I-\Lambda)^{-\frac{1}{2}}C_2)$$

Minimalisatie van $\sigma(Q,A)$ onder voorwaarde dat $\Sigma_K\Sigma_L A_L'Q_L'Q_KA_K=nI$ impliceert maksimalisatie van $tr(A_1'Q_1'Q_2A_2)\triangleq\sigma_1$.

Substitutie van A_1 en A_2 levert:

$$\sigma_1= tr(n/2(C_1(I+\Lambda)^{-\frac{1}{2}}\Lambda(I+\Lambda)^{-\frac{1}{2}} -C_2(I-\Lambda)^{-\frac{1}{2}}\Lambda(I-\Lambda)^{-\frac{1}{2}}C_2))$$

Dit is minimaal als alle elementen van C_2 gelijk aan nul zijn en C_1 gelijk is aan een ortogonale (pxp) matriks R, aangevuld met nullen en Λ de p grootste eigenwaarden van T bevat.

Hieruit volgt dat de oplossingen voor A_1 en A_2 te schrijven zijn als:

$$A_1=(\sqrt{n/2})(Q_1'Q_1)^{-\frac{1}{2}}Z_p(I+\Lambda_p)^{-\frac{1}{2}}R \quad\text{en}\quad A_2=(\sqrt{n/2})(Q_2'Q_2)^{-\frac{1}{2}}W_p(I+\Lambda_p)^{-\frac{1}{2}}R$$

Z_p en W_p bevatten de eigenvektoren van T behorende bij de p grootste singuliere waarden van T en Λ_p bevat deze singuliere waarden. De bovenstaande oplossingen voor A_1 en A_2 zijn ekwivalent met de oplossingen van A_1 en A_2, die uit CANALS komen (zie 7.1.2.).

7.4.1. Relatie OVERALS/CANALS met HOMALS

In hoofdstuk 2 is HOMALS geformuleerd als:

minimum $\Sigma_{j=1}^{M} SSQ(X-G_jY_j)$ over X en Y met $X'X=nI$.

HOMALS kan opgevat worden als een bijzonder geval van OVERALS, nl voor het geval dat elke set slechts één variabele bevat die meervoudig nominaal geïnterpreteerd wordt. Dwz dat de sets bestaan uit indikatormatriksen G_j. Het enige verschil tussen OVERALS en HOMALS is de normalisering van G_j. In OVERALS worden de G_j genormaliseerd en in HOMALS niet.

Andersom kan OVERALS/CANALS opgevat worden als een bijzonder geval van HOMALS. We kunnen namelijk de variabelen van één set herschrijven als één 'super'-variabele met evenveel kategorieën als het produkt van de aantallen kategorieën van de variabelen in deze set. Een nominale CANALS analyse van de sets komt overeen met een HOMALS analyse van twee 'super'-variabelen met additieve restrikties op de meerdimensionale kruistabellen van de sets (De Leeuw, 1973,par.5.9,blz.88). Bevat één der sets bv de matriks G_{1+2}, die bestaat uit twee indikatormatriksen G_1 en G_2 en zijn de kategorieskores gelijk aan $\phi=(\phi_1,\phi_2)$ en $\xi=(\xi_1,\xi_2,\xi_3)$ dan geldt:

$$\left[\begin{array}{cc|ccc} 1&0&0&1&0\\ 0&1&1&0&0\\ 0&1&0&0&1\\ 1&0&0&0&1\\ 1&0&1&0&0\\ 1&0&0&0&1\\ 0&1&1&0&0\\ 0&1&0&1&0\\ 1&0&0&0&1\\ 1&0&0&1&0 \end{array}\right] \quad \begin{bmatrix}\phi_1\\ \phi_2\end{bmatrix} \quad \begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\end{bmatrix} = \begin{bmatrix} 0&1&0&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&0&1\\ 0&0&1&0&0&0\\ 1&0&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&1&0&0&0\\ 0&1&0&0&0&0 \end{bmatrix} \begin{bmatrix}\phi_1+\xi_1\\ \phi_1+\xi_2\\ \phi_1+\xi_3\\ \phi_2+\xi_1\\ \phi_2+\xi_2\\ \phi_2+\xi_3\end{bmatrix}$$

$G_1 \cup G_2 = G_{1+2}$ $\qquad$ G_{1x2} 'super'-variabele $\qquad$ ↑ kategorieskores met additieve restrikties

Een illustratie van de verschillende manieren, waarop we de sets kunnen behandelen , zien we in het volgende voorbeeld

7.4.2 Voorbeeld CBS data

In 1977 startte het CBS met het onderzoek "Schoolloopbaan en herkomst van leerlingen bij het voortgezet onderwijs". Het onderzoek is longitudinaal, de leerlingen worden een aantal jaren gevolgd. De eerste resultaten zijn inmiddels geoubliceerd (CBS, 1979). We hebben uit deze publikatie tabel 2 gebruikt. De onderzoekssteekproef bestaat uit ongeveer 37000 leerlingen, in de tabellen zijn de getallen opgehoogd (naar voorkomen van de schoolsoort) tot landelijke totalen. Dit geeft een opgehoogde "steekproef" van ruim 120000. In tabel 2 wordt het verband weergegeven tussen beroep van de vader (BVA), prestatieskore (PRE), geslacht (SEX), en schoolkeuze na het LO (TON).

BVA heeft zeven kategorieen.

1: hogere employes, incl. vrije beroepen.
2: middelbare employes.
3: lagere employes.
4: zelfstandigen (waaronder landbouwers) met personeel.
5: zelfstandigen (waaronder landbouwers) zonder personeel.
6: arbeiders.
7: overig en onbekend.

PRE is een speciaal door het CITO ontworpen toets, met standard-five indeling.

TON heeft vier kategorieen.

1: vwo, havo.
2: mavo.
3: lts, lhno.
4: leao, lmo.

We hebben in onze analyses BVA, PRE, en SEX opgevat als voorspellers van TON. In totaal hebben we acht verschillende analyses gedaan, de analyses worden onderscheiden naar drie binaire kriteria.

1: de onafhankelijke variabelen kunnen lineair of niet-lineair getransformeerd worden.
2: de onafhankelijke variabelen kunnen additief of niet additief gekombineerd worden.
3: de afhankelijke variabele kan lineair of niet-lineair getransformeerd worden.

Model 1: $TON = \alpha_1 BVA + \alpha_2 PRE + \alpha_3 SEX$. $R^2 = .3082$.

Model 2: $\psi(TON) = \alpha_1 BVA + \alpha_2 PRE + \alpha_3 SEX$. $R^2 = .3475$.

Model 3: $TON = \alpha_1 BVA + \alpha_2 PRE + \alpha_3 SEX + \alpha_4 (BVA \times PRE) + \alpha_5 (BVA \times SEX) + \alpha_6 (PRE \times SEX) + \alpha_7 (BVA \times PRE \times SEX)$. $R^2 = .4463$.

Model 4: $\psi(TON) = \alpha_1 BVA + \alpha_2 PRE + \alpha_3 SEX + \alpha_4 (BVA \times PRE) + \alpha_5 (BVA \times SEX) + \alpha_6 (PRE \times SEX) + \alpha_7 (BVA \times PRE \times SEX)$. $R^2 = .4933$.

Model 5: $TON = \phi_1(BVA) + \phi_2(PRE) + \phi_3(SEX)$. $R^2 = .4411$.

Model 6: $\psi(TON) = \phi_1(BVA) + \phi_2(PRE) + \phi_3(SEX)$. $R^2 = .4978$.

Model 7: TON = Φ(BVA,PRE,SEX). R^2 = .4566.

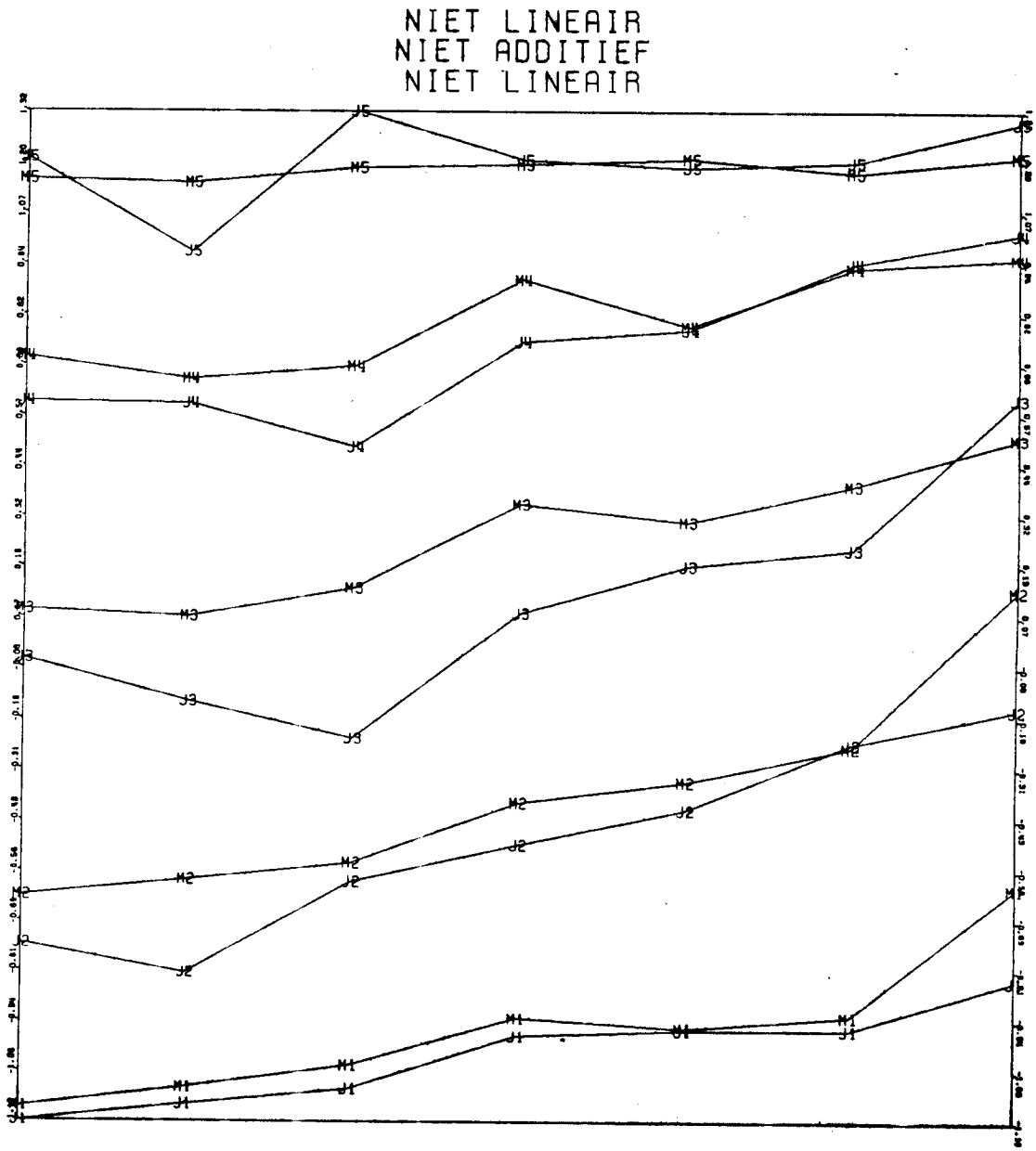
Model 8: ψ(TON) = Φ(BVA,PRE,SEX). R^2 = .5035.

We kunnen model 1 tot en met 6 fitten met CANALS met diverse opties. Het is daarvoor wel nodig dat lineaire interaktie variabelen zoals BVA PRE eerst apart uitgerekend worden. Model 6 is de meest voor de hand liggende toepassing van CANALS, alle variabelen zijn nominaal. Het aantal vrij te kiezen parameters is 13. Model 7 is CANALS met twee variabelen (identiek met HOMALS met twee variabelen, en dus met ANACOR). De eerste variabele is TON, wordt lineair getransformeerd. De tweede variabele heeft als kategorieen alle mogelijke kombinaties van BVA, PRE, en SEX, dus de tweede variabele is nominaal met 7 × 5 × 2 = 70 kategorieen. Het totale aantal vrije parameters is 68. Voor model 8 komen er daar nog twee bij, omdat we TON niet-lineair mogen transformeren. Het is natuurlijk onmogelijk om te zeggen wat het beste van de acht modellen is. Onze indruk is echter dat het zeer grote aantal ekstra vrije parameters in model 7 en model 8 nauwelijks iets oplevert. Niet-lineaire transformatie van TON levert in het algemeen 5% ekstra "verklaarde" variantie op, wat aardig veel lijkt voor twee ekstra parameters. Niet-lineaire transformatie van BVA, PRE, en SEX levert ongeveer 14% ekstra variantie, en kost acht parameters. Maar natuurlijk zijn de gevonden transformaties ook op zichzelf interessant, zelfs al is het duidelijk dat ze in termen van "verklaarde" variantie weinig van lineairiteit en additiviteit afwijken.

Bij de analyse van model 8 vinden we de volgende transformatie van TON.

1: vwo/havo +1.4901
2: mavo +0.2524
3: lts/lhno -1.3076
4: leao/lmo -0.8083

Het is duidelijk dat dit nogal afwijkt van lineariteit, dit verklaart de winst van 5% variantie. De 70 kategorieen van de samensgestelde BVA,PRE,SEX variabele staan weergegeven in figuur 7.16. In deze figuur staan 10 verschillende lijnen, vijf voor jongens met verschillende PRE, en vijf voor meisjes met verschillende PRE. De skores staan steeds uitgezet tegen BVA, iedere lijn bestaat dus uit zeven punten. Uit het feit dat de lijnen ongeveer gelijk verlopen volgt dat de afwijkingen van additiviteit gering zijn, het is duidelijk dat voor ieder PRE nivo de jongens en meisjeslijn weinig verschillen. De lijnen lopen niet erg sterk op, dus BVA heeft niet veel invloed (onafhaneklijk van PRE en SEX). En de vijf paren lijnen verschillen zeer behoorlijk, dus PRE heeft een zeer grote invloed op TON (wat niet zo verwonderlijk is). We gaan verder niet op de details van de figuur in.



Figuur 7.16. Volledig niet-lineaire analyse CBS data

7.5.Niet lineaire diskriminant analyse

7.5.1 CRIMINALS

We kunnen de klassieke kanonische diskriminant analyse uitbreiden met een schaling van de kategorieskores, dan hebben we een niet-lineaire kanonische diskriminant analyse. We noemen dit model CRIMINALS. Kanonische diskriminant analyse en kanonische korrelatie analyse zijn in elkaar om te rekenen, zodat we dus ook met een vrij eenvoudige ingreep CANALS in CRIMINALS kunnen veranderen. In de multivariate analyse wordt het kanonische diskriminant analyse probleem in het algemeen geformuleerd als het zoeken van richtingen die het verschil tussen de gemiddelden van groepen objekten maksimaliseren. Noemen we de binnengroepskovariantie-matriks W/n, de tussengroepskovariantie-matriks B/n , de variantie-kovariantie-matriks T/n en de kanonische richtingen K, dan geldt:

$BK=WK\Phi$, met $K'WK=nI$ en Φ een diagonale matriks.

Kanonische korrelatie analyse was:

Minimaliseer $SSQ(Q_1A_1-Q_2A_2)$ over A_1 en A_2 onder voorwaarde dat:

$A_1'Q_1'Q_1A_1=nI$ en $A_2'Q_2'Q_2A_2=nI$

We gebruiken de volgende notatie:

$Q_1=Q$ oorspronkelijke data

$\bar{Q}$ = matriks Q in afwijking van het gemiddelse

$Q_2=G$ indikatormatriks voor groepsindeling

$G'G=D$ en $u'G=d$, D een diagonale matriks met d op de diagonaal

Voor kanonische diskriminant analyse geldt:

$G'\bar{Q}$ bevat de groepstotalen

D de marginale frekwenties

$D^{-1}G'\bar{Q}K$ de projekties van de groepsgemiddelden op de kanonische assen

Korrelatiematriks $(\bar{Q},\bar{Q}K)$ de totale struktuur

Omgerekend in kanonische korrelatie termen is dit:

$K=(\mathrm{diag}(\bar{Q}'\bar{Q})^{-\frac{1}{2}}A_1(I-\Lambda)^{-\frac{1}{2}}$, Λ is een diagonale matriks met de kanonische korrelatiekoëfficienten op de diagonaal.

$D^{-1}G'\bar{Q}K=(I-\frac{ud'}{n})(\mathrm{diag}(D-\frac{dd'}{n}))^{-\frac{1}{2}}A_2\Lambda(I-\Lambda)^{-\frac{1}{2}}$

Korrelatiematriks $(\bar{Q},\bar{Q}K)$ = korrelatiematriks (Q,QA_1)

De eigenwaarden $\Phi=\Lambda^2/(I-\Lambda^2)$

Hiervoor een globaal bewijs.

De problemen A) tm F) kunnen in elkaar omgerekend worden.

A) $BK=WK\Phi$ met $K'WK=nI$

B) $BL=TL\Psi$ met $L'TL=I$ komt overeen met A) als

$K=L(I-\Psi)^{-\frac{1}{2}}$ en $\Phi=\Psi/(I-\Psi)$

Gebruik voor probleem C) $B=\bar{Q}'GD^{-1}\bar{Q}$ en $T=\bar{Q}'\bar{Q}$

C) $G'\bar{Q}(\bar{Q}'\bar{Q})\bar{Q}'GM=DM\Delta$ met $M'DM=I$ komt overeen met B) als

$L=(\bar{Q}'\bar{Q})^{-1}\bar{Q}'GM$ en $\Psi=\Delta$

D) $G'\bar{Q}P=DZ\Theta$ en $\bar{Q}'GZ=PV\Theta$ met $Z'DZ=nI$ en $P'TP=nI$

komt overeen met C) als $M=Z$ en $\Delta=\Theta^2$ en

komt overeen met B) als $L=P$ en $\Psi=\Theta^2$

E) $\bar{G}'\bar{Q}R=\bar{G}'\bar{G}S\Omega$ en $\bar{Q}'\bar{G}S=TR\Omega$ met $S'\bar{G}'\bar{G}S=nI$ en $R'TR=nI$, waarbij

$\bar{G}=(I-\frac{uu'}{n})G$, dwz matriks G in afwijking van het gemiddelde.

Dit komt overeen met D) als $P=R$, $Z=(I-\frac{uu'D}{n})S$ en $\Theta=\Omega$

F) $\bar{\bar{G}}'\bar{\bar{Q}}A_1=\bar{\bar{G}}'\bar{\bar{G}}A_2\Lambda$ en $\bar{\bar{Q}}'\bar{\bar{G}}A_1=\bar{\bar{Q}}'\bar{\bar{Q}}A_2\Lambda$ met $A_1'\bar{\bar{Q}}'\bar{\bar{Q}}A_1=I$ en $A_2'\bar{\bar{G}}'\bar{\bar{G}}A_2=I$

waarbij $\bar{\bar{G}}=\bar{G}(\mathrm{diag}(D-\frac{dd'}{n}))^{-\frac{1}{2}}$ matriks $\bar{G}$ gestandaardiseerd

$\bar{\bar{Q}}=\bar{Q}(\mathrm{diag}(\bar{Q}'\bar{Q})^{-\frac{1}{2}})$ matriks $\bar{Q}$ gestandaardiseerd

Dit komt overeen met E) als

$R=\mathrm{diag}(\bar{Q}'\bar{Q})^{-\frac{1}{2}}A_1$, $S=(\mathrm{diag}(D-\frac{dd'}{n}))^{-\frac{1}{2}}A_2$ en $\Omega=\Lambda$

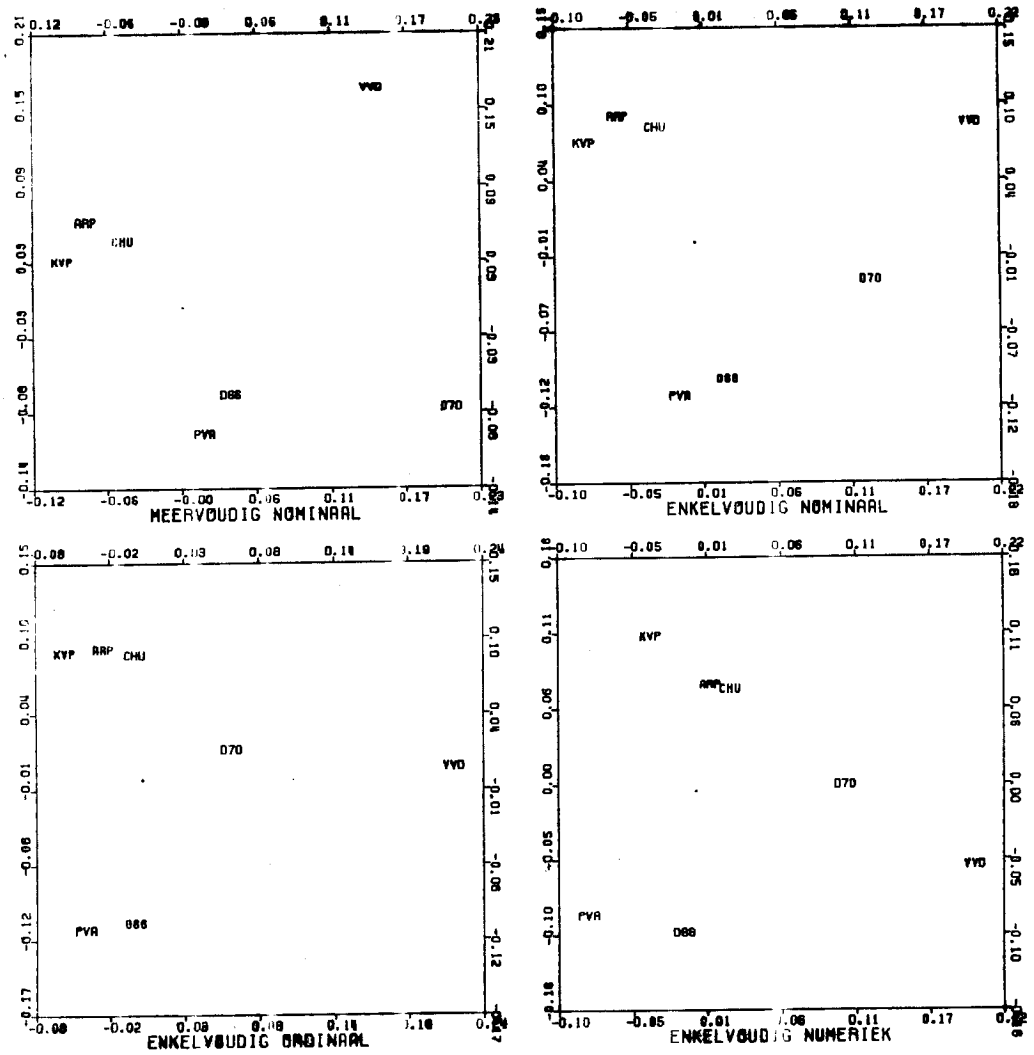
Probleem F) is niets anders dan het kanonische korrelatie probleem. Zoeken we voor de kolommen van Q een nieuwe representatie die in de ruimte ligt van de toegestane transformaties (zie 7.1.2.) en vatten we de kolommen van G meervoudig nominaal op dan krijgen we we terugrekenend van F) naar A) het CRIMINALS probleem. Er betsaat nog geen apart programma voor CRIMINALS, maar dat zal niet lang meer op zich laten wachten .

7.5.2 Toepassing van CRIMINALS

De parlementaire enkete 1972 hebben we gebruikt voor een (niet-linaire) diskriminant analyse. Voor een beschrijving van de data set zie hfdst 7.4.1.2. We hebben alleen de kamerleden van de zeven grootste partijen, die bovendien geen missing skores hadden, meegenomen in de analyse. Dit waren er 119. De meningen van de kamerleden over de zeven issues, laten zich goed voorspellen door hun partijlimaatschap. Wij hebben de issues op vier verschillende manieren geïnterpreteerd, nl meervoudig nominaal, enkelvoudig nominaal, ordinaal en numeriek. De stress van de verschillende analyses staat in tabel 7.7. de stressen lopen nogal uit een. Kijken we naar de ligging van de groepsgemiddelden in de kanonische ruimte, dan zien we dat het grote verschil tussen de konfiguraties de ligging van DS70 is (zie figuur 7.16)

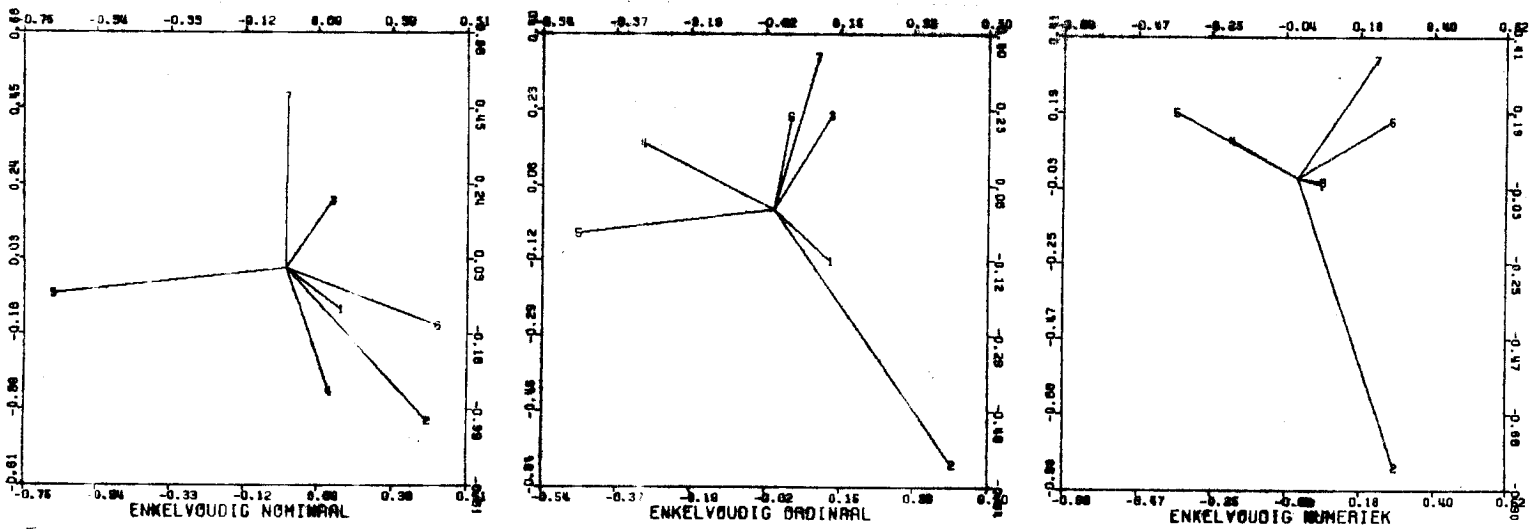
	stress
mv nom	.115
ev nom	.169
ev ord	.223
ev num	.304

Tabel 7.7 stress op vier manieren berekend mbv CRIMINALS



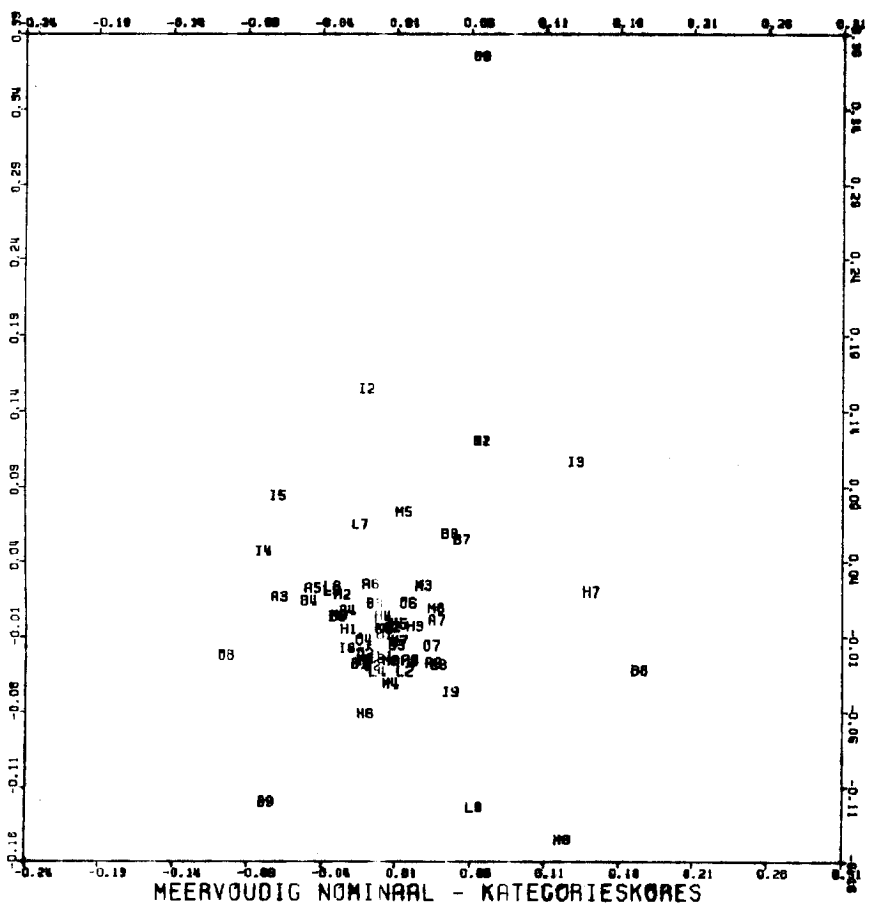
Figuur 7.16 Groepsgemiddelden van de zeven grootste partijen in de kanonische ruimtes. Parlement 1972

We zien ook verschillen tussen de benaderingen in de korrelaties van de variabelen met de kanonische assen (figuur 7.17).In het numerieke geval overheerst het abortusvraagstuk. In het ordinale geval tellen de opvattingen over medezeggenschap en legers ook mee en in het enkelvoudig nominale geval komen daar nog eens de opvattingen over belasting en inkomensverschillen bij.



Figuur 7.17 Korrelaties van de issues met de kanonische assen 1=ontwikkelingshulp, 2=abortus, 3=orde, 4=inkomensverschillens 5=medezeggenschap, 6=belastingen en 7=legers.

Voor het meervoudig nominale geval hebben we de herschaalde kategorieskores geplot (figuur 7.18).

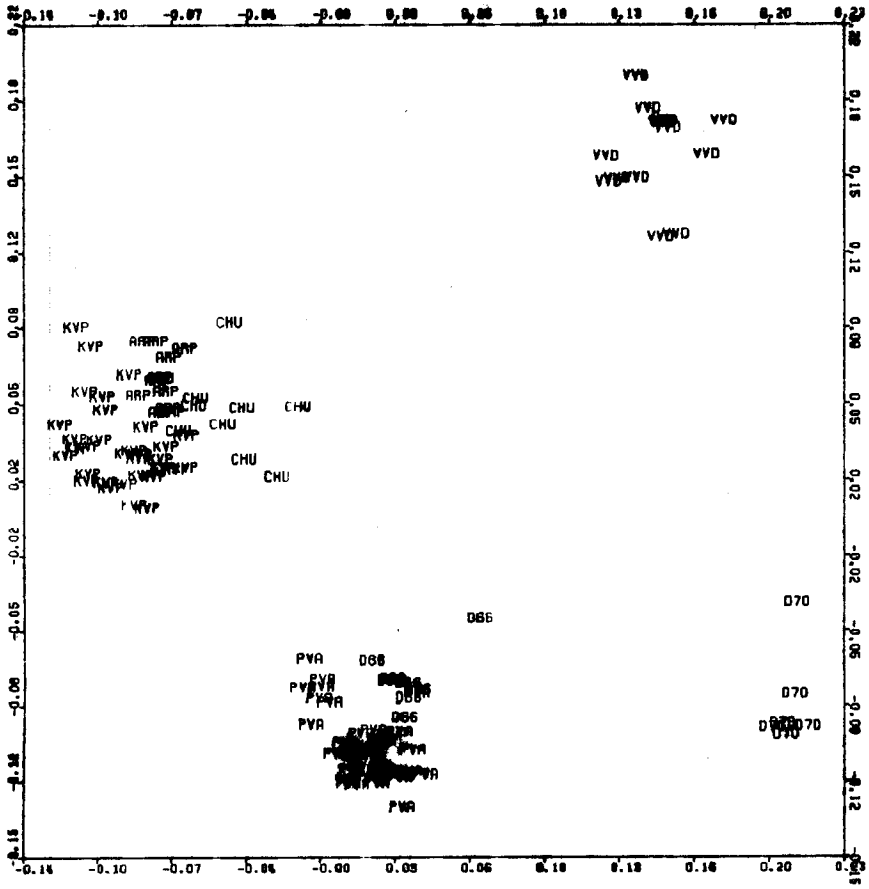


Figuur 7.18 herschaalde kategorieskores voor het meervoudig nominale geval

H=ontwikkelingshulp
A=abortus
O=orde
I=inkomens
M=medezeggenschap
B=belastingen
L=legers

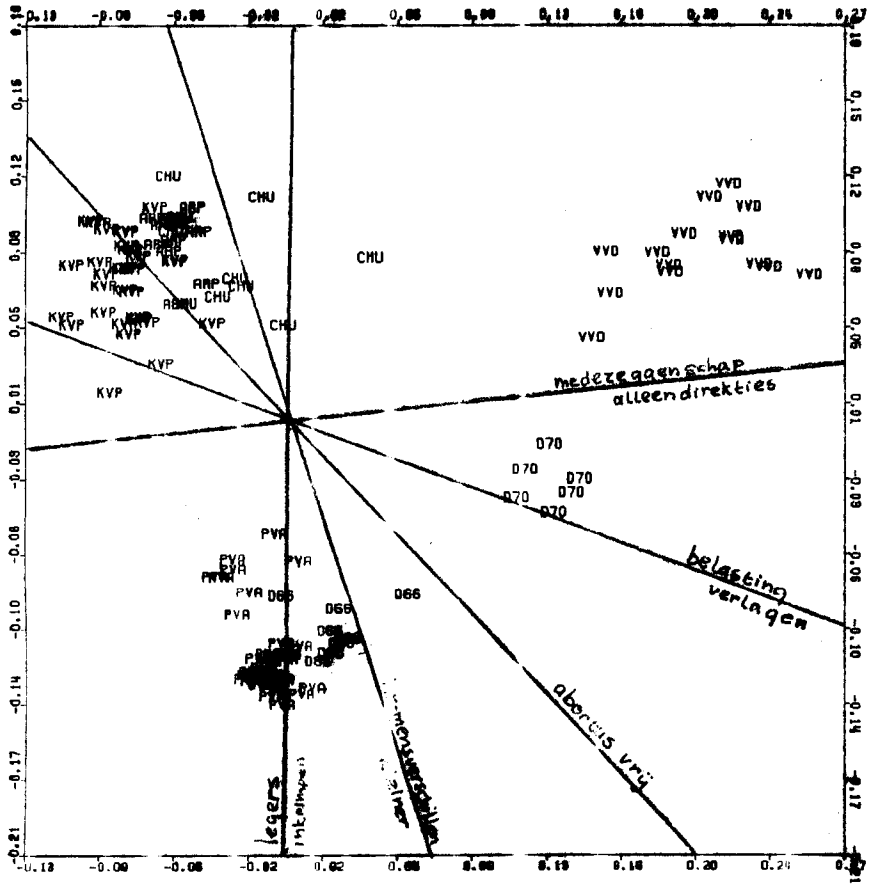
1,..9 oorspronkelijke kategorieskores

Bovenaan de plot ligt O9 (harder optreden). Onderaan liggen B9 (belasting verlagen) en L9 (legers inkrimpen). Rechts onder ligt H8 (minder geld voor ontwikkelingshulp). De socialisten en de liberalen worden voor het merendeel bepaald door ekstreme standpunten, de konfessionelen daarentegen meer door de gematigde skores. De individuen staan geplot in figuur 7.19.

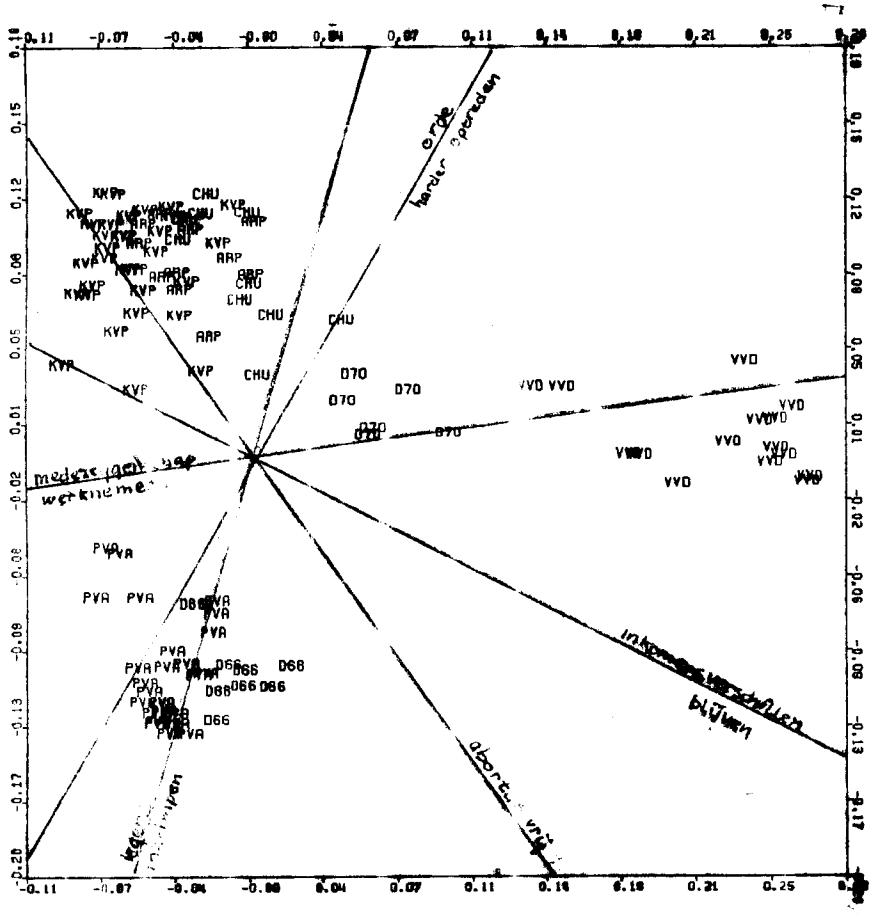


Figuur 7.19
Individueskores in de kanonische ruimte voor de meervoudig nominale oplossing
Parlement 1972

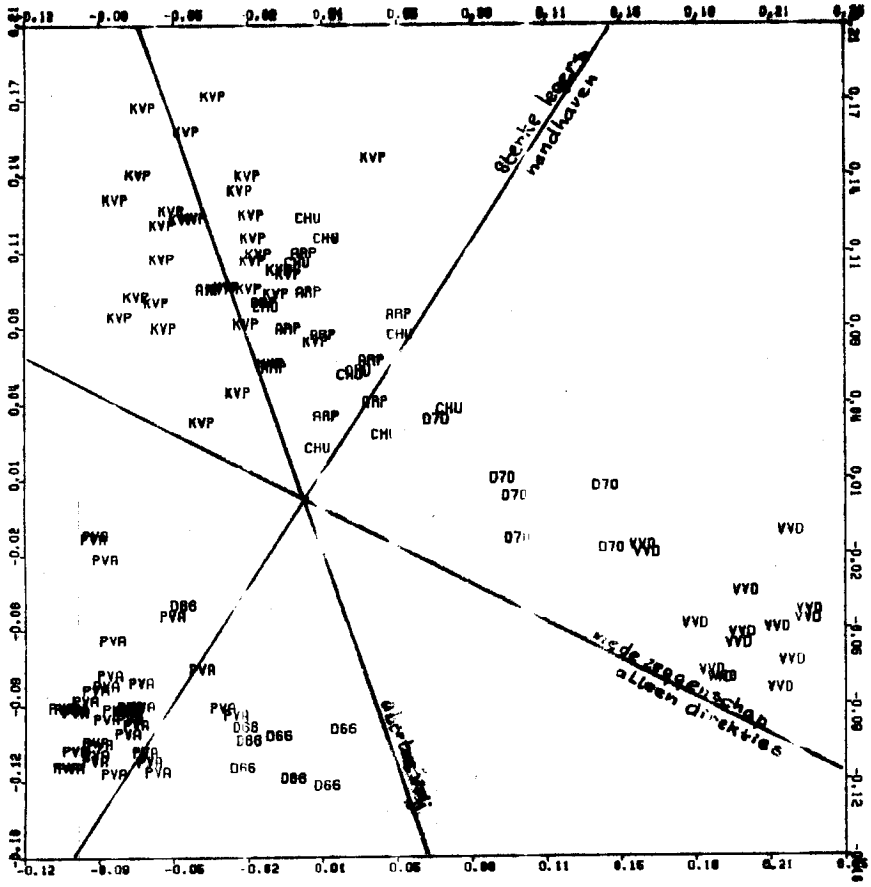
De konfessionelen en de socialisten vormen twee zeer duidelijke klusters. De liberalen daarentegen zijn opgesplitst in VVD'ers en DS70-leden.We hebben ook voor de enkelvoudig nominale, de ordinale en de numerieke oplossing de individuen in de kanonische ruimte geplot (figuur 7.20 tm figuur 7.22). De belangrijkste variabelen uit figuur 7.17 hebben we in de oplossingsruimtes getekend.De socialisten vormen in alle oplossingen een zeer hechte groep, die gemakkelijk van de anderen te scheiden is.De konfessionelen zijn in de numerieke oplossing nogal uitgewaaierd en de demokraten'70 vormen een brug tussen de konfessionelen en de liberalen.De plots spreken heel erg voor zichzelf, daarom zullen we geen verdere details bespreken.



Figuur 7.20
Individuskores in de kanonische ruimte voor de enkelvoudig nominale oplossing
Parlement 1972



Figuur 7.21
Individueskores in de kanonische ruimte voor de ordinale oplossing
Parlement 1972



Figuur 7.22
Individuskores in de kanonische ruimte voor de numerieke oplossing Parlement 1972

7.6. Monte Carlo studie

7.6.1. Inleiding

In de volgende studie bekijken we op drie verschillende manieren niet-lineaire multiple regressie voor gediskretiseerde kontinue variabelen. De eerste manier heeft betrekking op regressie analiese toegepast op de dummyvariabelen die voortkomen uit de diskretisering van de kontinue variabelen. De tweede manier kwantificeert de data eerst door niet-lineaire principale componenten analiese toe te passen op de dummyvariabelen en vervolgens lineaire regressie analiese op de gekwantificeerde variabelen. Voor de volledigheid voegen we een derde manier toe, nl. lineaire regressie analiese, toegepast op de dummyvariabelen die dan numeriek opgevat worden. De skores van de dummyvariabelen zijn rangnummers van het interval waartoe de desbetreffende objekten behoren (zie 7.6.3.). De eerste en derde regressie analieses zijn gedaan m.b.v. CANALS, de tweede analiese is gedaan m.b.v. HOMALS en een APL programma voor regressie analiese.

7.6.2. Data

De data zijn berekend uit een steekproef van een zeven-dimensionale normaalverdeling $(z_1,.....z_7)$. De z-variabelen zijn gestandaardiseerde random variabelen, gegenereerd m.b.v. de SSP procedures Gauss en Randu* (waarbij steeds dezelfde startwaarde is gebruikt). Onze data zijn lineaire kombinaties van de z-variabelen. De variabelen van de eerste set zijn gelijk aan $\alpha z_i + z_6$, $i = 1,...5$ en de variabele van de tweede set is gelijk aan $\sum_{i=1}^{5} z_i + z_7$. De parameter α neemt drie verschillende waarden aan zodat we verschillende interkorrelaties tussen de data vektoren hebben.

7.6.3. Kondities

De variabelen zijn op drie verschillende manieren gediskretiseerd. In drie kategorieën met de intervallen $(-\infty , -.7)$, $(-.7 , +.7)$, $(+.7 , +\infty)$. In vijf kategorieën met de intervallen $(-\infty , -1.5)$, $(-1.5 , -.5)$, $(-.5 , .5)$, $(.5 , 1.5)$ en $(1.5 , +\infty)$ en in tien kategorieën met de intervallen $(-\infty , -2.0)$, $(-2.0 , -1.5)$,..... $(1.5 , 2.0)$, $(2.0 , +\infty)$. De grenzen van de intervallen zijn zo gekozen dat ook in gediskretiseerde vorm de verdeling 'normaal'

* SSP = Scientific Subroutine Package. (360-CM-03x) Version II,IBM

verloopt. We hebben dus niet de optimale diskretisering genomen, zoals in hoofdstuk 4.2. besproken, aangezien we deze random studie uitgevoerd hebben nog voor we de optimale diskretisering berekend hadden. Naast de verschillende diskretiseringen hebben we ook verschillende steekproefgroottes, nl. 20, 100 en 1000 objekten. Tenslotte hebben we zoals hiervoor genoemd, ook nog drie verschillende waarden van de alpha parameter.Uitgaande van z-variabelen die echt kontinue normaal verdeeeld zijn, krijgen we interkorrelaties in de eerste set .1, .5 en .9 respektievelijk voor alpha is 3, 1 en 1/3.
De verschillende kondities zijn dus:

aantal kategorieën	3	5	10
steekproefgrootte	20	100	1000
alpha parameter	1/3	1	3

7.6.4. Benaderingen

Afgezien van negen kondities hebben we ook nog drie verschillende benaderingen van de data m.b.t. hun meetnivo, zoals al in de inleiding vermeld. We hebben nominaal, numeriek na schaling van de nominale data en numeriek. We noemen de verschillende benaderingen naar de komputerprogramma's die ervoor gebruikt zijn, nl. CANALS S(ingle) N(ominal), M(ultiple) R(egressie) na HOMALS en CANALS M(etries).

7.6.5. Random studie

We hebben de resultaten van de verschillende regressie analieses en de schalingen onderling vergeleken en vergeleken met de regressie analiese van de theoretiese data. De theoretiese data zijn de variabelen gebaseerd op z-variabelen die kontinu normaal verdeeld zijn met verwachting nul en variantie één (zie 7.6.2.) De theoretiese data hebben oneindig veel objekten en oneindig veel kategorieën. De enige konditie die op de theoretiese data van toepassing is, is de alpha paramater (zie 7.6.2.). De korrelaties tussen de theoretiese variabelen van de eerste set en tussen de variabelen van de eerste en de tweede set kunnen uitgedrukt worden in alpha: respektievelijk $r_1 = 1/(\alpha^2+1)$ en $r_2 = \alpha/\{6(\alpha^2+1)\}^{\frac{1}{2}}$
De gekwadrateerde multiple korrelatie koëfficient (ρ^2) en de re-

gressie gewichten a_i (identiek voor alle variabelen van de eerste set en gestandaardiseerd zodanig dat $a' Q_1' Q_1 a = 1$) zijn respektievelijk: $\rho^2 = 5\alpha^2/((5 + \alpha^2)6)$ en $(1 + \alpha^2)^{\frac{1}{2}}/\{5(5 + \alpha^2)\}^{\frac{1}{2}}$.
We krijgen de volgende waarden voor r_1, r_2, en a_i voor verschillende waarden van α:

α	1/3	1	3
r_1	.9	.5	.1
r_2	.129	.289	.387
ρ	.135	.373	.732
a_i	.209	.258	.378

We geven een overzicht van de multiple korrelatie koëfficienten, weergegeven in percentages van hun werkelijke skores (d.w.z. de multiple korrelatie koëfficienten van de theoretiese data) voor alle kondities en alle benaderingen in tabel 7.8., 7.9, en 7.10. De drie tabellen bevatten dezelfde getallen geordend op verschillende wijzen om er op verschillende manieren naar te kunnen kijken. We verzamelden ook de regressiegewichten voor alle kondities en alle benaderingen (tabel 7.12.). In het nominale geval zijn de gewichten zoals ze door CANALS berekend worden vrij om van teken te wisselen, althans samen met de kategorieskores. Voor sommige variabelen is het volkomen duidelijk welk teken gekozen moet worden, ondat de skores ordinaal geschaald zijn (voor alle gevallen $\alpha = 3$), voor andere variabelen is het niet duidelijk. Daarom kiezen we alle tekens positief. We vergelijken de kategorieskores van CANALS SN met die van HOMALS. De kategorieskores van CANALS M zijn niet interessant omdat dit gestandaardiseerde skores zijn van de oorspronkelijke kategorieën (gewogen met de marginale frekwenties). We hebben een plot gemaakt van de resultaten van 1000 objekten, drie α waarden en drie diskretiseringen (figuur 7.23. en 7.24). We hebben eveneens een plot gemaakt van de HOMALS kategorieskores van 100 objekten (figuur 7.26.). In het theoretiese geval van gediskretiseerde normaal verdeelde variabelen, liggen de kategorieskores van een variabele op een rechte lijn, omdat zowel HOMALS als CANALS de data benaderen met een multinormaalverdeling om een optimale homogeniteit of stress te vinden. Dus als we de CANALS en HOMALS plots vergelijken moeten we in gedachten houden dat we in het ideale geval rechte lijnen hadden gehad.

7.6.6. Resultaten

a. Multiple regressie koëfficienten

Tabel 7.8., 7.9. en 7.10. geven de multiple regressie koëfficienten voor elke konditie en benadering van de ware multiple korrelatie koëfficient weer (zie 7.6.5.). Tabel 7.8. bestaat uit drietallen geordend naar steekproefgrootte. De T betekent twintig, H = honderd en D = duizend objekten. Het getal na de letter is het totaal aantal variabelen dat bij de analiese betrokken is en daarna volgt het aantal kategorieën. Voor bijna alle drietallen is het waar dat de multiple korrelatie koëfficient afneemt naarmate de steekproefgrootte toeneemt. Dit gebeurt onafhankelijk van de diskretisering, de korrelatie en de benaderingswijze. We kunnen zelfs zeggen dat het multiple regressie probleem beter opgelost wordt als het aantal objekten toeneemt, omdat de benadering van de kontinu normaal verdeelde variabele ook beter is. We kunnen dit ook zien aan de frekwenties van de variabelen. We hebben ze alleen voor vijf kategorieën weer gegeven (tabel 7.11). De steekproeven met drie en tien kategorieën zijn hetzelfde, alleen de diskretisering is verschillend.
Er zijn enkele uitzonderingen op de regel : hoe groter de steekproef hoe lager de multiple korrelatie koëfficient. Voor $\alpha = 1/3$ en CANALS M ziet de waarde 54 (H63) er erg laag uit. Voor MR na HOMALS en $\alpha = 3$ zijn de waarden 51 (T63) en 94 (H610) aan de lage kant.
Tabel 7.9. geeft dezelfde percentages als tabel 7.8., maar in een andere volgorde. Hier hebben we de drietallen aan de hand van het aantal kategorieën geordend. We zien dat voor drietallen geldt dat hoe lager het aantal kategorieën is, hoe lager de multiple korrelatie koëfficient. Het is duidelijk dat we de beste schatter van de multiple korrelatie koëfficient hebben bij het grootste aantal kategorieën. Door te diskretiseren krijgen we een onderschatting van de multiple korrelatie koëfficient en door het nemen van een steekproef een overschatting van de multiple korrelatie koëfficient. We hebben ook in deze tabel een paar uitzonderingen, maar alleen voor steekproefgroottes van 20 of 100 objekten en niet voor 1000 objekten. In de twee kleinste steekproeven zien onze data er nog niet erg normaal verdeeld uit, de grootste steekproef ziet er wel normaal verdeeld uit (zie tabel 7.11.).
De derde tabel kan op twee manieren bekeken worden, vertikaal en

	-- alfa = 1/3 -			--- alfa = 1 --			--- alfa = 3 --		
T63	444	376	323	213	148	117	121	51	99
H63	227	161	45	118	106	97	98	90	87
D63	116	99	99	95	93	93	86	85	86
T65	741	733	521	268	265	212	137	116	115
H65	415	217	201	181	122	131	108	101	99
D65	148	108	102	102	99	97	93	92	93
T610	741	696	449	268	265	197	137	112	111
H610	583	308	213	235	126	125	120	94	103
D610	217	117	104	118	106	102	102	99	100
	CANALS SN	MR na HOMALS	CANALS M	CANALS SN	MR na HOMALS	CANALS M	CANALS SN	MR na HOMALS	CANALS M

tabel 7.8. Multipele korrelatie koëfficienten uitgedrukt in percentages van de ware waarde van de MC koëfficient (versie 1).

	-- alfa = 1/3 -			--- alfa = 1 --			--- alfa = 3 --		
T63	444	376	323	213	148	117	121	51	99
T65	741	733	521	268	265	212	137	116	115
T610	741	696	499	268	265	197	137	112	111
H63	227	161	45	118	106	97	98	90	87
H65	415	217	201	181	122	131	108	101	99
H610	583	308	213	235	126	125	120	94	103
D63	116	99	99	95	93	93	86	85	86
D65	148	108	102	102	99	97	93	92	93
D610	217	117	104	118	106	102	102	99	100
	CANALS SN	MR na HOMALS	CANALS M	CANALS SN	MR na HOMALS	CANALS M	CANALS SN	MR na HOMALS	CANALS M

tabel 7.9. Multipele korrelatie koëfficienten uitgedrukt in percentages van de ware waarde van de MC koëfficient (versie 2).

	3 kategorieën			5 kategorieën			10 kategorieën		
α=1/3	444	376	323	741	733	521	741	696	499
α=1	213	148	117	268	265	212	268	265	197
α=3	121	51	99	137	116	115	137	112	111
α=1/3	227	161	45	415	217	201	583	308	213
α=1	118	106	97	181	122	131	235	126	125
α=3	98	90	87	108	101	99	120	94	103
α=1/3	116	99	99	148	108	102	217	117	104
α=1	95	93	93	102	99	97	118	106	102
α=3	86	85	86	93	92	93	102	99	100
	CANALS SN	MR na HOMALS	CANALS M	CANALS SN	MR na HOMALS	CANALS M	CANALS SN	MR na HOMALS	CANALS M

tabel 7.10. Multipele korrelatie koëfficienten uitgedrukt in percentages van de ware waarde van de MC koëfficient (versie 3).

2	6	9	2	1	3	27	47	18	5	55	230	395	254	66
1	8	6	5	0	3	27	41	23	6	49	242	391	258	60
2	8	5	5	0	6	30	37	25	2	58	233	395	254	60
1	10	5	4	0	3	32	39	23	3	47	241	401	248	63
1	8	7	4	0	2	26	44	24	4	52	216	404	266	62
1	2	14	3	0	11	19	42	24	4	75	224	375	262	64

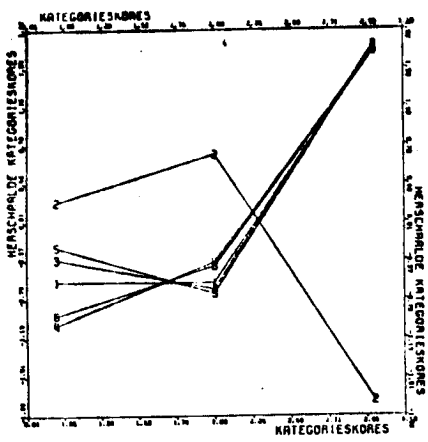
tabel 7.11. Marginale frekwenties 5 kat.; 20 obj.(links), 100 obj.(midden) en 1000 obj.(rechts).

vloed op de gewichten en van alle drie benaderingen zien de gewichten voor alpha = 3 er het beste uit omdat zij het meest op de ware regressie gewichten lijken. Het positieve teken bij de CANALS SN gewichten is zoals eerder vermeld kunstmatig ingevoerd (zie 7.6.5.).

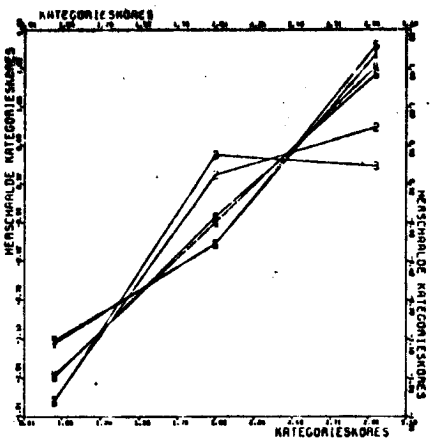
c. Kategorieskores

We hebben plots gemaakt van de kategorieskores voor duizend objekten voor CANALS SN en HOMALS. In het ideale geval verwachten we rechte lijnen zowel voor CANALS als voor HOMALS (zie 7.6.5.) Figuur 7.23. bevat negen CANALS plots. We zien dat CANALS beter fit als de korrelaties tussen de variabelen van de eerste set afnemen, voor HOMALS lijkt het net omgekeerd, voor kleinere interkorrelaties fitten de kategorieskores slechter dan voor grotere (zie figuur 7.24.)* . We zien ook dat de skores van variabele 6, die in de tweede set zit, zich anders gedragen dan de skores van de eerste vijf variabelen. Dit komt omdat de korrelaties tussen de variabelen van de eerste en de tweede set verschillen van de interkorrelaties in de eerste set (zie 7.6.5.). Als we de CANALS plots vergelijken met de HOMALS plots dan zien we dat de HOMALS plots beter zijn in die zin dat de kategorieskores van HOMALS meer op een rechte lijn liggen dan de kategorieskores van CANALS. We hebben daarom van de CANALS kategorieskores geen verdere plots gemaakt, maar wel voor de HOMALS kategorieskores, nl. voor 100 objekten (figuur 7.26.). Vergeleken met de 'duizend'-plots van HOMALS zijn deze plots niet zo goed. Maar vergeleken met de "duizend"-plots van CANALS zijn deze plots niet slecht uitgevallen. We kunnen hieruit konkluderen dat de schaling van HOMALS veel stabieler is dan die van CANALS. Figuur 7.25. geeft de beste plot van CANALS en de slechtste van HOMALS voor 10 kategorieën gezamenlijk weer. Behalve dat HOMALS ongevoeliger is voor de steekproefgrootte dan CANALS is HOMALS ook ongevoeliger voor de onder-

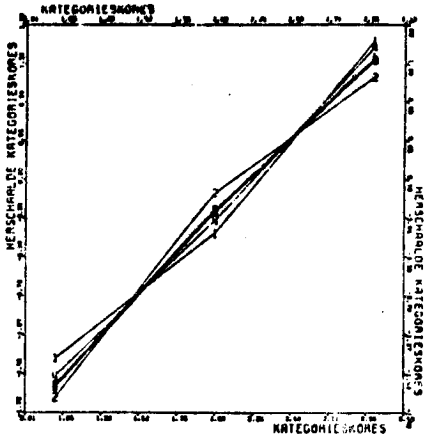
* Dit stemt overeen met de theorie over HOMALS en CANALS. De principale assen van HOMALS zijn gevoelig voor kleine veranderingen in de data bij onafhankelijkheid van de variabelen. Bij CANALS zijn de kanonische assen juist stabiel bij onafhankelijkheid van de variabelen en de assen zijn gevoelig voor kleine veranderingen in de data bij grote multikollineariteit (Björck en Golub, 1973).



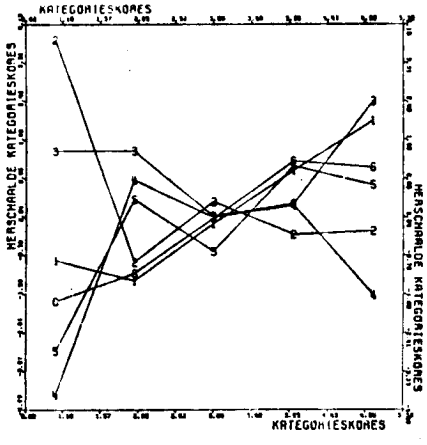
CANALS RANDOM STUDIE D63 ALFA=1/3



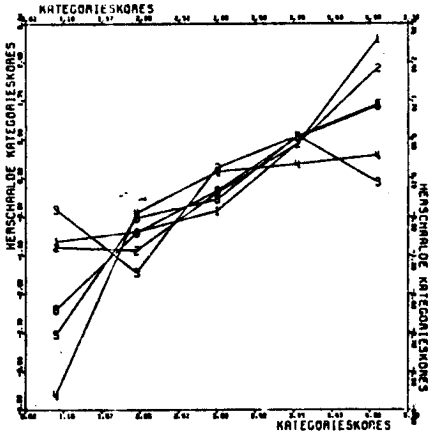
CANALS RANDOM STUDIE D63 ALFA=1



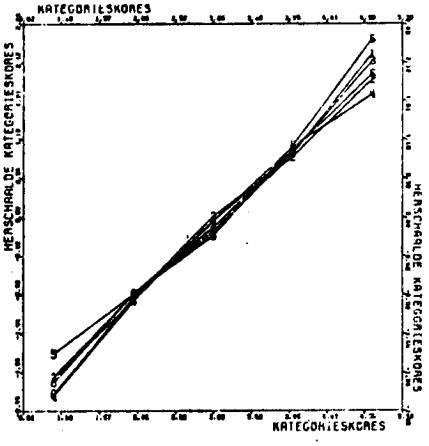
CANALS RANDOM STUDIE D63 ALFA=3



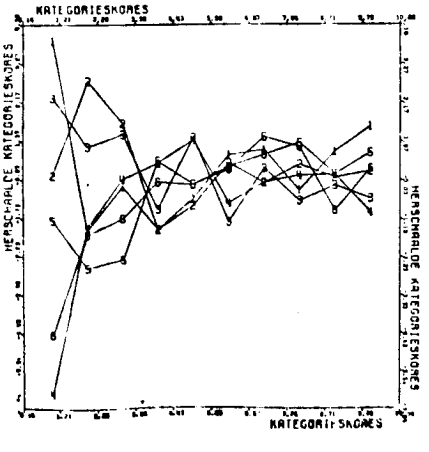
CANALS RANDOM STUDIE D65 ALFA=1/3



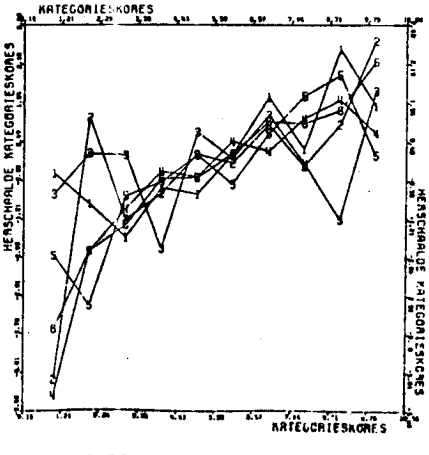
CANALS RANDOM STUDIE D65 ALFA=1



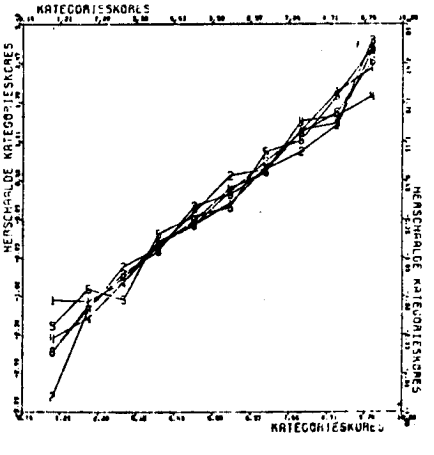
CANALS RANDOM STUDIE D65 ALFA=3



CANALS RANDOM STUDIE D610 ALFA=1/3

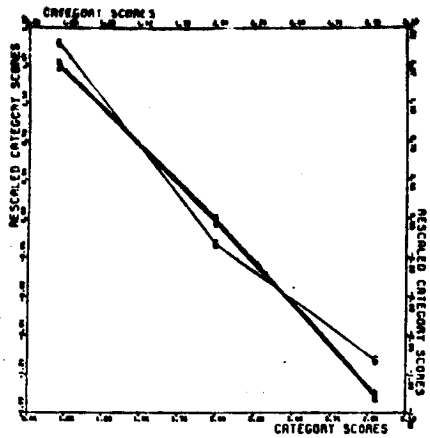


CANALS RANDOM STUDIE D610 ALFA=1

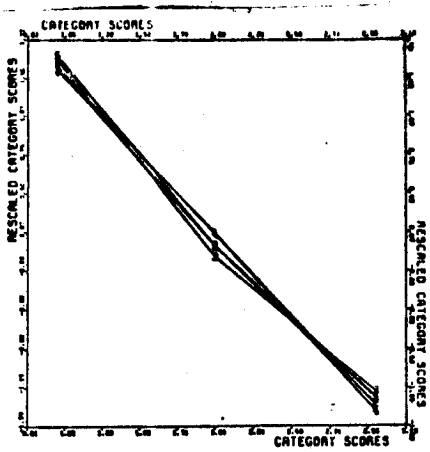


CANALS RANDOM STUDIE D610 ALFA=3

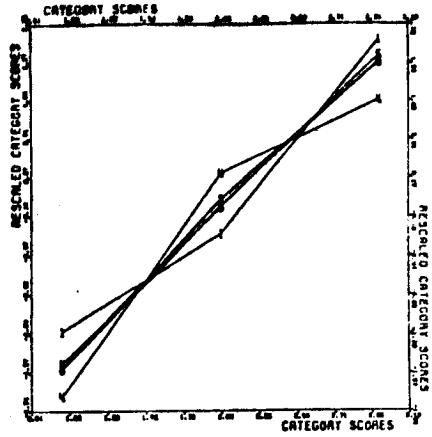
figuur 7.23. CANALS kategorie-skores 1000 objekten.



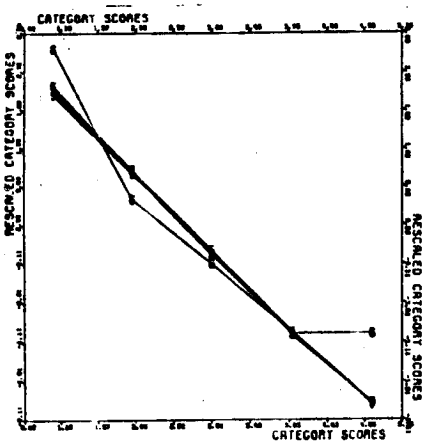
HOMALS RANDOM STUDY D63 ALFA=1/3



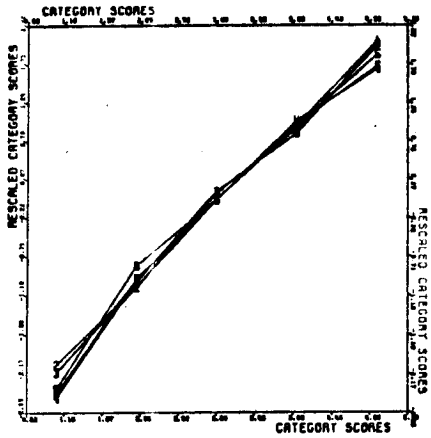
HOMALS RANDOM STUDY D63 ALFA=1



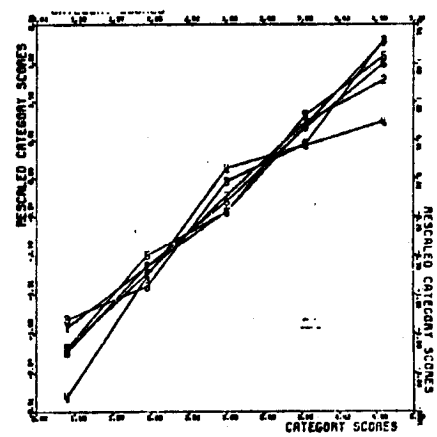
HOMALS RANDOM STUDY D63 ALFA=3



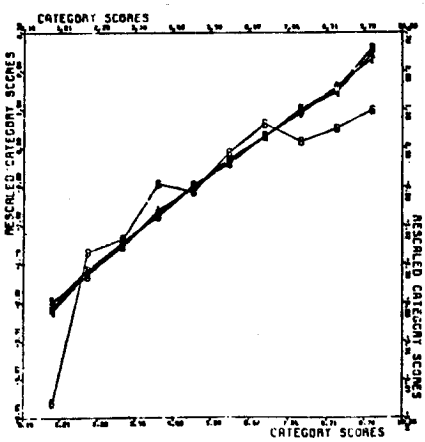
HOMALS RANDOM STUDY D65 ALFA=1/3



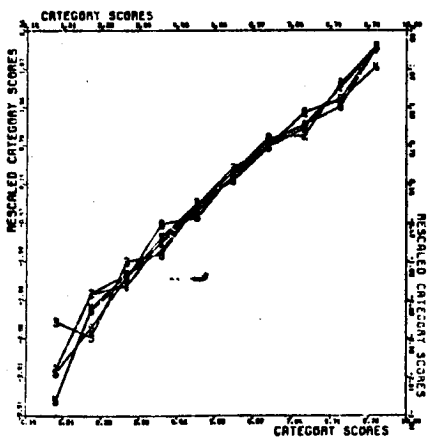
HOMALS RANDOM STUDY D65 ALFA=1



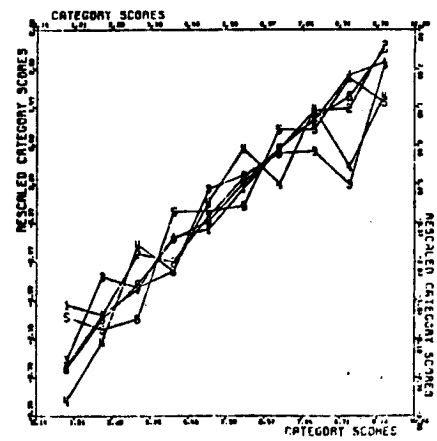
HOMALS RANDOM STUDY D65 ALFA=3



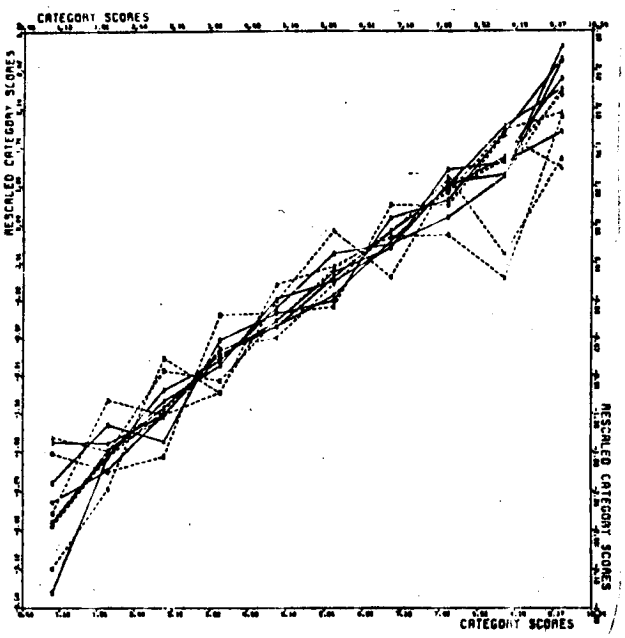
HOMALS RANDOM STUDY D610 ALFA=1/3



HOMALS RANDOM STUDY D610 ALFA=1

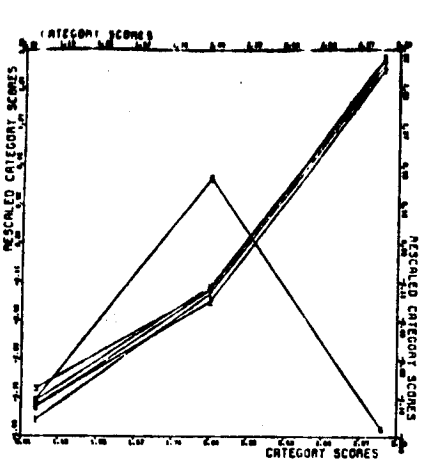


HOMALS RANDOM STUDY D610 ALFA=3

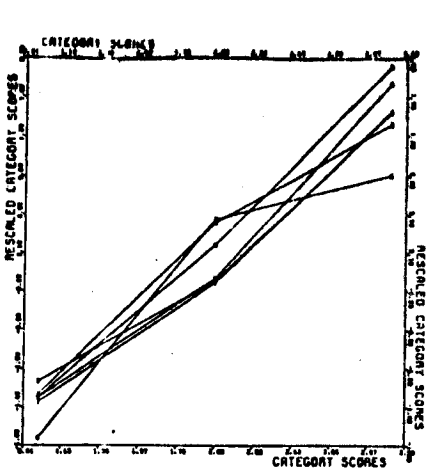


figuur 7.24. HOMALS kategorie-skores 1000 objekten.

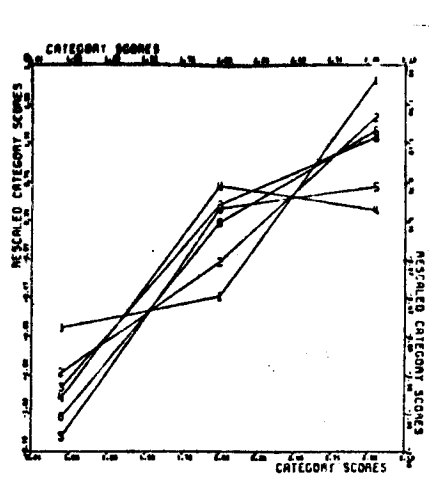
figuur 7.25. CANALS en HOMALS skores (lijnen resp. streepjes); random studie D610, alfa=3.



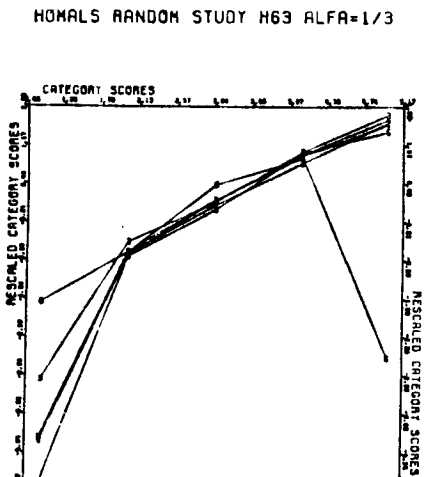
HOMALS RANDOM STUDY H63 ALFA=1/3



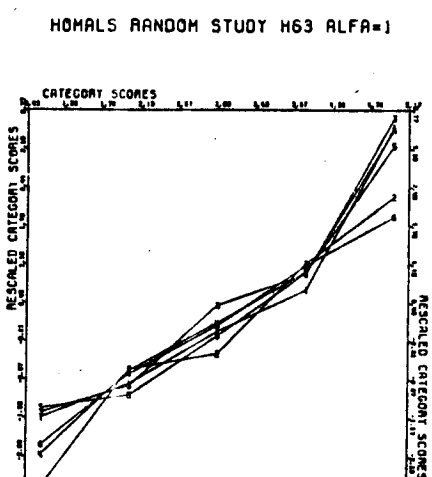
HOMALS RANDOM STUDY H63 ALFA=1



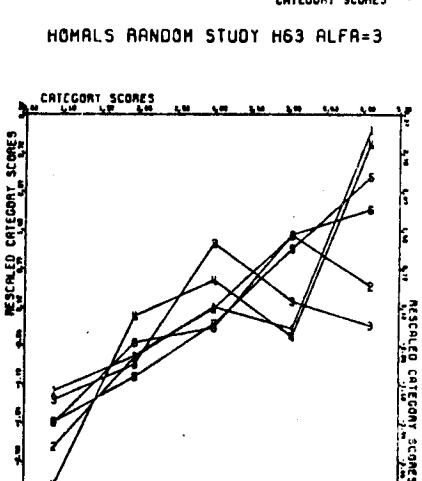
HOMALS RANDOM STUDY H63 ALFA=3



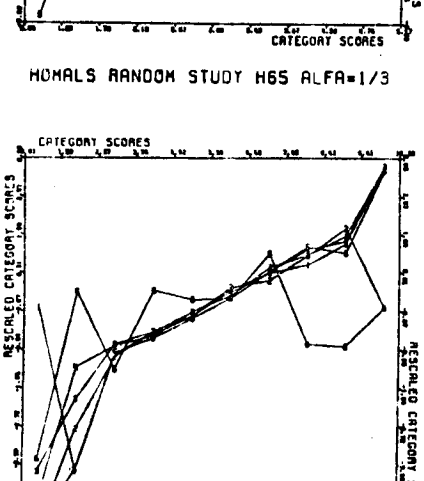
HOMALS RANDOM STUDY H65 ALFA=1/3



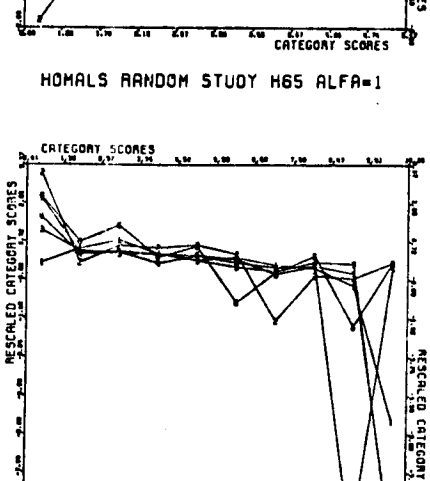
HOMALS RANDOM STUDY H65 ALFA=1



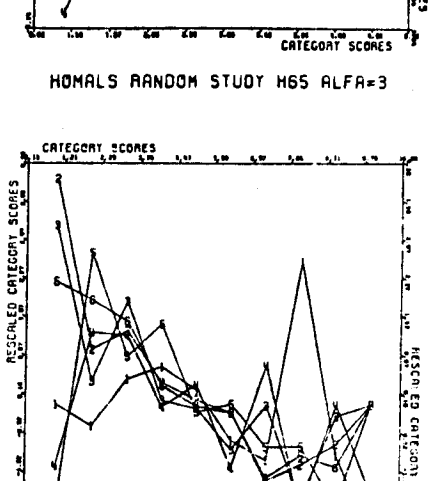
HOMALS RANDOM STUDY H65 ALFA=3



HOMALS RANDOM STUDY H610 ALFA=1/3



HOMALS RANDOM STUDY H610 ALFA=1



HOMALS RANDOM STUDY H610 ALFA=3

figuur 7.26. HOMALS kategorie-skores 100 objekten.

horizontaal. Wederom hebben we de drietallen met een speciaal patroon. In de vertikale richting neemt de multiple korrelatie koëfficient binnen drietallen af als de alpha toeneemt (d.w.z. de interkorrelaties tussen de variabelen van de eerste set kleiner worden). Dit is een bekend verschijnsel bij multiple regressie technieken. Grote multikollineariteit onder de variabelen leidt tot instabiele oplossingen, zie opmerking in 7.6.6.c. Er is slechts één uitzondering op, hoe groter alpha hoe lager de multiple korrelatie koëfficient. In het geval H63, CANALS Metries en alpha = 1/3, is de waarde 45 erg laag zoals we al eerder in tabel 7.8. gezien hebben. Als we de tabel in horizontale richting beschouwen, vergelijken we de verschillende benaderingen. De multiple korrelatie koëfficient neemt toe naarmate de benadering meer nominaal is. We overschatten als we de data als nominaal i.p.v. metries opvatten. Er zijn nogal wat uitzonderingen op de laatste regel, maar de meeste uitzonderingen zien er niet erg serieus uit. Alleen in het T63 en H610 geval voor alpha = 3 zijn de waarden 51 en 94 laag, wat we reeds gekonstateerd hebben bij de bespreking van tabel 7.8.

Tabel 7.8., 7.9. en 7.10. leiden tot de volgende konklusie. De beste benadering van de multiple korrelatie koëfficient vindt plaats bij 1000 onjekten, tien kategorieën, de kleinste interkorrelaties in de eertse set (alpha = 3) en de meest metrische benaderingswijze (CANALS Metries), maar álle drie de benaderingen zijn redelijk onder de optimale kondities. Zelfs alle "duizend"-oplossingen geven een redelijk resultaat, alhoewel er onmiskenbaar een invloed van de diskretisering en de interkorrelaties is op de multiple korrelatie koëfficient. Om een duidelijker beeld te krijgen van de resultaten hebben we de kategorieskores van alle "duizend"-oplossingen geplot (zie figuur 7.26.).

b. Regressiegewichten

We hebben alle berekende regressiegewichten met de waarde van de ware regressie gewichten (tab.7.12.) vergeleken.Deze waarde wordt genoemd bovenaan tabel 7.12. tesamen met de alpha parameter.Voor elke kolom geldt dat de gewichten beter teruggevonden worden als de steekproefgrootte toeneemt. Het is moeilijk om iets over het aantal kategorieën te zeggen m.b.t. de gewichten. Het lijkt alsof de gewichten niet al te zeer beïnvloed worden door de mate van diskretisering. De korrelaties hebben zo te zien een lichte in-

	-----α = 1/3 a_1=.209----			-----α = 1 a_1 = .258-----			---α = 3 a_1 = .378----		
T63	.707	1.020	.883	.346	.486	.433	.374	.088	.318
	.221	.689	.477	1.226	.476	.407	.794	.829	.691
	.209	-.413	-.616	1.197	-.038	.050	.344	.090	.296
	.667	-.260	.146	.410	.404	.499	.391	.083	.542
	.537	-.378	.175	.441	-.224	.160	.291	.167	.032
H63	.820	-.743	.824	.355	-.216	-.226	.299	.008	.244
	.292	.351	.622	.829	.849	.824	.482	.485	.453
	.788	-.159	-.999	.435	.416	.391	.440	.440	.488
	.800	.524	.176	.094	-.067	-.089	.263	.183	.218
	1.132	.972	-.324	.148	.118	.248	.519	.518	.464
D63	.717	.574	.657	.366	.350	.359	.415	.411	.415
	.848	-.266	-.293	.254	.243	.225	.353	.353	.349
	.153	-.438	-.476	.130	.115	.110	.339	.339	.344
	.091	.446	.521	.339	.341	.348	.438	.411	.440
	.025	.654	.542	.350	.345	.349	.447	.452	.443
T65	.009	.11	-.383	.202	-.079	-.250	.163	-.141	.215
	.804	.374	.531	.364	.316	.327	.210	.22	.444
	.015	-.164	.212	.282	.015	.301	.669	.534	.486
	.166	.684	.842	.130	.045	.343	.359	-.002	.462
	.044	-.005	-.185	.444	.682	.474	.702	.580	.591
H65	.555	-1.170	-1.246	.733	-.058	-.207	.162	.126	.212
	.312	.815	1.196	.475	.620	.703	.618	.531	.583
	.144	-.301	.357	.326	.054	.174	.302	.232	.347
	.463	.824	-.606	.332	-.072	-.154	.228	.229	.211
	.597	.429	.675	.766	.574	.552	.502	.538	.518
D65	.499	.829	.826	.340	.314	.322	.407	.404	.560
	.482	-.244	-.228	.256	.250	.269	.363	.364	.509
	.243	.205	.251	.174	.133	.140	.350	.348	.485
	.666	.005	-.149	.289	.259	.241	.399	.379	.476
	.364	.220	.305	.393	.382	.370	.411	.407	.550
T610	.156	.222	-.897	.493	.017	-.043	.327	.049	.194
	.241	.696	.552	.020	.421	.477	.564	.784	.510
	.372	.035	.347	.212	-.088	.160	.201	.306	.426
	.457	.658	.791	.384	.364	.408	.376	.005	.473
	.385	-.207	.107	.495	.276	.281	.355	-.078	.367
H610	.755	-.033	-1.304	.313	.773	-.168	.428	.113	.260
	.457	1.340	1.701	.411	.577	.733	.628	.509	.547
	.547	-.768	.047	.220	-.169	.077	.221	.132	.339
	.222	-.433	-1.018	.244	.687	-.070	.378	.224	.260
	.796	.668	.000	.385	-.710	.535	.457	.452	.441
D610	.636	.652	.007	1.066	.297	.333	.407	.393	.416
	.335	.182	.009	.291	.336	.282	.373	.366	.371
	.440	-.223	-.244	.530	.091	.104	.351	.399	.352
	.825	-.247	-.291	.390	.251	.270	.390	.359	.394
	.501	.632	.589	.449	.324	.325	.404	.404	.397

Tabel 7.12. Regressie gewichten, T,H,D = 20, 100, 1000 obj., 6 variabelen, 3, 5, 10 = aantal kategorieën; de kolommen bij elke waarde van alpha staan in de volgorde CANALS-SN, MR na HOMALS, CANALS-M.

linge korrelaties. Uit de plots valt verder af te leiden dat de diskretisering zoals gedefinieerd in deze studie niet al te veel invloed heeft op de kategorieskores.

7.6.7. Konklusie

De plots van de kategorieskores leiden tot de veronderstelling dat we beter HOMALS kunnen gebruiken als schalingstechniek vóór-dat we CANALS toepassen dan CANALS direkt gebruiken met de nominale optie. De vergelijking van de multiple korrelatie koëfficienten ondersteunt deze veronderstelling ook, omdat de meer metrische benaderingen dichter bij de ware waarde van de multiple korrelatie koëfficienten komen dan de niet metrische benaderingen. Het feit dat HOMALS minder gevoelig is voor verschillende interkorrelaties dan CANALS is nog een ondersteuning van onze veronderstelling. Bij gebruik van multivariate technieken voor numerieke data adviseren veel onderzoekers om een multiple regressie analiese te doen op de principale komponenten i.p.v. op de ruwe data, wat op hetzelfde neer komt als onze veronderstelling. We willen nl. geen foutenvariantie verklaren maar alleen echte variantie en we kunnen daarvan zeker zijn door de foutenvariantie 'kwijt te raken' door de principale komponenten te nemen. Bovendien zijn principale komponenten onderling onafhankelijk.
De resultaten van de gewichten geven geen duidelijke aanwijzing over hoe we onze data moeten behandelen. Het enige duidelijke feit dat uit de gewichtentabel naar voren komt is het feit dat duizend objekten betere resultaten opleveren dan twintig of honderd.
De steekproefgrootte komt ook uit de vergelijking van de multiple korrelatie koëfficienten naar voren als een belangrijke faktor. Om die reden is het aan te raden om bij gebruik van CANALS Single Nominal een niet te kleine steekproef te hebben. De laatste opmerking heeft betrekking op het aantal kategorieën. We zagen enige invloed van de mate van diskretisering op de multiple korrelatie koëfficient, maar bij de gewichtentabellen en de kategorieplots was deze niet zichtbaar. Daarom konkluderen we voorlopig dat de diskretisering er niet al te veel toe doet, maar dat zou nog beter onderzocht moeten worden.

8: Van alles en nog wat

8.0 Inleiding

Een aantal onderwerpen die in hoofdstuk 1 en ook in latere hoofdstukken kort genoemd zijn komen hier nogmaals wat uitvoeriger aan de orde. Met name wordt aandacht besteed aan de relatie van onze technieken met andere vormen van MVA, en aan de theorie van optimale schaling.

8.1 Ontbinding van bivariate verdelingen

In hoofdstuk 2 hebben we gezien dat bivariate verdelingen in het algemeen kanonische ekspansies hebben. Daarmee bedoelen we het volgende. Stel $\underline{h}_1$ en $\underline{h}_2$ zijn stochastische variabelen en $\phi_1(\underline{h}_1)$ en $\phi_2(\underline{h}_2)$ zijn reele funkties met eindige variantie. Dan kunnen we $\phi_1(\underline{h}_1)$ en $\phi_2(\underline{h}_2)$ schrijven in de vorm

$$\phi_1(\underline{h}_1) = \sum \alpha_{s1} g_{1s}(\underline{h}_1),$$

$$\phi_2(\underline{h}_2) = \sum \alpha_{s2} g_{2s}(\underline{h}_2),$$

waarbij $g_{1s}(\underline{h}_1)$ en $g_{2s}(\underline{h}_2)$ ortogonale bases zijn, dus

$$E(g_{1s}(\underline{h}_1) g_{1t}(\underline{h}_1)) = \delta^{st},$$

$$E(g_{2s}(\underline{h}_2) g_{2s}(\underline{h}_2)) = \delta^{st}.$$

Gegeven de bases kunnen we de α_s berekenen door

$$\alpha_{s1} = E(g_{1s}(\underline{h}_1)\phi_1(\underline{h}_1)),$$

$$\alpha_{s2} = E(g_{2s}(\underline{h}_2)\phi_2(\underline{h}_2)).$$

We veronderstellen bovendien dat $g_{10}(\underline{h}_1) = g_{20}(\underline{h}_2) \equiv 1$, zodat voor alle $s > 0$

$E(g_{1s}(\underline{h}_1)) = E(g_{2s}(\underline{h}_2)) = 0,$

en zodat

$$\alpha_{01} = E(\phi_1(\underline{h}_1)),$$

$$\alpha_{02} = E(\phi_2(\underline{h}_2)).$$

Definieer

$$\lambda_{st} \triangleq E(g_{1s}(\underline{h}_1) g_{2t}(\underline{h}_2)),$$

dan geldt natuurlijk dat $\lambda_{00} = 1$, $\lambda_{0s} = \lambda_{s0} = 0$ voor alle $s > 0$, en λ_{st} voor $s > 0$ en $t > 0$ kan opgevat worden als de <u>korrelatie</u> tussen $g_{1s}(\underline{h}_1)$ en $g_{2t}(\underline{h}_2)$.

De boven ontwikkelde notatie kan gebruikt worden om een eenvoudige representatie tussen de kovariantie van funkties te geven. Immers

$$E(\phi_1(\underline{h}_1)\phi_2(\underline{h}_2)) = \sum_{s=0}^{\infty} \sum_{t=0}^{\infty} \alpha_{s1}\alpha_{s2}\lambda_{st},$$

en daardoor ook

$$C(\phi_1(\underline{h}_1)\phi_2(\underline{h}_2)) = \sum_{s=1}^{\infty} \sum_{t=1}^{\infty} \alpha_{s1}\alpha_{t2}\lambda_{st}.$$

Meer in het bizonder

$$V(\phi_1(\underline{h}_1)) = \sum_{s=1}^{\infty} \alpha_{s1}^2,$$

$$V(\phi_2(\underline{h}_s)) = \sum_{s=1}^{\infty} \alpha_{s2}^2.$$

Een representatie van de kovarianties van deze vorm heet kanonisch wanneer we onze bases zo gekozen hebben dat $\lambda_{st} = 0$ wanneer $s \neq t$. In dat geval noemen we de λ_{ss} (die we dan natuurlijk gewoon λ_s kunnen noemen) kanonische korrelatie-koefficienten. De representatie is dan dus

$$C(\phi_1(\underline{h}_1)\phi_2(\underline{h}_2)) = \sum_{s=1}^{\infty} \lambda_s \alpha_{s1}\alpha_{s2}.$$

Het prettige is dat zo'n kanonische representatie voor een zeer grote klasse van bivariate verdelingen mogelijk is (Cambanis en Liu, 1971, Chesson, 1976).

Wat gebeurt er met deze theorie in het multivariate geval ? We beperken ons even tot drie variabelen, omdat het duidelijk is hoe we naar meer dan drie variabelen kunnen generaliseren, als we maar diep genoeg in onze voorraad indices tasten. Het is nog steeds waar dat we altijd een representatie kunnen vinden van de vorm

$$C(\phi_1(\underline{h}_1)\phi_2(\underline{h}_2)\phi_3(\underline{h}_3)) = \sum_{s=1}^{\infty}\sum_{t=1}^{\infty}\sum_{u=1}^{\infty} \lambda_{stu}\,\alpha_{s1}\alpha_{t2}\alpha_{u3}.$$

Ongelukkigerwijs is dit echter aanzienlijk minder interessant, en wel om twee redenen. In de eerste plaats zijn de $C(\phi_1(\underline{h}_1)\phi_2(\underline{h}_2)\phi_3(\underline{h}_3))$ moeilijker te interpreteren dan gewone kovarianties. Dat komt ongetwijfeld door onze multinormale bias, en door het feit dat we als psychometrici jarenlang blootgesteld worden aan een bombardement met korrelaties en kovarianties. In de tweede plaats is er geen eenvoudige kanonische vorm voor drieweg-matriksen, om ze zo maar even te noemen, dat wil zeggen het is niet mogelijk om bases te kiezen zodat $\lambda_{stu} = 0$ wanneer niet $s = t = u$. Er zijn wel data analytische technieken denkbaar op basis van deze representatie, we kunnen bijvoorbeeld de α's zo kiezen dat $C(\phi_1(\underline{h}_1)\phi_2(\underline{h}_2)\phi_3(\underline{h}_3))$ zo groot mogelijk wordt, maar de resultaten zijn moeilijk te interpreteren. En weer voornamelijk omdat we niet inzien wat deze techniek doet in bekende gevallen.

Het is daarom dat onze technieken ook wanneer we multivariate gegevens hebben toch bivariaat blijven. Als $\underline{h}_1,\ldots,\underline{h}_m$ de variabelen zijn, dan gebruiken we de representatie

$$C(\phi_j(\underline{h}_j)\phi_\ell(\underline{h}_\ell)) = \sum_{s=1}^{\infty}\sum_{t=1}^{\infty} \lambda_{j\ell st}\alpha_{sj}\alpha_{t\ell}.$$

We hebben ons daarbij in het bizonder geinteresseerd voor situaties (zoals de

multinormale) waarin we onze bases kunnen kiezen op zo'n manier dat $\lambda_{j\ell st} = 0$ wanneer $s \neq t$ voor alle j en ℓ. In dat geval vereenvoudigt een groot deel van onze theorie, en is het eenvoudig in te zien wat onze algoritmes berekenen. We hebben gezien dat deze theoretisch interessante situatie ook in de praktijk bij benadering dikwijls voorkomt (de zogenaamde hoefijzers-situaties).

8.2 Ontbinding van multivariate verdelingen

We bekijken nu, voorlopig even voor het geval van twee variabelen, een wat algemenere situatie. Stel $\phi(h_1,h_2)$ is een funktie van beide variabelen, niet meer noodzakelijk van de vorm $\phi_1(h_1)\phi_2(h_2)$. We gebruiken onze bases nu om $\phi(h_1,h_2)$ te representeren in de vorm

$$\phi(h_1,h_2) = \sum_{s=0}^{\infty}\sum_{t=0}^{\infty} \beta_{st} g_{1s}(h_1) g_{2t}(h_2),$$

waarbij de koefficienten β_{st} gegeven zijn door

$$\beta_{st} = E_*(\phi(\underline{h}_1,\underline{h}_2) g_{1s}(\underline{h}_1) g_{2t}(\underline{h}_2)).$$

Hierbij betekent E_* dat we de verwachte waarde berekenen met gebruik making van het produkt van de marginalen. Dat wil zeggen we hebben tot zover nog helemaal geen multivariate verdeling nodig, we gebruiken alleen de marginalen. Een dergelijke representatie is mogelijk wanneer

$$E_*(\phi^2(\underline{h}_1,\underline{h}_2)) < \infty.$$

Omdat weer $g_{10}(h_1) = g_{20}(h_2) \equiv 1$ aangenomen wordt, geldt dat

$$\beta_{00} = E_*(\phi(\underline{h}_1,\underline{h}_2)).$$

Het is interessant te kijken naar de β_{s0} met $s > 0$. Deze hangen natuurlijk af van de keuze van de basis. Maar stel nu dat we willen minimaliseren

$$V_*(\phi(\underline{h}_1,\underline{h}_2) - \psi(\underline{h}_1))$$

over ψ zodanig dat $E(\psi(\underline{h}_1)) = 0$. De oplossing is

$$\psi(h_1) = \sum_{s=1}^{\infty} \beta_{s0} g_{1s}(h_1).$$

En die oplossing is natuurlijk onafhankelijk van de keuze van de basis. Op dezelfde manier kunnen we nu inzien dat we $V_*(\phi(\underline{h}_1,\underline{h}_2))$ kunnen partioneren volgens

$$V_*(\phi(\underline{h}_1,\underline{h}_2)) = V_1(\phi(\underline{h}_1,\underline{h}_2)) + V_2(\phi(\underline{h}_1,\underline{h}_2)) + V_{12}(\phi(\underline{h}_1,\underline{h}_2)),$$

met

$$V_1(\phi(\underline{h}_1,\underline{h}_2)) = \sum_{s=1}^{\infty} \beta_{s0}^2,$$

$$V_2(\phi(\underline{h}_1,\underline{h}_2)) = \sum_{s=1}^{\infty} \beta_{0s}^2.$$

$$V_{12}(\phi(\underline{h}_1,\underline{h}_2)) = \sum_{s=1}^{\infty}\sum_{t=1}^{\infty}\beta_{st}^2.$$

De variantiekomponenten zijn onafhankelijk van de keuze van de basés, omdat het de gekwadrateerde afstanden zijn tot deelruimtes die niet in termen van de bepaalde basés die wij toevallig gebruiken gedefinieerd zijn. Vanzelfsprekend geldt wel dat de grootte van de komponenten afhankelijk is van de marginale verdelingen die we gebruikt hebben om de basés te definieren. Dat geldt echter niet voor uitspraken over het verdwijnen van komponenten. Als we zeggen dat $V_{12}(\phi(\underline{h}_1,\underline{h}_2)) = 0$, dan betekent dit dat $\phi(h_1,h_2)$ van de vorm is $\phi(h_1,h_2) = \psi_1(h_1) + \psi_2(h_2)$. Als we zeggen dat $V_{12}(\phi(\underline{h}_1,\underline{h}_2) = V_2(\phi(\underline{h}_1,\underline{h}_2)) = 0$ dan betekent dit dat $\phi(\underline{h}_1,h_2) = \psi_1(h_1)$, en als we zeggen dat alle drie de komponenten nul zijn dan betekent dat dat $\phi(\underline{h}_1,h_2)$ een konstante is (allemaal met *-waarschijnlijkheid gelijk aan één natuurlijk, we hebben het steeds over gelijkheid van stochastische veranderlijken).

Merk op dat we in deze paragraaf nog steeds de bivariate verdeling nergends gebruikt hebben, alleen nog maar de marginalen. Daarom is het ook geen kunst om alles wat we tot nu toe behandeld hebben te generaliseren naar multivariate verdelingen. Er zijn dan wat meer marginalen, β krijgt wat meer indices, en er zijn meer variantiekomponenten, maar wezenlijk verandert er niets. Om een beetje precieser te zijn: als er m variabelen $\underline{h}_1,\ldots,\underline{h}_m$ zijn, en een multivariate funktie $\phi(h_1,\ldots,h_m)$, dan definieren we

$$\beta_{s_1\ldots s_m} \triangleq E_*(\phi(\underline{h}_1,\ldots,\underline{h}_m)g_{1s_1}(\underline{h}_1)\ldots g_{ms_m}(\underline{h}_m)).$$

We hebben $2^m - 1$ variantiekomponenten, allemaal onafhankelijk van de keuze van de bases. Zo'n komponent heeft als index een deelverzameling van $\{1,\ldots,m\}$. Bijvoorbeeld

$$V_J(\phi(\underline{h}_1,\ldots,\underline{h}_m)) \triangleq \sum\{\beta^2_{s_1\ldots s_m} : s_j \neq 0 \text{ als } j \in J\}.$$

Het is ook nuttig om te definieren

$$\overline{V}_r(\phi(\underline{h}_1,\ldots,\underline{h}_m)) \triangleq \sum\{V_J(\phi(\underline{h}_1,\ldots,\underline{h}_m)) : \text{card}(J) = r\},$$

waarbij card(J) het aantal elementen in J is. Met behulp hiervan kunnen we de <u>rang</u> van de funktie $\phi(h_1,\ldots,h_m)$ definieren. We zeggen dat $\phi(h_1,\ldots,h_m)$ van de rang r is als $\overline{V}_s(\phi(\underline{h}_1,\ldots,\underline{h}_m)) = 0$ voor alle $s > r$. Evenals in het bivariate geval is de rang een eigenschap van de funktie, en niet van de marginalen. Een funktie is van de rang nul als hij konstant is, van de rang één als hij van de vorm is $\phi(h_1,\ldots,h_m) = \psi_1(h_1) + \ldots + \psi_m(h_m)$, van de rang twee als hij van de vorm is $\phi(h_1,\ldots,h_m) = \psi_{12}(h_1,h_2) + \ldots + \psi_{m,m-1}(h_m,h_{m-1})$, enzovoorts.

Na al dit gegoochel wordt het tijd om de multivariate verdeling er eens bij te halen. Daar ging het ons tenslotte om. Er zijn nu twee verschillende benaderingen in omloop. In de eerste benadering kiezen we voor $\phi(h_1,\ldots,h_m)$ de dichtheid van de multivariate verdeling ten opzichte van het produkt van de marginalen (in het diskrete geval: de kruistabel gedeeld door het produkt van de marginalen, in het absoluut kontinue geval de multivariate dichtheid gedeeld door het produkt van de marginale dichtheden). Het is handig om ons weer even te beperken tot het bivariate geval. Nu geldt

$$\beta_{s0} = E_*(\phi(\underline{h}_1,\underline{h}_2)g_{1s}(\underline{h}_1)) = E(g_{1s}(\underline{h}_1)) = 0$$

voor alle $s > 0$. En zo natuurlijk ook $\beta_{0s} = 0$ voor alle $s > 0$. Dus geldt voor de dichtheid $p(h_1,h_2)$

$$p(h_1,h_2) = p_1(h_1)p_2(\underline{h}_2)\{1 + \sum_{s=1}^{\infty}\sum_{t=1}^{\infty}\beta_{st}g_{1s}(\underline{h}_1)g_{2t}(h_2)\}.$$

Hier kunnen we de kanonische analyse weer even om de hoek laten kijken. De bases kunnen nu weer zo gekozen worden dat $\beta_{st} = 0$ voor $s \neq t$, en we hebben de bekende ontbinding van de bivariate verdeling teruggevonden. Onder veel restriktievere kondities dan in 8.1 natuurlijk, we moeten hier aannemen dat er een dichtheid bestaat, en dat de dichtheid kwadratisch integreerbaar is. Er is trouwens nog een subtiel verschil tussen 8.1 en 8.2, wat soms blijkt uit het ontbreken van de streepjes onder de h_j. Daar zijn de h_j geen variabelen, maar waarden die de variabelen aannemen. Pas als we verwachte waarden of varianties uitrekenen vatten we de $\underline{h}_j$ als stochastische variabelen op.

In het multivariate geval leidt deze keuze van $\phi(h_1,\ldots,h_m)$ tot de <u>additieve definitie</u> van interaktie. Dit zijn de variantiekomponenten die we eerder gedefinieerd hebben. We zeggen in een trivariaat geval bijvoorbeeld dat tweede orde interaktie ontbreekt wanneer $V_{123}(\phi(\underline{h}_1,\underline{h}_2,\underline{h}_3)) = 0$, dat wil zeggen als en alleen als $\phi(h_1,h_2,h_3)$ van de rang twee is. Naast de additieve definitie is er ook een <u>multiplikatieve</u>. We nemen dan voor $\phi(h_1,\ldots,h_m)$ niet de dichtheid, maar de logaritme van de dichtheid. De belangrijkste reden om die logaritme te nemen is dat onafhankelijkheid in termen van de waarschijnlijkheden multiplikatief gedefinieerd is, en door een logaritmische transformatie additief wordt.

Het idee om onafhankelijkheid in multivariate verdelingen te bestuderen met behulp van ortogonale funkties is aardig oud. In zekere zin gaat het terug tot Pearson (1916), maar de eerste systematische moderne uitwerking kwam van Lancaster. Hij gebruikte de additieve definitie van interaktie (Lancaster 1959, 1960a, 1960b). Toen de diskrete multivariate analyse op gang begon te komen, voornamelijk door het werk van Mosteller en Goodman, werd langzamerhand de multiplikatieve definitie van interaktie, gekoppeld aan loglineaire modellen,

populairder. Lancaster (1971, 1975) en Darroch (1974) vergelijken de twee. De strijd is nog niet beslist. In het speciale geval van binaire gegevens werd de additieve ekspansie van de 2 × ... × 2 tabel ontdekt door Lazarsfeld rond 1950, en herondekt en gepubliceerd door Bahadur (1961). Voor binaire gegevens is de additieve ekspansie van de logits de grootste konkurrent. Cox (1972) vergelijkt de twee. Het formalisme in 8.2 is geinspireerd door Benzécri (1967), die zich op zijn beurt weer baseert op Good (1963).

8.3 Relatie met HOMALS/CANALS

De in 8.2 besproken technieken hebben gemeen dat ze steekproefwaarden $\hat{\beta}_{s_1 \ldots s_m}$ berekenen voor een gegeven $\phi(h_1,\ldots,h_m)$, en vervolgens kijken of een bepaald patroon van de β's gelijk is aan nul. In additieve diskrete MVA bekijken we de kwadratensom van de β's, en vergelijken die met chi-kwadraat met het bijpassende aantal vrijheidsgraden. In multiplikatieve diskrete MVA moeten we eerst de β's nog korrigeren voor varianties en kovarianties alvorens we chi-kwadraat kunnen berekenen. Multiplikatieve analyse heeft trouwens nauwe banden met multinomiale maksimale aannemelijkheid, de verbanden worden misschien het duidelijkst uitgelegd in Good (1963), Benzécri (1967), en Haberman (1974).

In plaats van deze 'globale' tests kunnen we echter ook, met name in het additieve geval, technieken gebruiken die optimale koefficienten berekenen. Vergelijk de bespreking van het boek van Harris in 1.1.9. Als $\underline{h}_j$ in de steekproef de waarden h_{ij} aanneemt (i=1,...,n; j=1,...,m), en we willen dat $\beta_{s_1 \ldots s_m} = 0$, dan definieren we

$$b_{i,s_1 \ldots s_m} \triangleq g_{1s_1}(h_{i1}) \ldots g_{ms_m}(h_{im})$$

Eén manier om uit te vinden of $\beta_{s_1 \ldots s_m} = 0$ is vinden van een lineaire kombinatie

$$u_i = \sum \alpha_{s_1 \ldots s_m} b_{i,s_1 \ldots s_m},$$

met sommatie over alle $s_1,\ldots,s_m$ waarvoor we willen dat $\beta_{s_1 \ldots s_m} = 0$, zodanig dat SSQ(u) zo groot mogelijk is. Als de g_{js} indikatoren zijn, en de $s_1,\ldots,s_m$ waarover we sommeren bevatten steeds maar één $s_j \neq 0$ dan hebben we HOMALS teruggevonden. Als we sommeren over alle m-vouden van indices met twee indices ongelijk aan nul hebben we een generalisatie van HOMALS, die we zoals in 7.5 ook kunnen implementeren door HOMALS toe te passen op alle 'super'-variabelen samengesteld uit twee oorspronkelijke variabelen. Deze generalisatie geeft natuurlijk automatisch een hogere homogeniteit. In de notatie van hoofdstuk 2 willen we hier het verlies

$$\sigma(\underline{z};\Phi) = \frac{1}{M} \sum_{j<\ell}^{m} \sum^{m} V(\underline{z} - \phi_{j\ell}(\underline{h}_j,\underline{h}_\ell))$$

met $M = \frac{1}{2}m(m-1)$ zo klein mogelijk maken. CANALS kan, volkomen analoog hieraan, ingevoerd worden via de verliesfunktie van OVERALS (vergelijk 7.3). Zoals uiteen-

gezet in 7.5 is CANALS hetzelfde als HOMALS met lineaire restrikties. Dat komt in de terminologie van dit hoofdstuk erop neer dat we voor ten minste één van de twee sets eisen dat de transformatie van rang één is.

8.4 Vereniging en doorsnede

In 8.2 hebben we gezien dat multivariate MVA gebruik maakt van wat in de multilineaire algebra het tensor produkt van deelruimtes van een lineaire ruimte heet (bijvoorbeeld Marcus, 1973, hoofdstuk 1). Er zijn echter ook andere manieren om verschillende deelruimtes van een lineaire ruimte te kombineren, en die manieren leiden ons regelrecht naar HOMALS en CANALS. Ze hebben bovendien veel te maken met het verschil tussen analyse van samenhang en ahalyse van afhankelijkheid, wat in hoofdstuk 1 vele malen aan de orde kwam.

De deelruimtes van een lineaire ruimte definieren namelijk een raamwerk (lattice), in de zin dat als L_1 en L_2 deelruimtes zijn van L, dan kunnen we definieren

$L_1 \cap L_2 \triangleq \{x : x \in L_1 \ \& \ x \in L_2\}$,

$L_1 \oplus L_2 \triangleq \{z :$ er is een $x \in L_1$ en een $y \in L_2$ zodanig dat $z = x + y\}$.

De eerste kombinatie noemen we de doorsnede (Engels: meet), de tweede de vereniging (Engels: join). Merk op dat de vereniging niet hetzelfde is als de vereniging in verzamelingstheoretische zin, die vereniging hoeft helemaal geen deelruimte te zijn. De doorsnede is de grootste deelruimte die in zowel L_1 als L_2 ligt, de vereniging is de kleinste deelruimte die zowel L_1 als L_2 bevat. In hoofdstuk 1 hebben we in dit verband al losjes de termen "grootste gemene deler" en "kleinste gemene veelvoud" gebruikt.

Stel nu dat $\underline{h}_1,\ldots,\underline{h}_m$ stochastische variabelen zijn op een waarschijnlijkheidsruimte. Definieer L als de verzameling van alle stochastische variabelen op die ruimte, en geef L het gebruikelijke inwendige produkt. We nemen dus aan dat de elementen van L eindige variantie hebben, als inwendig produkt gebruiken we de kovariantie. Iedere $\underline{h}_j$ definieert een deelruimte L_j van L van funkties $\phi_j(\underline{h}_j)$ met eindige variantie. De $\underline{h}_j$ kunnen vektor-waarden aannemen, in het algemeen is $\underline{h}_j$ een stochastische k_j-vektor.

We definieren nu de 'meet'-rang van de $\underline{h}_j$ als de dimensie van de doorsnede van de L_j, en de 'join'-rang van de $\underline{h}_j$ als de dimensie van de vereniging van de L_j. Het eerste probleem van de lineaire multivariate analyse is een basis te vinden voor de doorsnede van de L_j, het tweede probleem om een basis te vinden voor de vereniging van de L_j. CANALS (of algemener: OVERALS) lost het eerste probleem (het meet-probleem) op, HOMALS lost het join-probleem op. Of preciezer: principale komponenten analyse probeert een basis te vinden voor de join van de $\underline{h}_j$. In termen van verliesfunkties: we willen een basis vinden van zo klein mogelijke dimensie waarin een zo groot mogelijk deel van de $\underline{h}_j$ past. Kanonische analyse probeert een basis te vinden voor de meet van de $\underline{h}_j$: we willen een basis

vinden van zo groot mogelijke dimensie die zo goed mogelijk in alle $\underline{h}_j$ past. Deze formulering wijst op een fundamenteel verschil tussen HOMALS/PRINCALS en CANALS/OVERALS. In CANALS/OVERALS zoeken we naar zoveel mogelijk dimensies. Het is inderdaad dikwijls zo dat de eerste dimensies van CANALS/OVERALS triviale oplossingen geven die kapitaliseren op toevallige aspekten van de gegevens. In HOMALS/PRINCALS zou dat vervelend zijn, omdat we daar laag-dimensionele oplossingen zoekn, maar in CANALS/OVERALS is het theoretisch gezien juist prettig omdat we zoveel mogelijk dimensies in de meet willen vinden. Er is dus duidelijk een zekere diskrepantie tussen kanonische analyse zoals het in de praktijk toegepast wordt (onder andere in deze klapper), en zoals het uit de join-meet formulering naar voren komt (zo veel mogelijk dimensies).

We gaan hier niet in detail in op het probleem hoe we join-loss en meet-loss definieren. Het betoog verloopt ongeveer analoog aan dat in 5.2.1. De verliesfunktie σ_1 die daar gedfinieerd wordt is de meest natuurlijke kandidaat voor join-loss, de verliesfunktie σ_2 de meest natuurlijke kandidaat voor meet-loss. Zoals aangetoond in 5.2.1 zijn join-problemen een bizonder soort meet-problemen, of, anders gezegd, kunnen we altijd de verliesfunktie σ_2 gebruiken. Dus: ook PCA, een join-probleem, kan opgelost worden door σ_2, meet-loss, te minimaliseren. De HOMALS-verliesfunktie, en daardoor de PRINCALS-verliesfunktie, is dus het speciale geval van meet-loss waarin de $\underline{h}_j$ indikator-vektoren zijn. De CANALS/OVERALS verliesfunktie is meer algemeen (vergelijk het eerste deel van 7.5). Over het minimaliseren van meet-loss in diverse omstandigheden kunnen we hier ook kort zijn, om de eenvoudige reden dat vrijwel de gehele klapper daar over gaat. Alleen is het nuttig om nogmaals de nadruk te leggen op het feit dat we in CANALS/OVERALS <u>zo veel mogelijk</u> dimensies willen vinden (als voor ons gevoel verantwoord is), terwijl we in HOMALS <u>zo weinig mogelijk</u> dimensies willen vinden (als voor ons zelfde gevoel verantwoord is).

8.5: Over optimale schaling

Op verschillende plaatsen in deze klapper is al aan de orde gekomen dat er verschillende schaalnivo's zijn, dat zo'n schaalnivo beperkingen oplegt aan de kwantifikaties van de kategorieen van de desbetreffende variabele, en dat we dit soort kwantifikatieproblemen oplossen door iets op een kegel te projekteren. In de volgende paragrafen willen we dit alles wat specifieker behandelen. We beperken ons in eerste instantie tot enkelvoudige kwantifikatie, dat wil dus zeggen dat we het meerdimensionale kwantifikatieprobleem wat de variabelen betreft tot één dimensie terugbrengen door op één enkele richting in de meerdimensionale ruimte te projekteren. Meervoudige kwantifikatie komt in het kort aan het eind van deze paragrafen nog ter sprake.

8.5.1: Meetnivo en procesnivo

In De Leeuw, Young, en Takane (1976) wordt een eksplisiet onderscheid gemaakt tussen meetnivo en procesnivo. Voor meetnivo gebruiken we het bekende onderscheid tussen numerieke, ordinale, en nominale variabelen. Voor procesnivo onderscheiden we, in navolging van De Leeuw, Young, en Takane, diskrete en kontinue processen, en in navolging van De Leeuw (1978) ook nog het zwaartepuntsproces. Het verschil tussen meetnivo en procesnivo komt voort uit het feit dat we zowel individuen als kategorieen van variabelen op een schaal willen afbeelden. De eisen die we stellen aan de posities van de kategoriekwantifikaties volgen uit onze aannamen over het meetnivo, de eisen die we stellen aan de relatie tussen kategoriekwantifikaties en individukwantifikaties volgen uit onze aannamen over het procesnivo. Praten over procesnivo heeft alleen maar zin bij kategorische variabelen, waarbij het aantal individuen behoorlijk veel groter is dan het aantal kategorieen, dat wil dus zeggen in situaties waarin we bij voorkeur gebruik maken van een indikator matriks. Overigens moeten we één misverstand bij voorbaat uit de weg ruimen. In sommige psychologische literatuur wordt er over meetnivo gepraat als of het een onvervreembare eigenschap van een variabele is, die daardoor bijvoorbeeld ook allerlei methodologische verbodsbepalingen tot gevolg heeft. Zoals onder meer Lord (1953) duidelijk aangetoond heeft is dit een onhoudbaar soort essentialisme. Het meetnivo is, evenals de dimensionaliteit van de faktorruimte of de graad van de interpolatiepolynoom of de vorm van de itemkarakteristiek, een hypotese die onderzocht moet worden. Natuurlijk verwerken we in deze hypothese de kennis die we hebben over de manier waarop de gegevens tot stand gekomen zijn, en natuurlijk dikteert de hypothese (samen met allerlei andere aannamen) de aard van de techniek die we gaan toepassen. Maar een hypotese is van belang omdat er alternatieve hypothesen bestaan, en daarom kunnen aannamen over meetnivo nooit tot methodologische verbodsbepalingen leiden.

8.5.2: Ordinale gegevens

Het lijkt misschien wat vreemd om met het ordinale meetnivo te beginnen, maar er zijn historische redenen voor. De niet-metrische "rekenkundige doorbraak" in de psychometrie wordt over het algemeen verbonden met de naam van Shepard (1962). Dat lijkt ons niet helemaal juist, dat artikel was ongetwijfeld van groot belang, het toonde de mogelijkheid van een "rekenkundige doorbraak", maar de "rekenkundige doorbraak" zelf was Kruskal (1964a, b). Kruskal introduceerde drie fundamentele innovaties: ten eerste de nadruk op een eksplisiet gedefinieerde en eksplisiet te minimaliseren verliesfunktie, ten tweede het gebruik van gradient methoden, en ten derde de algorithmische benadering van schaalnivo's. Deze laatste bijdrage was misschien wel het belangrijkste, en vanwege de kontekst waarin Kruskal schreef lag daarbij duidelijk de nadruk op ordinale gegevens. Kruskal introduceerde monotone regressie in de psychometrie, en gaf een eerste voorbeeld van het belang van procesnivo met zijn onderscheid tussen de primaire en sekundaire behandeling van gelijken. De latere bijdragen van De Leeuw, Young, en Takane zijn er voornamenlijk op gericht deze theorie aan te vullen en te kompleteren.

Stel G is een n × k indikator matriks, we zoeken een kwantifikatie y van de individuen en een kwantifikatie z van de kategorieen. De vektor y heeft dus n elementen, z heeft k elementen. We kiezen y en z op zon'n manier dat aan de meet- en procesvoorwaarden voldaan wordt, en dat bovendien $SSQ(x - y)$ zo klein mogelijk is, waarbij x een gegeven n-vektor is. In de PRINCALS kontekst bevat x de projekties van de individuen op de richting behorende bij de variabele. De <u>diskreet ordinale</u> voorwaarden zijn

$z_1 \leq \ldots \leq z_k$,

$y = Gz$.

De kategoriekwantifikaties moeten dus in de 'goede' volgorde staan, en individuen krijgen dezelfde kwantifikatie als de kategorie waar ze in vallen. Dus individuen in dezelfde kategorie hebben dezelfde kwantifikatie, Kruskal noemt dit de <u>sekundaire benadering van gelijken</u>. Definieer $D = G'G$, als gewoonlijk, en $\bar{z} = D^{-1}G'x$. We schrijven $y = G\bar{z} + G(z - \bar{z})$, en vinden

$$SSQ(x - y) = SSQ(x - G\bar{z}) + (z - \bar{z})'D(z - \bar{z}).$$

Het probleem wordt nu om de laatste term, die we ook wel schrijven als $SSQ_D(z - \bar{z})$, te minimaliseren over z die voldoet aan $z_1 \leq \ldots \leq z_k$, een aanzienlijk kleiner probleem dan we eerst hadden. We komen verderop terug op de oplossing, maar we behandelen eerst wat alternatieve procesnivo's.

De <u>zwaartepunts ordinale</u> voorwaarden zijn

$z_1 \leq \ldots \leq z_k$,

$z = D^{-1}G'y$.

We gebruiken nu de partitionering $x - y = G(\bar{z} - z) + (x - G\bar{z}) - (y - Gz)$. Dit

geeft

$$SSQ(x - y) = SSQ_D(z - \bar{z}) + SSQ\{(x - G\bar{z}) - (y - Gz)\} .$$

De tweede term kunnen we nul maken door y te kiezen als $\hat{y} = x - G(\bar{z} - \hat{z})$, de eerste term minimaliseren we door z te kiezen op precies dezelfde manier als bij ordinaal diskreet. De twee procesnivo's geven dus een zelfde $\hat{z}$, maar een verschillende y. Voor ordinaal diskreet geldt $\hat{y} = G\hat{z}$, voor zwaartepunts ordinaal geldt $\hat{y} = x - G(\bar{z} - \hat{z})$. In de laatste formules gebruiken we symbolen met dakjes voor de oplossingen van de regressieproblemen.

Blijft over <u>kontinu ordinaal</u>. Daarbij is het gemakkelijk twee vektoren z_+ en z_-, beide van lengte k, te gebruiken. We eisen

$$z_1^- \leq z_1^+ \leq \ldots \leq z_k^- \leq z_k^+,$$

$$Gz_- \leq y \leq Gz_+,$$

waarbij de laatste ongelijkheid tussen n-vektoren elementsgewijs opgevat moet worden. We kunnen de eisen van kontinu ordinaal interpreteren in termen van intervallen (z_r^-, z_r^+). Deze intervallen moeten op de juiste wijze langs de reele lijn geordend liggen, en de kwantifikatie van het individu moet liggen in het interval behorend bij de kategorie waar hij in valt. Het probleem met kontinu ordinaal is dat deze eisen een partiele ordening over de y_i definieren, en om algoritmische redenen zouden we graag een komplete ordening hebben. Gelukkig is het eenvoudig een komplete ordening te vinden waaraan de oplossing $\hat{y}$ moet voldoen. In De Leeuw (1978) wordt bewezen dat als $y_i \underline{\simeq} y_j$, dat wil zeggen als y_i en y_j in dezelfde kategorie vallen, en als $x_i \leq x_j$, dat dan ook $\hat{y}_i \leq \hat{y}_j$. Met behulp van deze regel is het gemakkelijk de komplete orde te definieren waaraan $\hat{y}$ moet voldoen. We ordenen de kategorieen als voorgeschreven, en binnen de kategorieen ordenen we volgens de data x_i.

We hebben nu het regressieprobleem voor alle drie de vormen van procesnivo's gereduceerd tot het minimaliseren van een funktie van de vorm $SSQ_W(x - y)$ over alle zwak geordende y, met W een diagonale matriks met niet-negatieve gewichten. Dit is een heel speciaal kwadratisch programmeerprobleem dat monotone of isotone regressie genoemd wordt. De meest gebruikelijke metoden om het probleem op te lossen zijn gebaseerd op de volgende regel. We nemen aan dat de eis is $y_1 \leq \ldots \leq y_n$. De regel is: als $x_i > x_{i+1}$ dan $\hat{y}_i = \hat{y}_{i+1}$. Deze regel definieert in feite een efficient algoritme: we lopen de data vektor door tot we een foute ordening vinden, die definieert een gelijkheid in de oplossing, en we gebruiken die gelijkheid om het probleem te reduceren tot een probleem met vektoren van de lengte n - 1. We herhalen dit, en vinden of geen foutjes, of aan het eind een probleem van de orde n - 2. Enzovoorts. In het meest ongunstige geval hebben op laatst een probleem van de orde één, wat betekent dat alle elementen van $\hat{y}$ gelijk zijn.

Bij diskrete en zwaartepuntsregressie is dit zeer efficient: we vormen de gemiddelden $\overline{z} = D^{-1}G'x$, en doen monotone regressie op deze k-vektor. Bij kontinue regressie moeten we monotone regressie doen op een vektor van lengte n, en dat kan een aanzienlijk onaangenamer probleem zijn. Zoals eerder aangeduid kan n bij HOMALS, CANALS, en PRINCALS in de duizenden lopen. Dit is één van de belangrijkste redenen waarom in de huidige versies van deze programmaas alleen nog maar de diskrete opties opgenomen zijn. Een andere belangrijke reden heeft te maken met de sterkte van de eisen. Het is duidelijk dat de diskrete eisen sterker zijn dan de kontinue, en dat de kontinue eisen weer sterker zijn dan de zwaartepuntseisen. Zelfs met diskrete eisen, met name in CANALS, zien we echter soms dat ordinale variabelen door de programmaas 'gedegenereerd' worden, dat wil zeggen dat bijvoorbeeld één individu in oneindig gelegd wordt, terwijl alle andere individuen een gelijke schaalwaarde krijgen. Bij zwakkere eisen zijn er veel meer mogelijkheden om gedegenereerde transformaties te vinden, die weliswaar optimaal zijn in de zin van het kleinste kwadratenverlies, maar die data analytisch weinig interessant, en statistisch weinig stabiel zijn.

8.5.3: Nominale gegevens

Gegeven onze bespreking van ordinale gegevens kunnen we kort zijn over nominale. De nominale voorwaarden ontstaan uit de ordinale door de restrikties op de kategorie kwantifikaties te laten vallen. Dus: nominale gegevens hebben alleen procesnivo restrikties, geen meetnivo restrikties. Diskreet nominaal wordt dus bijvoorbeeld gedefinieerd als $y = Gz$, zonder enige restriktie op z. Op dezelfde manier wordt zwaartepunts nominaal $z = D^{-1}G'y$, maar omdat er geen restrikties zijn op z betekent dit dat zwaartepunts nominaal helemaal geen restrikties oplegt. Kontinu nominaal tenslotte eist $Gz_- \leq y \leq Gz_+$, tezamen met $z_- \leq z_+$. De y_i moeten dus in hun kategorie intervallen liggen, de intervallen mogen niet overlappen, maar de volgorde van de intervallen op de reele lijn ligt niet vast. Het regressie probleem voor diskreet nominaal heeft vanzelfsprekend de oplossing $\hat{z} = D^{-1}G'x$ en $\hat{y} = G\hat{z}$. Er is geen zwaartepunts regressie probleem, de oplossing is triviaal $\hat{y} = x$. Voor kontinu nominaal is de situatie wat onplezierig. We kunnen kontinue nominaal opvatten als kontinu ordinaal met een onbekende volgorde van kategorieen. Het regressie probleem is dan om die volgorde van kategorieen te kiezen waarvoor het kontinu ordinale algoritme de beste fit geeft. Een naar probleem, zowel wat rekenwerk betreft als wat eigenschappen van de oplossing betreft. We zijn momenteel bezig met het ontwerpen van betere metoden om zowel kontinu nominaal als kontinu ordinaal te fitten, maar we doen daar hier nog geen verdere mededelingen over.

8.5.4 Numerieke gegevens

Numerieke gegevens zijn er in vele soorten en maten. We beperken ons hier tot

de intervalschaal, die zeker voor onze doeleinden het belangrijkste is. Er is nu een numerieke k-vektor v gegeven, en de eisen voor diskreet interval zijn

$z = \alpha v$,

$y = Gz$.

We veronderstellen hierbij dat $u'Dv = 0$, zodat ook $u'y = 0$ (u is de vektor met alle elementen gelijk aan één). Zwaartepunts interval is

$z = \alpha v$,

$z = D^{-1}G'y$.

Voor kontinu interval zijn de eisen

$z_- = \alpha v_-$,

$z_+ = \beta v_+$,

$Gz_- \leq y \leq Gz_+$.

Om deze eisen beter te kunnen relateren aan zwaartepunts interval enerzijds en kontinu ordinaal anderszijds merken we op dat we in kontinu ordinaal zonder verlies van algemeenheid mogen eisen dat $z_r^+ = z_{r+1}^-$ voor alle $r=1,\ldots,k-1$. En bovendien kunnen we eisen $z_1^- = -\infty$ en $z_k^+ = +\infty$. De intervallen delen dan de reele lijn op in k stukken. We krijgen een zelfde situatie bij kontinu interval als we eisen dat $\alpha = \beta$, en $v_r^+ = v_{r+1}^-$, en $v_1^- = -\infty$, en $v_k^+ = +\infty$.

Het regressieprobleem voor diskreet interval is het minimaliseren van $SSQ(x - \alpha Gv)$ over α, en iedereen weet wel hoe dat moet. Voor zwaartepunts interval gebruiken we dezelfde partitionering als bij zwaartepunts ordinaal. We vinden dan

$SSQ(x - y) = SSQ_D(\alpha v - \bar{z}) + SSQ\{(x - G\bar{z}) - (y - Gz)\}$.

Eerst vinden we dus α door de eerste term te minimaliseren, dan berekenen we $\hat{z} = \hat{\alpha} v$, en dan tenslotte $\hat{y} = x - G(\bar{z} - \hat{z})$, net als in het ordinale geval. Zoals gewoonlijk is kontinu interval wat lastiger, maar we hebben een relatief eenvoudig en efficient algoritme ontwikkeld. We bespreken het hier niet, omdat onze programmaas zoals vermeld nog geen kontinue opties hebben.

8.5.5 Ontbrekende gegevens

Ontbrekende gegevens worden door ons behandeld als individuen met een unieke kategorie. Aan de kategoriekwantifikaties worden bovendien geen ordinale of numerieke eisen gesteld. Hieruit volgt dat het regressieprobleem met ontbrekende gegevens opgelost kan worden door regressie over de niet-ontbrekende gegevens uit te voeren, en het stuk van x korresponderend met de ontbrekende gegevens vervolgens in y te kopieren. Dat is dus nogal eenvoudig.

8.5.6 Meervoudige kwantifikatie

Zoals uiteengezet in het geometrische stuk over schaalnivo's, elders in dit hoofdstuk, kunnen we meervoudig diskreet nominaal beschrijven door een variabele met k kategorieen op te delen in k variabelen met twee kategorieen, en door deze k variabelen diskreet nominaal te behandelen. Meervoudig kontinu nominaal komt overeen met opdelen van een variabele met k kategorieen

in $\binom{k}{2}$ binaire variabelen, die vervolgens allemaal kontinu ordinaal behandeld worden. De details vindt men elders in dit hoofdstuk. Het is overigens ongebruikelijk om in het meervoudige geval nog van schalen te spreken, de geometrische interpretatie is hier veel overheersender. Overigens zijn er eerder in de literatuur (bijvoorbeeld door Coombs en Kao) al meervoudig ordinale modellen voorgesteld. In hoofdstuk 10 wordt op dit soort meervoudige modellen nader ingegaan.

9. Bootstrap en Jackknife.

9.1. Theorie.

9.1.1. Inleiding.

De resultaten van multivariate analyse technieken worden inzichtelijker, wanneer we iets over de spreiding van de resultaten weten. Waar analytiese methoden (nog) niet bestaan, of te gekompliceerd zijn, bieden simulatiemethoden uitkomst. Om redenen die verderop in dit onderdeel aan de orde komen, gebruiken we hiervoor de bootstrap-methode (Efron, 1979). Wegens de overeenkomst van de bootstrap met de Quenouille-Tukey-jackknife (Miller, 1974) worden beide hier besproken.

Voor we de methoden data-analyties en statisties beschrijven, vergelijken we ze aan de hand van een voorbeeld. We hebben een reeks getallen afkomstig uit een onbekende verdeling:

0.530 2.059 0.244 1.949 0.487 1.498 1.077 1.686 1.354 0.245

Het gemiddelde is 1.0652 en de variantie $\hat{\sigma}^2=0.5701$. Om iets te kunnen zeggen over de variantie van de populatie, doen we het volgende: we laten uit de data telkens een groepje van twee weg, en berekenen de variantie van de overige getallen. We vinden:

	gemiddelde	variantie	pseudo-waarde
nr. 1 en 2 weg	1.0675	$\hat{\sigma}_1^2=0.4455$	$\tilde{\sigma}_1^2=1.0685$
nr. 3 en 4 weg	1.0574	$\hat{\sigma}_2^2=0.5250$	$\tilde{\sigma}_2^2=0.7505$
nr. 5 en 6 weg	1.0834	$\hat{\sigma}_3^2=0.6581$	$\tilde{\sigma}_3^2=0.2181$
nr. 7 en 8 weg	0.9861	$\hat{\sigma}_4^2=0.6708$	$\tilde{\sigma}_4^2=0.1673$
nr. 9 en 10weg	1.1316	$\hat{\sigma}_5^2=0.6199$	$\tilde{\sigma}_5^2=0.3709$

Hoe we de 'pseudo-waarden' vinden komt in 9.1.2 aan de orde.

We bepalen nu het gemiddelde van de pseudo-waarden, dit is .5151, en de variantie ervan, .1479. Volgens de jackknife-theorie ligt nu de populatie parameter σ^2 met 95% betrouwbaarheid in het interval 0.5151 ± 0.478.

Een bootstrap kan als volgt gaan: we trekken vijf maal een rijtje van tien getallen uit de getallen 1 t/m 10, we nemen de hierbij behorende getallen en bepalen de variantie:

trekking	gemiddelde	variantie
2,3,5,2,10,8,5,9,8,4	1.2256	$\hat{\sigma}_1^2$=0.5966
9,9,2,6,9,6,7,1,1,6	1.1798	$\hat{\sigma}_2^2$=0.4131
10,5,9,2,9,2,9,6,5,7	1.1974	$\hat{\sigma}_3^2$=0.3977
2,3,8,3,10,5,2,3,2,2	1.1386	$\hat{\sigma}_4^2$=0.8125
10,8,9,8,9,3,4,3,9,4	1.2065	$\hat{\sigma}_5^2$=0.4894

Het gemiddelde van de $\hat{\sigma}_j^2$ is 0.5419 en de variantie 0.0291. Volgens de bootstrap-theorie ligt nu de populatie parameter σ^2 met 95% betrouwbaarheid in het interval 0.5419 ± 0.474.

9.1.2. Data analytiese beschrijving van de methoden.

We gaan uit van een data set van n observaties, en een willekeurige analyse techniek. De jackknife werkt als volgt:

- verdeel de waarnemingen in s subsets van gelijke grootte (we nemen aan dat s een deler is van n);
- pas de analyse techniek s maal toe en wel telkens op de data waaruit één subset is weggelaten;
- als $\hat{\theta}$ geschat is uit de oorspronkelijke analyse, en $\hat{\theta}_j$ bij de j-de herhaling, noemen we $\tilde{\theta}_j = s\,\hat{\theta} - (s-1)\hat{\theta}_j$ de j-de pseudo-waarde;
- het gemiddelde van de pseudo-waarden $\tilde{\theta} = 1/s\ \Sigma_j^s\ \tilde{\theta}_j$ is een schatter voor de populatie parameter θ en

$$\frac{\tilde{\theta} - \theta}{\{\frac{1}{s(s-1)}\ \Sigma_j^s\ (\tilde{\theta}_j - \tilde{\theta})^2\}^{\frac{1}{2}}}$$

heeft bij benadering een t-verdeling met s-1 vrijheidsgraden (Miller, 1974).

De bootstrap methode bestaat uit de volgende stappen:

- ken aan elk van de observaties kans 1/n toe en trek een steekproef ter grootte n (met teruglegging) uit de data. Dit levert een nieuwe data set;
- pas de analyse techniek toe op deze nieuwe data set;
- herhaal deze procedure een aantal malen, zeg s maal;
- als $\hat{\theta}$ geschat is uit de oorspronkelijke analyse, en $\hat{\theta}_j$ bij de j-de bootstrap, dan is $\hat{\theta}_j$ een schatter van $\hat{\theta}$, en dit is een schatter van een populatie parameter θ, dus $\tilde{\theta} \triangleq 1/s\ \Sigma_j^s\ \hat{\theta}_j$ is

ook een schatter van θ

$$\frac{\hat{\theta} - \theta}{\{ 1/s \ \Sigma_j^s \ (\hat{\theta}_j - \hat{\theta})^2 \}^{\frac{1}{2}}}$$

heeft bij benadering een t-verdeling met s-1 vrijheidsgraden.

9.1.3. Statistiese beschrijving van de methoden.

Als we ons beperken tot analyse technieken van kategoriese gegevens, dan bestaan er slechts eindig veel mogelijke profielen. Als er q profielen mogelijk zijn, en onze data zijn afkomstig van trekkingen uit een zekere populatie, dan kunnen we een kansveld definiëren: π_i is de kans dat bij een random trekking uit de populatie we een individu tegenkomen dat profiel i heeft (i=1,...,q). We doen nu een steekproef ter grootte n met onafhankelijke trekkingen en vinden $\underline{n}_i$ individuen die profiel i hebben. Als $\underline{x}_i \triangleq \underline{n}_i/n$, dan is $\underline{x}_i$ een schatter van π_i.

Nu de analyse techniek. Die gaat uit van n gegevens, steeds te koderen als een $x \in \mathbb{R}^q$, en resulteert in een aantal uitkomsten. Elk van deze uitkomsten (eigenwaarde, stress, skore in een bepaalde dimensie etc.) is een funktie van de gegevens, oftewel we kunnen in plaats van over analyse uitkomsten praten over funkties $f(x)$, $x \in \mathbb{R}^q$. We gaan nu één zo'n f bekijken, en we willen nu weten hoe goed $f(\underline{x})$ $f(\underline{\pi})$ schat. We nemen aan dat f in een omgeving van π kontinu differentieerbaar is. Deze aanname is voor enkele gevallen bewezen, en lijkt ook voor de overige aannemelijk.

$$f(x) = f(\pi) + (x - \pi)' f_\pi + o(x-\pi) \qquad (x \to \pi)$$

We noteren hier f_π voor de afgeleide van f naar x in het punt π, omdat het aksent al als transpositie teken gebruikt wordt. Voor de steekproef schatter $\underline{x}$ geldt

$$E(\underline{x}) = \pi$$
$$V(\underline{x}) = 1/n \ (\Pi - \pi\pi')$$

waarin $\Pi \triangleq \text{diag}(\pi_1,...,\pi_q)$ en $\pi \triangleq \Pi u$.
Voor grote n nadert $\underline{x}$ in waarschijnlijkheid tot π, en is dus

$$f(\underline{x}) = f(\pi) + (\underline{x} - \pi)' f_\pi + o(\underline{x}-\pi)$$

De verwachting hiervan is

$$E(f(\underline{x})) = f(\pi) + E(\underline{x} - \pi)'f_\pi + Eo(\underline{x}-\pi)$$
$$= f(\pi) + Eo(\underline{x}-\pi) \to f(\pi) \qquad (n\to\infty),$$

dus $f(\underline{x})$ is een konsistente schatter van $f(\pi)$. Wat is de variantie van $f(\underline{x})$?

$$V(f(\underline{x})) = E(f(\underline{x}) - f(\pi))^2 = E((\underline{x} - \pi)'f_\pi + o(\underline{x}-\pi))^2$$
$$\approx E((\underline{x} - \pi)'f_\pi)^2 = f_\pi'V(\underline{x})f_\pi = 1/n\ f_\pi'(\Pi - \pi\pi')f_\pi$$

Dit alles ter inleiding op bootstrap en jackknife. Het moge duidelijk zijn dat bootstrap en jackknife (beide op hun eigen manier) nieuwe random vektoren $\underline{y}\in\mathbb{R}^q$ gebruiken die door $\underline{x}$ bepaald worden, en ook (voor grote n) dicht bij $\underline{x}$ liggen. Hoe deze vektoren $\underline{y}$ bepaald worden, bekijken we verderop; eerst gaan we $f(\underline{y})$ onderzoeken.

$$f(\underline{y}) = f(\pi) + (\underline{y} - \pi)'f_\pi + o(\underline{y}-\pi) \quad (\underline{y}\to\pi)$$

Als we de o-termen weglaten moet nu bij benadering gelden

$$f(\underline{y}) = f(\underline{x}) + (\underline{y} - \underline{x})'f_\pi$$

Dit volgt eenvoudig als we de twee vergelijkingen van $f(\underline{y})$ en $f(\underline{x})$ van elkaar aftrekken. Ook geldt

$$f(\underline{y}|\underline{x}) = f(\underline{x}) + ((\underline{y}|\underline{x}) - \underline{x})'f_\pi$$

De verwachting en variantie hiervan zijn, als $E(\underline{y}|\underline{x})=\underline{x}$,

$$E(f(\underline{y}|\underline{x})) = f(\underline{x})$$
$$V(f(\underline{y}|\underline{x})) = f_\pi'E((\underline{y}|\underline{x}) - \underline{x})((\underline{y}|\underline{x}) - \underline{x})'f_\pi = f_\pi'V(\underline{y}|\underline{x})f_\pi$$

Om nu verder te kunnen, moeten we weten hoe $\underline{y}$ van $\underline{x}$ afhangt, en dat wordt bepaald door bootstrap of jackknife. We vertalen de recepten uit 9.1.2 in de nu gebruikte terminologie.

Jackknife: neem aan dat s een deler is van n. Verdeel de observaties random in s groepen van gelijke grootte. Voor groep j bepalen we $\underline{x}_j$ als volgt: het i^e element van $\underline{x}_j$ is de proportie observaties in groep j dat profiel i skoort (i=1,...,q). Merk op dat $\underline{x}= 1/s\ \Sigma_j^s\ \underline{x}_j$ (de proportie observaties van profiel i is het gemiddelde van de groepsproporties van profiel i, voor elke i).

We moeten steeds een groep weglaten. Dit betekent dat we $\underline{x}^j$ moe-

ten maken zodanig dat het i^e element van $\underline{x}^j$ gelijk is aan de proportie observaties in het komplement van groep j dat profiel i skoort (i=1,...,q). Maar dat is juist het gemiddelde over alle groepen behalve j:

$$\underline{x}^j = \frac{1}{s-1} \sum_{\ell \neq j} \underline{x}_\ell$$

Dit kunnen we ook schrijven als

$$\underline{x}^j = \frac{1}{s-1} \left(\sum_{\ell=1}^{s} \underline{x}_\ell - \underline{x}_j \right) = \frac{1}{s-1} (s\underline{x} - \underline{x}_j)$$

Op $\underline{x}^j$ moeten we nu de analyse techniek toepassen; dat komt overeen met het bepalen van $f(\underline{x}^j)$. Hiervan weten we

$$f(\underline{x}^j|\underline{x}) = f(\underline{x}) + ((\underline{x}^j|\underline{x}) - \underline{x})' f_\pi$$

en omdat

$$(\underline{x}^j|\underline{x}) - \underline{x} = \frac{1}{s-1} (s\underline{x} - \underline{x}_j) - \underline{x} = \frac{1}{s-1} (\underline{x} - \underline{x}_j)$$

is

$$f(\underline{x}^j|\underline{x}) = f(\underline{x}) + \frac{1}{s-1} (\underline{x} - (\underline{x}_j|\underline{x}))' f_\pi$$

Nu moeten we de j^e <u>pseudo-waarde</u> $\underline{\vartheta}_j$ bepalen:

$$\begin{aligned} \underline{\vartheta}_j &= sf(\underline{x}) - (s-1) f(\underline{x}^{-j}|\underline{x}) \\ &= sf(\underline{x}) - (s-1) f(\underline{x}) - (\underline{x} - (\underline{x}_j|\underline{x}))' f_\pi \\ &= f(\underline{x}) - (\underline{x} - (\underline{x}_j|\underline{x}))' f_\pi \end{aligned}$$

Het gemiddelde van de pseudo-waarden is een schatter voor de populatie parameter $f(\pi)$:

$$\begin{aligned} \underline{\vartheta} &= 1/s \sum_{j=1}^{s} \underline{\vartheta}_j = f(\underline{x}) - 1/s \sum_{j=1}^{s} (\underline{x} - (\underline{x}_j|\underline{x}))' f_\pi \\ &= f(\underline{x}) + 1/s\, f_\pi' \sum_{j=1}^{s} ((\underline{x}_j|\underline{x}) - \underline{x}), \end{aligned}$$

en hieraan zien we dat $\underline{\vartheta}$ inderdaad een schatter van $f(\pi)$ is. Voor het bewijs dat

$$\frac{\underline{\vartheta} - f(\pi)}{\left\{ \frac{1}{s(s-1)} \sum_j^s (\underline{\vartheta}_j - \vartheta)^2 \right\}^{\frac{1}{2}}}$$

asymptoties een t-verdeling met s-1 vrijheidsgraden volgt, ver-

wijzen we naar Miller(1974).

Bootstrap: $\underline{y}$ wordt nu bepaald door aan profiel i kans $\underline{x}_i$ toe te kennen (i=1,...,q) (let op: $\underline{x}_i$ is nu een getal, terwijl zojuist $\underline{x}_j$ een vektor was), en een steekproef te doen ter grootte n uit de profielen; $\underline{y}_i$ is de proportie observaties van profiel i uit deze 'bootstrap'-steekproef. Nu is

$$E(\underline{y}|\underline{x}) = \underline{x} \qquad \text{en} \qquad V(\underline{y}|\underline{x}) = 1/n\ (\underline{x}^D - \underline{x}\underline{x}')$$

Hierin is $\underline{x}^D$ de diagonaalmatrix waarvoor geldt $\underline{x}^D u=\underline{x}$. Het toepassen van de analyse op $\underline{y}$ betekent het bepalen van $f(\underline{y}|\underline{x})$. Hiervoor geldt

$$E(f(\underline{y}|\underline{x})) = f(\underline{x}) \text{ en} \qquad V(f(\underline{y}|\underline{x}) = f_\pi' V(\underline{y}|\underline{x}) f_\pi = \frac{1}{n} f_\pi' (\underline{x}^D - \underline{x}\underline{x}') f_\pi$$

Efron (1979) bewijst dat $f(\underline{y}|\underline{x})$ ssymptoties normaal verdeeld is voor $n\to\infty$. Voor eindige n kunnen we een aantal schatters $f(\underline{y}^1),\ldots,$ $f(\underline{y}^s)$ bepalen en in plaats van normaliteit een t-verdeling hanteren, zoals in 9.1.2 aangegeven.

9.1.4. Welke van de twee?

De technieken worden vergeleken voor gelijke s, omdat beide dan even 'duur' zijn. Bij de vergelijking gaat de tweede orde term, die we in de vorige paragraaf als $o(\underline{y}-\pi)$ weglieten, een rol spelen. Om niet te verdwalen in formule-gegoochel, zullen we de jackknife en de bootstrap wat globaler vergelijken.

Zoals uit 9.1.3 blijkt, is de verdeling van een bootstrap-schatter $\underline{y}|\underline{x}$ analoog aan de verdeling van $\underline{x}$, en dat is $\underline{x}|\pi$. Voor jackknife-schatters $\underline{x}^j|\underline{x}$ geldt dit niet; de variantie van $\underline{x}^j|\underline{x}$ is (globaal) 1/(s-1) maal de variantie van $\underline{x}|\pi$.

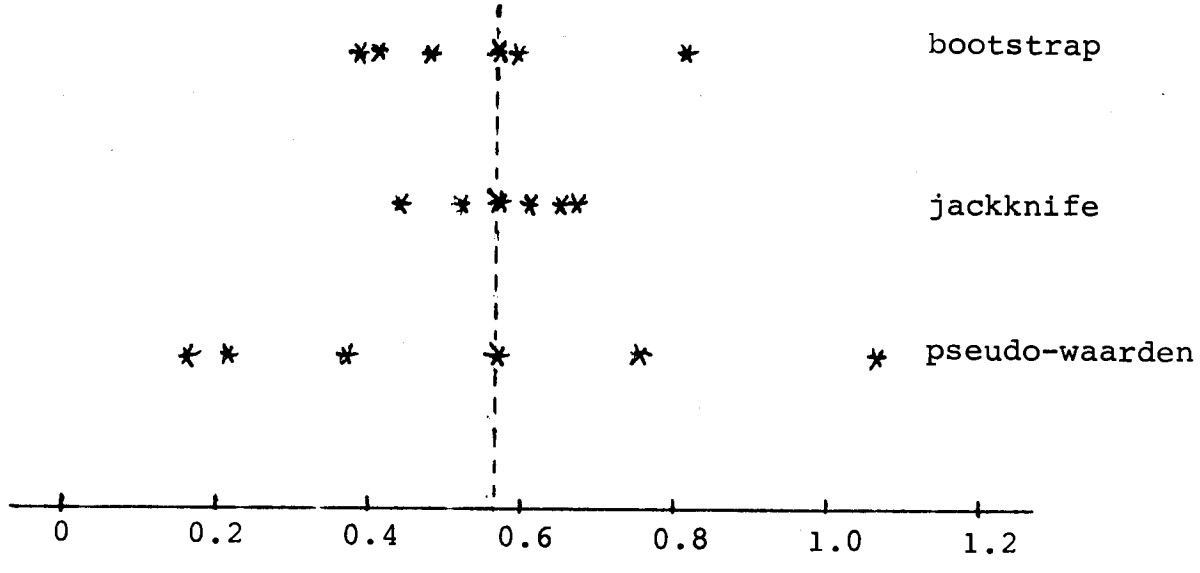
Stellen we ons nu voor, dat de s bootstrap-schatters $\underline{y}^1,\ldots,\underline{y}^s$ in een wolk rond $\underline{x}$ liggen, evenals de jackknife-schatters $\underline{x}^1,\ldots,$ $\underline{x}^s$, dan is de laatste wolk kleiner dan de eerste. Hoe groter s is, des te groter dit verschil. Het gevolg hiervan is, dat het weglaten van de tweede orde term, wat we in de vorige paragraaf deden, voor de bootstrap ernstiger gevolgen heeft dan voor de jackknife. We kunnen konkluderen dat de jackknife een betere schatter van $V(f(\underline{y}|\underline{x}))$ geeft.

Maar geldt dit nu ook voor $V(f(\underline{x}|\pi)$? Het verschil tussen deze varianties wordt nu net door de tweede orde term bepaald! We

kunnen het ook anders zeggen: met de jackknife wordt de variantie nauwkeuriger geschat dan hij te bepalen is (het is als het afdrukken van een rekenresultaat in meer decimalen dan er betrouwbaar zijn). Het voordeel van de jackknife ten opzichte van de bootstrap is dus maar schijn.

Kijken we nu naar een tweede gevolg van het feit, dat de bootstrap wolk groter is dan de jackknife wolk, en dat de eerste bij benadering gelijk is aan de $\underline{x}$-wolk om π (die we niet kennen, omdat we maar één observatie van $\underline{x}$ hebben). Hierdoor liggen de punten $f(\underline{y}^j|\underline{x})$ net zo ten opzichte van het punt $f(\underline{x})$ als meerdere punten $f(\underline{x})$ (als we die kenden) zouden liggen ten opzichte van $f(\pi)$. Doordat de jackknife wolk veel kompakter om $\underline{x}$ ligt, geldt deze prettige eigenschap niet bij deze techniek, en zeker niet bij de pseudo-waarden, welke slechts transformaties zijn van de $f(\underline{x}^j)$.

We hebben de verdeling van de bootstrap-, jackknife- en pseudowaarden voor ons voorbeeld getekend in figuur 9.1 (hierbij is de oorspronkelijke observatie van σ^2 met een stippellijn aangegeven).



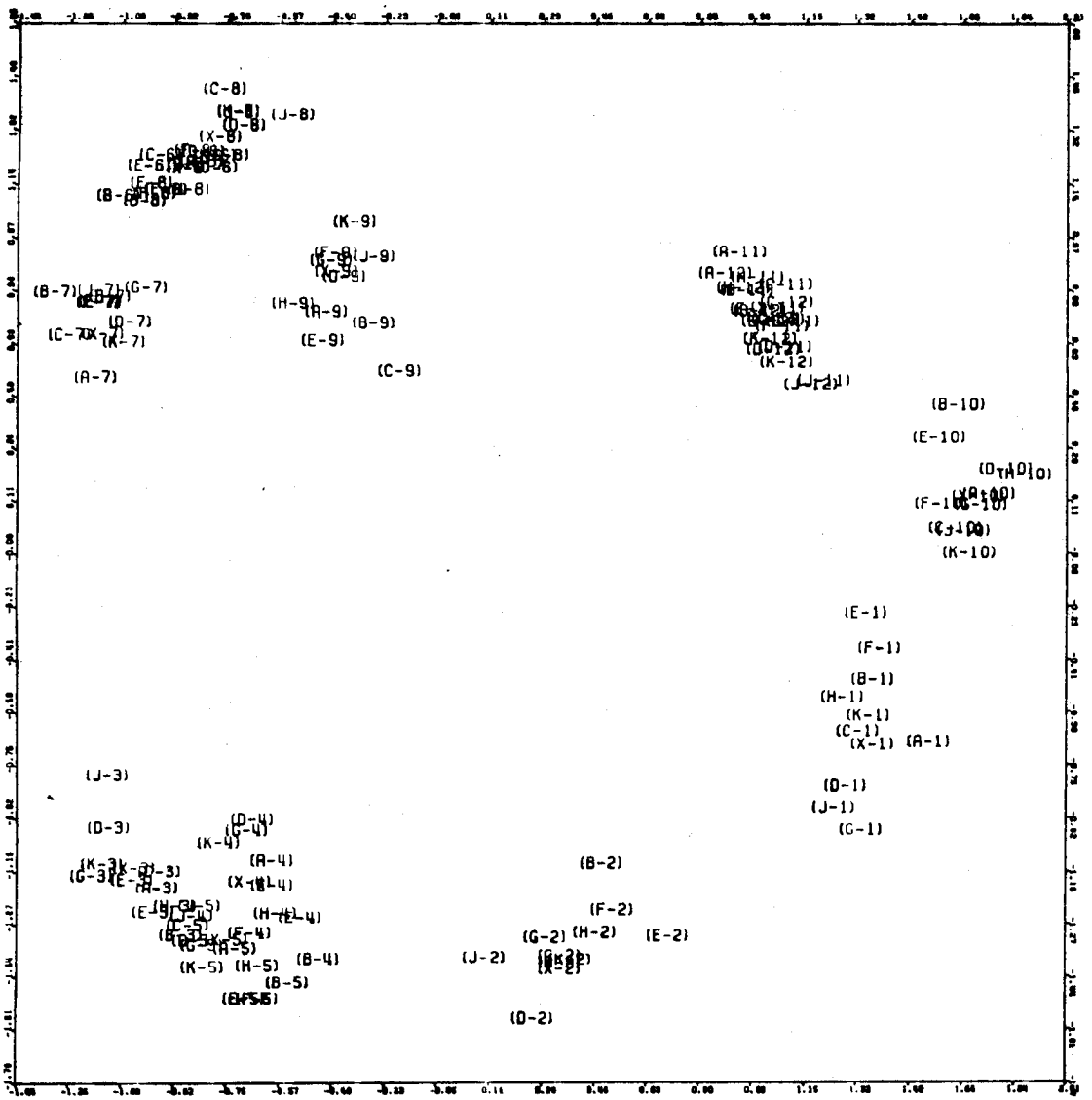
figuur 9.1. Drie verdelingen rond σ^2.

9.2. Voorbeelden

We geven nu een aantal voorbeelden van de bootstrap op PRINCALS, ANACOR en HOMALS. De resultaten van de verschillende steekproeven zijn steeds, op spiegelingen na, zonder meer over elkaar heen geplot. Er zijn dus bijvoorbeeld geen extra rotaties toegepast om de verschillende konfiguraties onderling zo goed mogelijk op elkaar te doen lijken. De voorbeelden zijn inhoudelijk al verschillende malen aan de orde geweest, dus we volstaan met enige summiere kommentaren.

9.2.1. PRINCALS bootstrap op voorkeuren in de Tweede Kamer (1968)

Het aantal steekproeven is hier 10, genummerd in de plot (zie figuur 9.2.) van A t/m K, waarbij I niet gebruikt wordt. Met X is



Figuur 9.2. PRINCALS bootstrap op voorkeuren 2e kamer

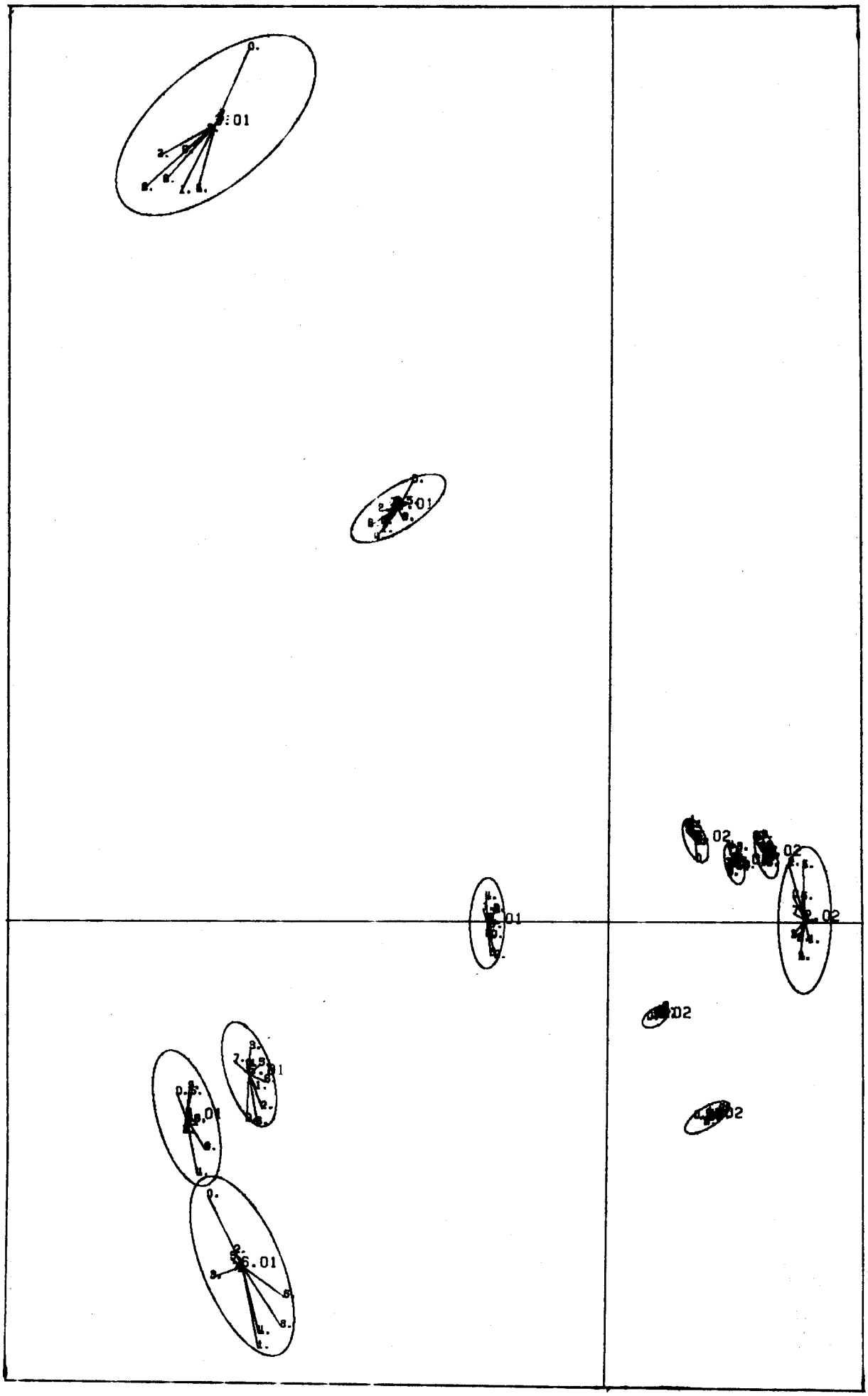
1 - CPN
2 - PSP
3 - PvdA
4 - D'66
5 - PPR
6 - KVP
7 - ARP
8 - CHU
10 - Boerenpartij
11 - SGP
12 - GPV

is de oorspronkelijke oplossing aangeduid. De gegevens betreffen de voorkeursrangordeningen van 141 kamerleden voor de genoemde partijen (zie 5.4.2.). In alle analyses zijn ordinale opties gebruikt. Uit figuur 9.2. blijkt, dat de puntenwolken voor KVP en CHU, SGP en GPV, en in iets mindere mate PvdA, PPR en D'66 elkaar overlappen, wat dus betekent dat hun onderlinge ligging niet erg stabiel is. Bovendien maakt deze bootstrap duidelijk, meer nog dan uit figuur 5.11. blijkt, dat CPN en Boerenpartij elkaar in hun onpopulariteit zeer dicht naderen. Het "gat" tussen de CDA partijen en het linkse blok is duidelijk stabiel.

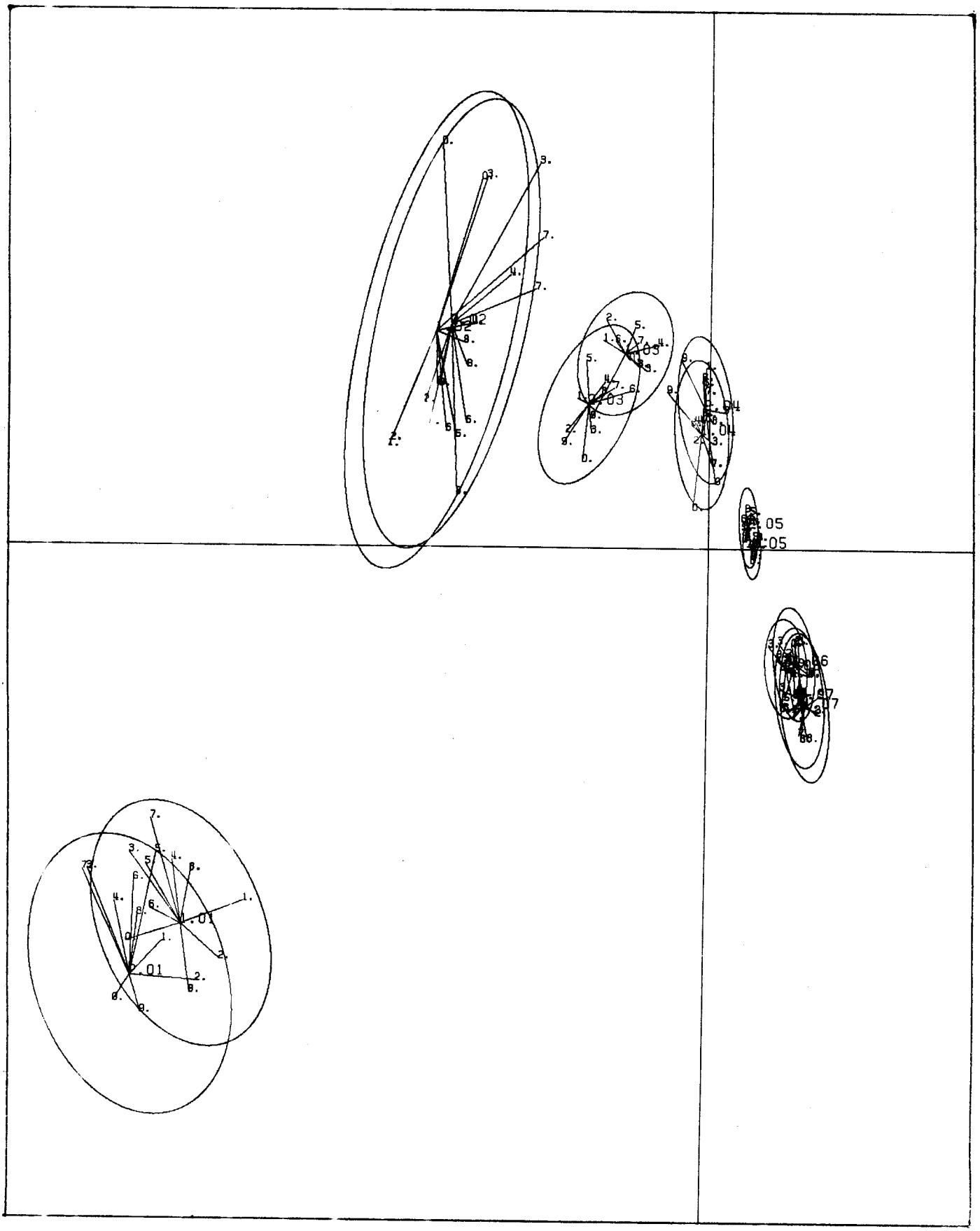
9.2.2. ANACOR bootstrap op Japanners & Sociale Mobiliteit

De gegevens van Sugiyama over godsdienstige gebruiken in Japan bestaan uit antwoorden van 4243 individuen op 6 vragen (vgl. 3.2.2.). De Sociale Mobiliteitgegevens van Glass (vgl. 3.2.3.) zijn gebaseerd op een steekproef van 3497 gezinnen. We hebben hier dus behoorlijk grote n, en we zijn vooral benieuwd of de benadering met ellipsen (vgl. 4.1.5.) en die met de bootstrap overeenkomstige resultaten geven. Daarom hebben we in figuur 9.3 en 9.4 zowel ellipsen als de bootstrap-punten weergegeven. Uit beide plaatjes blijkt, dat vrijwel zonder uitzondering alle bootstrap punten binnen de ellipsen vallen. We kunnen dus uit de bootstrap resultaten hetzelfde soort konklusies trekken als we op grond van de ellipsen al deden. Voor de Japanners geldt dat het kontrast tussen items 1 en 3 en 4,5 en 6 stabiel is, en de nee-kategorie van item 2 (GRAVE) het meest gespreid is van de nee-kategorieën. Merk op dat het niet perse zo is, dat de bootstrap punten "homogeen" verspreid liggen binnen hun ellips; ja op item 3 komt bij de bootstrap een beetje meer naar beneden te liggen dan we op grond van de ellipsen zouden verwachten. Iets dergelijks geldt voor item 6.

Uit de resultaten voor de Sociale Mobiliteitsgegevens zien we dat opnieuw 6 en 7 (semi-skilled en unskilled) voor vader én zoon ononderscheidbaar zijn, dat we de meeste asymmetrie hebben bij 3 (higher supervisory) en de meeste spreiding bij 2 (managerial and executive). Voor vader en zoon 1 (professional) lig-



Figuur 9.3. ANACOR bootstrap op Japanners



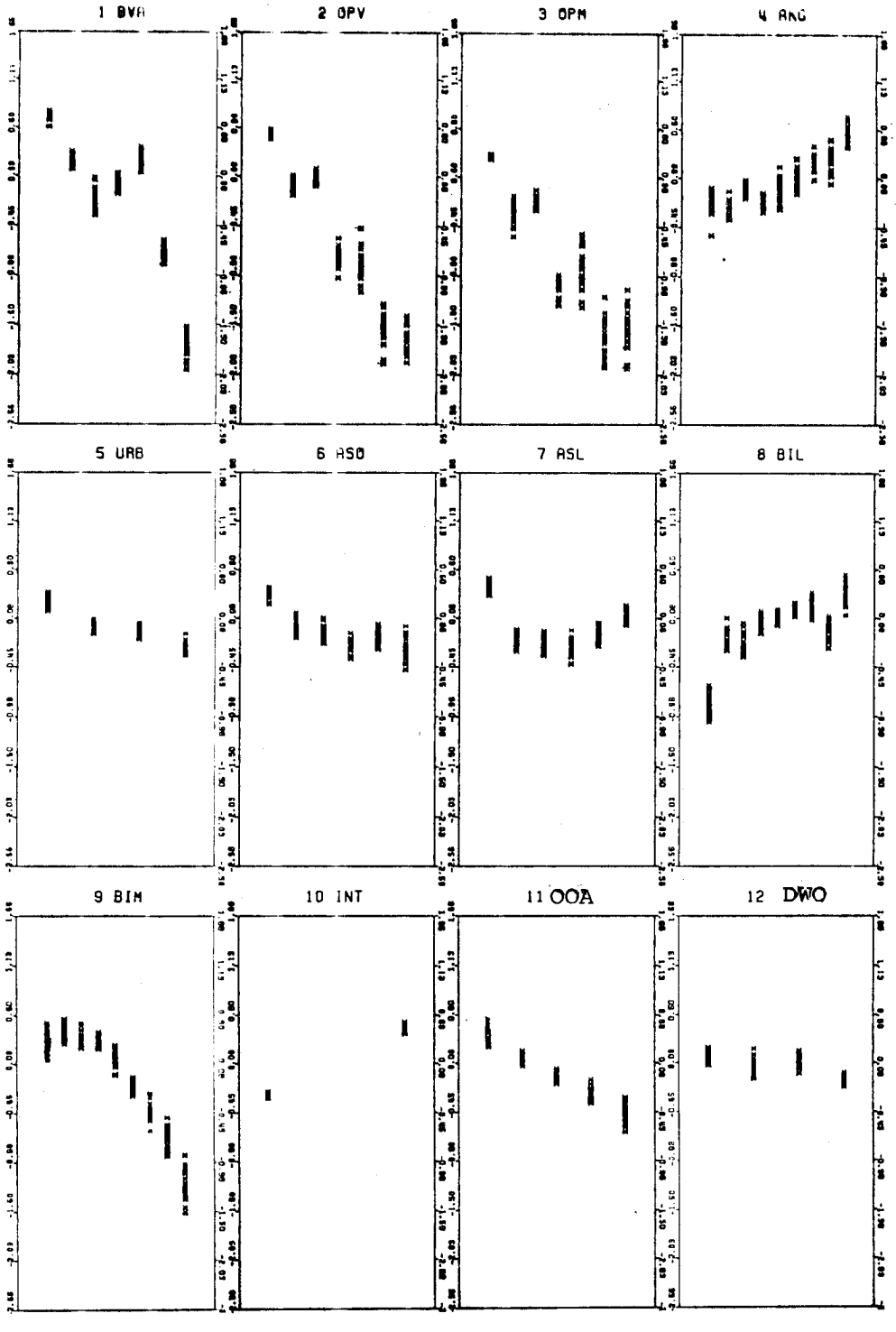
Figuur 9.4. ANACOR bootstrap op Sociale Mobiliteit

gen de bootstrap punten weer vrijwel alleen in de bovenkant van de ellipsen. De "vorm" van de bootstrap puntenwolken kan dus soms iets meer informatie geven dan de ellips. Aan punt 2 is overigens prima te zien dat bootstrap en ellips dezelfde kant op wijzen.

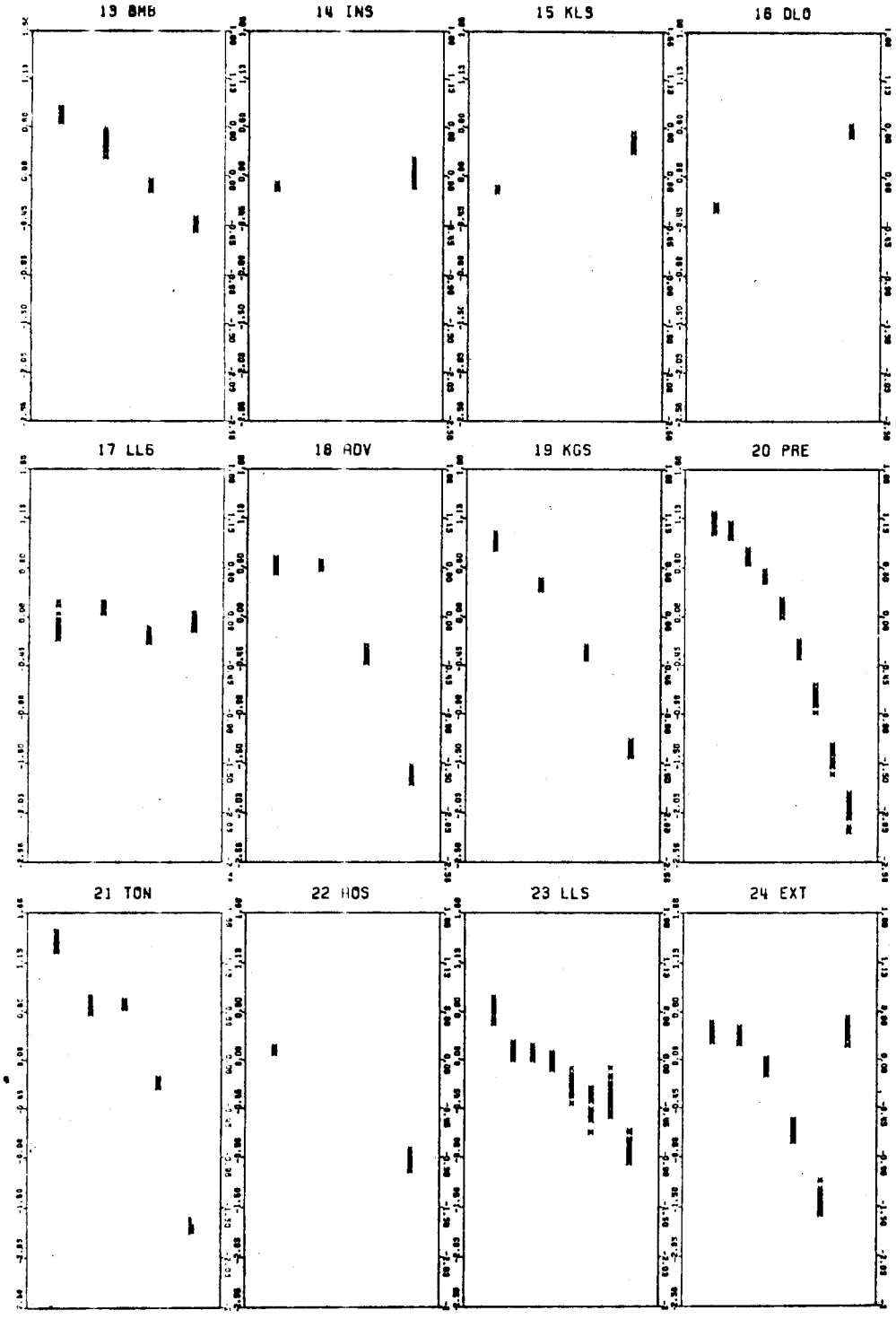
9.2.3. HOMALS bootstrap op Van Jaar tot Jaar

In het van Jaar tot Jaar onderzoek hebben we n = 1845, en we hebben s = 25 gebruikt. De bootstrap geeft ons nu een goede indruk van de stabiliteit van de transformaties van de variabelen (zie figuren 9.5.a,b,en c) die op SEX na vergelijkbaar zijn met figuur 2.11; door spiegeling van de as lopen ze wel allemaal precies andersom. We bespreken niet alle variabelen, maar lopen er een paar na.

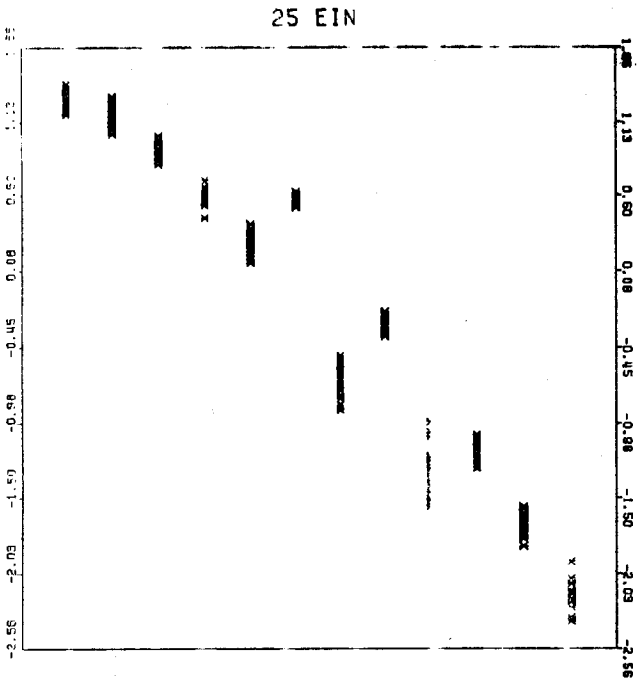
De volgorde wisseling van BVA is gezien de bootstrap stabiel; ook de kategorieën 2,3 en 4,5 en 6,7 van OPV overlappen elkaar duidelijk. BIM's transformatie is prachtig stabiel, BIL's uitschietende kategorieën 1 en 8 ook. Van de variabelen ADV, PRE en TON, die een hoge diskriminatiewaarde hebben, spreiden de punten ook zeer weinig en is goed te zien dat er òf overlap is òf onderscheid. Bij bijv. LLS, die slechter diskrimineert, markeren de bootstrap punten toch nog het kontrast tussen 2,3 en 4 versus 5,6, en 7. Als we de bootstrap kategorie transformaties vergelijken merken we op dat hoge marginale frekwenties van een kategorie vaak samen gaan met kleine spreiding van transformaties en vice versa (vgl. DLO-1, n = 1295 en bv. OPV-1, n =761 en OPV-6, n =56). Verder lijkt de spreiding van de diskriminatiematen (zie figuur 9.6.) van een variabele samen te hangen met de hoogte van de diskriminatiemaat: variabelen die het "slecht" doen, doen het vaak even slecht, terwijl "middelmatige" variabelen weer een grotere spreiding vertonen dan zowel de "goede" als de "slechte". We zijn dit verschijnsel ook bij analyse op deelbestanden tegen gekomen,b.v. de belangijke variabele TON is bij onderverdeling in jongens/ meisjes of in de diverse beroepsgroepen ook steeds belangrijk, terwijl b.v. DWO systematisch onbelangrijk is. De diskriminatiematen van AOS en OPV variëren meer.



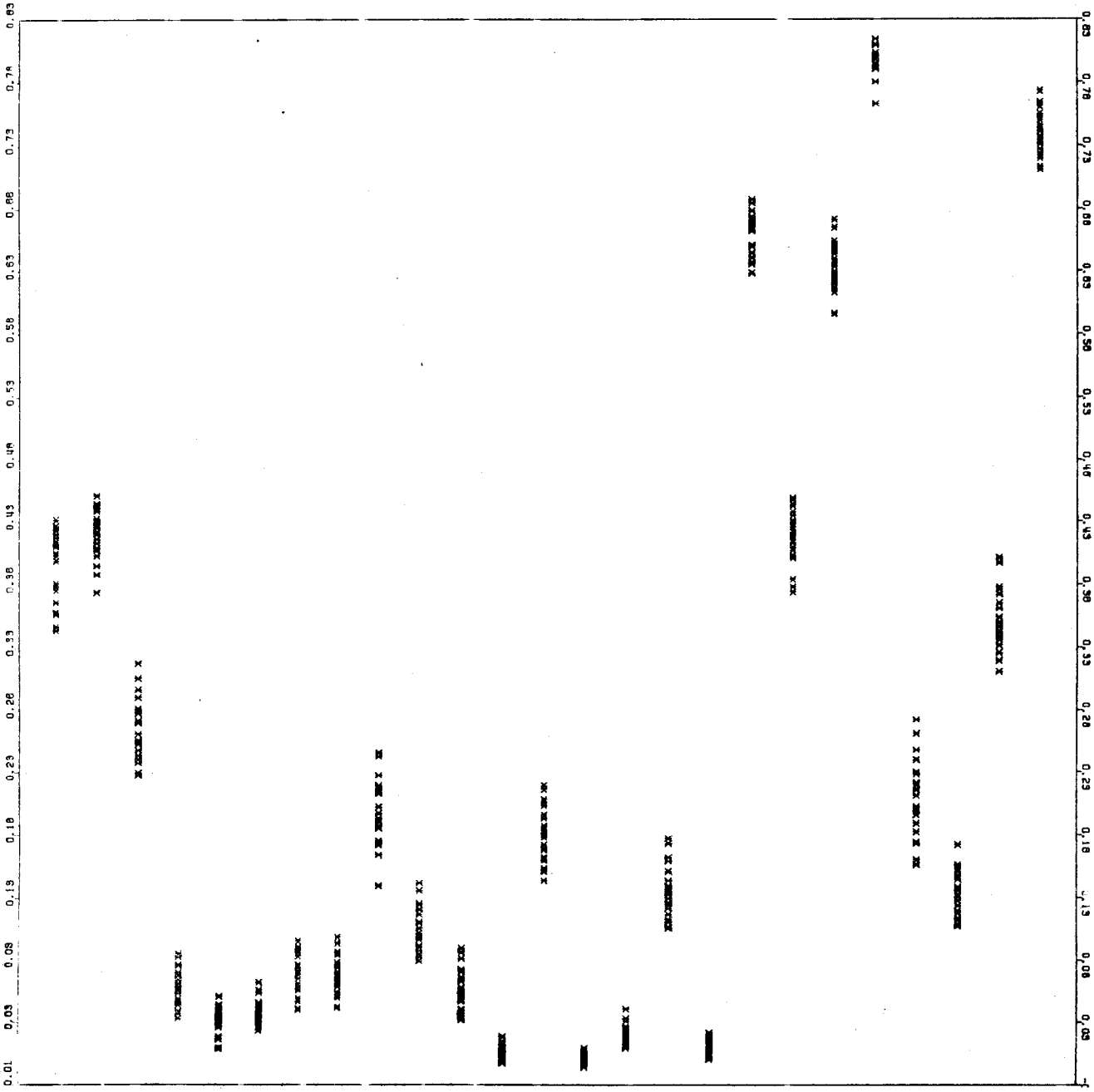
Figuur 9.5.a HOMALS bootstrap JJ;transformaties van variabelen



Figuur 9.5.b. HOMALS bootstrap JJ; transformaties van variabelen



Figuur 9.5.c. HOMALS bootstrap JJ; transformaties van de variabele Eindnivo



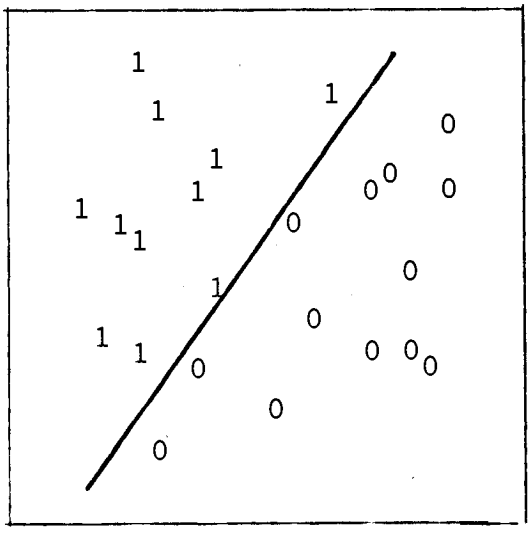
Figuur 9.6. HOMALS bootstrap JJ;diskriminatiematen

10.0. HOMALS en MDS.

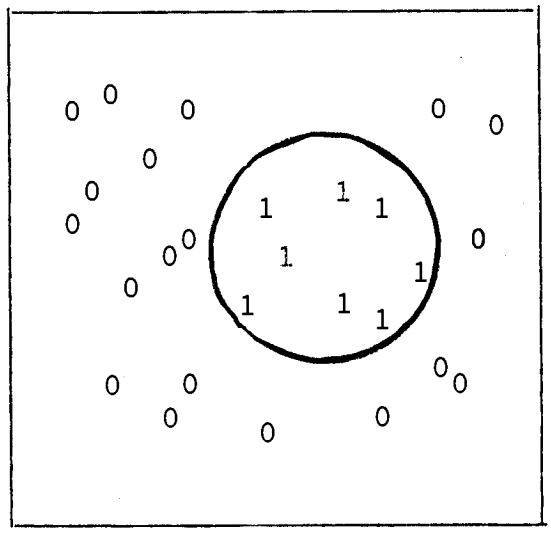
Bij de geometriese benadering van HOMALS is al naar voren gekomen, dat de relaties tussen individuen en kategorien worden weergegeven als afstanden. Omdat het begrip afstand zo'n prominente rol speelt in Multidimensional Scaling (MDS), dringt de vraag zich op, of we HOMALS kunnen zien als een lid van deze familie van technieken. Het antwoord is: ja, HOMALS is ekwivalent met een bepaalde vorm van Metries Multidimensionaal Ontvouwen van binaire gegevens.

In de zestiger jaren zijn diverse pogingen ondernomen (Coombs,1964, Lingoes, 1968; de Leeuw, 1969) om de toendertijd snel in populariteit groeiende Niet-metriese (lees: ordinale) MDS technieken toe te passen op rechthoekige binaire matrixen. In deze voorstellen was men er steeds op uit, zo zwak mogelijke aannamen te doen over de data, en werd de rol van rijen en kolommen altijd asymmetries opgevat. Het algemene idee was, de rij(kolom)-objekten als punten in de ruimte weer te geven en dan de kolom(rij)-objekten elk te laten korresponderen met een of ander eenvoudig scheidingsvlak. Op deze manier werden de data kolom(rij)-konditioneel gebruikt, en de scheidingsvlakken zouden steeds punten korresponderend met nullen moeten onderscheiden van punten korresponderend met enen. Een aantal vormen die voor de scheidingsvlakken kunnen worden gekozen staan in figuur 10.1. Deze diverse vormen kunnen allemaal als varianten van het Ontvouwings model van 10.1.b gezien worden: de kontoer van 10.1.a kunnen we benaderen door het middelpunt van de ontvouwingscirkel denkbeeldig naar oneindig te laten gaan, de kontoer van 10.1.c krijgen we in een ontvouwings model met zgn. Dominantie metriek en ook 10.1.d kunnen we benaderen met stukken cirkel en restrikties op de rij(kolom)-punten.

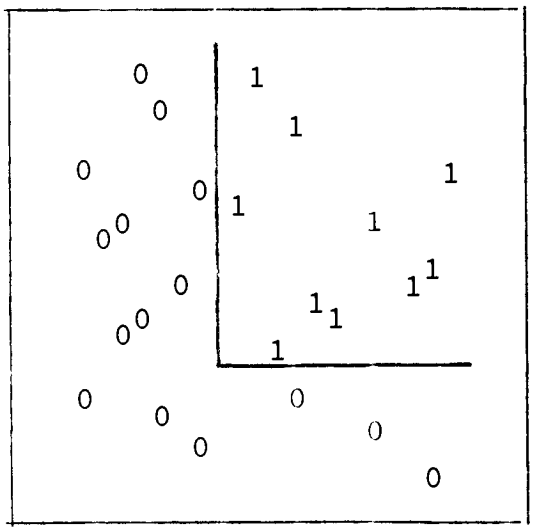
Nu is het tot op heden zeer moeilijk gebleken een aanvaardbaar werkend algoritme te ontwikkelen voor het Niet-metriese Ontvouwings model. Dit heeft vermoedelijk te maken met het geringe aantal restrikties dat uit het model volgt (zie bv. Kruskal en Carroll, 1969; Heiser en de Leeuw, 1979). Voor binaire gegevens wordt wat dit betreft de zaak nog erger en we kunnen in deze situatie nog meer vormen van 'degeneratie', nog meer 'lokale minima' en nog meer alternatieve oplossingen met vrijwel dezelfde stress verwachten dan bij 'gewone' ontvouwing.



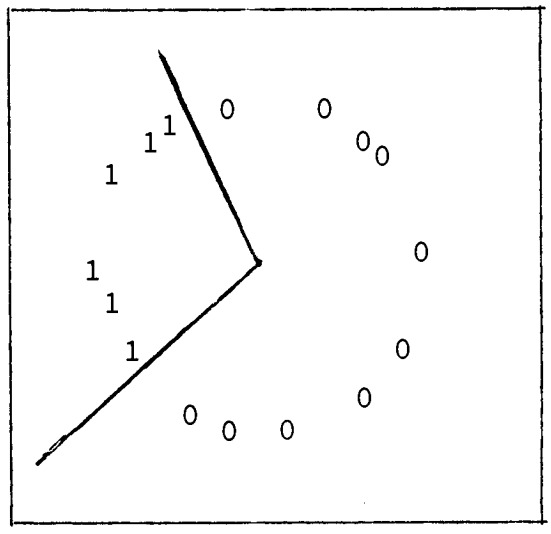
a. Hypervlakken/lijnen (PCA-model).



b. Bollen/cirkels (Unfolding model).



c. Rechte hoeken (Konjunktief model)



d. Kegels vanuit de oorsprong (circumplex)

figuur 10.1. Vorm van de scheidingsvlakken onder vier geometriese modellen.

Het blijkt dat het zwaartepunt-principe van HOMALS goed van pas komt. We zullen in dit hoofdstuk eerst laten zien hoe HOMALS als een MDS techniek voor binaire gegevens geformuleerd kan worden, waarin de aannamen van 'non-metriciteit' en 'konditionaliteit' over boord gezet zijn; daarna bespreken we hoe niet-metriese unfolding via HOMALS benaderd kan worden. Vervolgens bekijken we de relaties tussen HOMALS en andere vormen van MDS voor binaire gegevens, te weten de preferentie analyse techniek van (onder meer) Carroll en Chang (1964), MDPREF, en de Meer-dimensionale Scalogram Analyse techniek van Guttman en Lingoes, MSA (Lingoes, 1968). We besluiten het hoofdstuk met een aantal toepassingen van de Radex theorie van Guttman.

10.1. Unfolding van binaire gegevens.

Ons uitgangspunt is de indikator supermatrix G, waarbij we de indeling van de kolommen in groepjes (die in HOMALS de variabelen vormen) verwaarlozen; we komen hier in de radex voorbeelden nog op terug. G is binair, met elementen g_{ij}, van de orde nxm en met Gu = mu. We kunnen alleen de rij-objekten schalen, alleen de kolom-objekten, of beide tegelijkertijd. We beginnen met de eerste mogelijkheid (vgl. de 'drie wegen').

Elke kolom van G vatten we op als een ekwivalentierelatie $=_j$ (geinterpreteerd als "lijken op volgens j") op de rij-objekten $r_1, r_2, \ldots, r_i, r_k, \ldots, r_n$:

$$r_i =_j r_k \quad \text{als} \quad g_{ij} = g_{kj} = 1 \tag{1}$$

De bedoeling is nu, om de rij-objekten als punten $x_1, x_2, \ldots, x_i, \ldots x_n$ af te beelden in de p-dimensionale ruimte, zodanig dat voor alle j geldt

$$\text{als} \quad r_i =_j r_k \quad \text{dan} \quad x_i = x_k \tag{2}$$

Dus in woorden: rij-objekten die volgens j op elkaar lijken moeten samenvallen. Daarom koderen we (2) in m symmetriese nxn dissimilarity-matrixen, en definieren we bijbehorende matrixen van gewichten:

$$\left.\begin{aligned}\delta_{ikj} &= 0\\ w_{ikj} &= 1/d_j\end{aligned}\right\} \quad \text{als} \quad r_i =_j r_k \tag{3}$$

$$\left.\begin{aligned}\delta_{ikj} &= 1\\ w_{ikj} &= 0\end{aligned}\right\} \quad \text{als niet} \quad r_i =_j r_k$$

Hierbij is d_j gelijk aan de j'de kolomsom van G, dus het j'de element van u'G. De j'de dissimilarity matrix geeft dus aan, voor ieder paar rij-objekten, of ze volgens kolom j ekwivalent zijn of niet, en grote ekwivalentie klassen worden lichter gewogen dan kleine. Het standaard MDS probleem is, willekeurige dissimilarity matrixen te benaderen met afstanden in een ruimte van zo'n klein mogelijke dimensionaliteit (Torgerson, 1958; Shepard, 1962; Kruskal, 1964, 1977; Guttman, 1968; voor een recent overzicht van varianten en toepassingen, zie Carroll en Arabie (1979); een meer theoreties overzicht is te vinden in de Leeuw en Heiser, 1980).

We verzamelen de rij-punten x_i in een konfiguratie matrix X en noteren $\gamma_{ik}(X)$ voor de euclidiese afstanden tussen de rijen van X. De gebruikelijke MDS kleinste kwadraten verliesfunktie is dan

$$\sigma_1(X) = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{n} \sum_{k=1}^{n} w_{ikj}(\delta_{ikj} - \gamma_{ik}(X))^2 \tag{4}$$

en het metriese MDS probleem is, een X te vinden die $\sigma_1(X)$ minimaliseert onder een of andere normalisatie konditie. We kunnen het probleem zoals in (4) gesteld aanzienlijk reduceren door gebruik te maken van de bijzondere struktuur in (3). Immers, voor alle i,k,j geldt $w_{ikj}\delta_{ikj} = 0$, en dus vallen in

$$\begin{aligned}\sigma_1(X) = &\frac{1}{2} \sum_j \sum_i \sum_k w_{ikj}\delta^2_{ikj} + \frac{1}{2} \sum_j \sum_i \sum_k w_{ikj}\gamma^2_{ik}(X) -\\ &- \sum_j \sum_i \sum_k w_{ikj}\delta_{ikj}\gamma_{ik}(X)\end{aligned} \tag{5}$$

de eerste en de laatste term aan de rechterkant weg. We krijgen

$$\sigma_1(X) = \frac{1}{2} \sum_j \sum_i \sum_k w_{ikj}\gamma^2_{ik}(X) = \frac{1}{2} \sum_i \sum_k \gamma^2_{ik}(X) \sum_j w_{ikj} \tag{6}$$

De matrix $W_.$, met elementen $w_{ik.}$, is gelijk aan onze vertrouwde $GD^{-1}G'$. Hij telt voor ieder paar i,k hoe vaak ze bij dezelfde kolom horen, gewogen voor de marginalen van de kolommen. Omdat G een indikator super matrix is geldt

$$u'W_. = u'GD^{-1}G' = dD^{-1}G' = mu' \tag{7}$$

d.w.z. de rijen en kolommen van $W_.$ sommeren tot m. We zetten de vereenvoudigde verliesfunktie nu om in matrix notatie:

$$\begin{aligned}\sigma_1(X) &= \frac{1}{2}\sum_i\sum_k w_{ik.}\sum_s (x_{is}-x_{ks})^2 \\ &= \frac{1}{2}\sum_i\sum_s x_{is}^2\sum_k w_{ik.} + \frac{1}{2}\sum_k\sum_s x_{ks}^2\sum_i w_{ik.} - \sum_i\sum_k\sum_s w_{ik.}x_{is}x_{ks} \\ &= m\ \mathrm{tr}\ X'X - \mathrm{tr}\ X'GD^{-1}G'X \end{aligned} \tag{8}$$

Het minimaliseren van deze $\sigma_1(X)$ over alle X van de vorm $u'X = 0$ en $X'X = I$ is ekwivalent aan het HOMALS probleem voor de rij-objekten. De normalisatie konditie is hierbij wel strenger dan gebruikelijk is in MDS (te weten $\mathrm{tr}X'X = 1$).

Een geheel analoge redenering kunnen we opzetten wanneer we de kolomobjekten in plaats van de rij-objekten als punten willen representeren. In dat geval definieren we een ekwivalentie relatie over de produktverzameling van $c_1, c_2, \ldots, c_j, c_l, \ldots, c_m$:

$$c_j =_i c_l \quad \text{als} \quad g_{ij} = g_{il} = 1 \tag{9}$$

en de n symmetriese mxm dissimilarity matrixen

$$\left.\begin{aligned}\delta_{jli} &= 0 \\ w_{jli} &= 1/m\end{aligned}\right\} \quad \text{als} \quad c_j =_i c_l$$

$$\left.\begin{aligned}\delta_{jli} &= 1 \\ w_{jli} &= 0\end{aligned}\right\} \quad \text{als niet } c_j =_i c_l \tag{10}$$

Dit keer willen we over alle genormaliseerde Y minimaliseren

$$\sigma_2(Y) = \frac{1}{2}\sum_i\sum_j\sum_l w_{jli}(\delta_{jli} - \gamma_{jl}(Y))^2 \tag{11}$$

en dit kan gereduceerd worden tot

$$\sigma_2(Y) = \text{tr } Y'DY - \frac{1}{m} \text{tr } Y'G'GY \tag{12}$$

en dit probleem is ekwivalent aan HOMALS voor de kolom objekten.

We gaan nu de derde weg bewandelen en kijken of de drie wegen weer op hetzelfde uitkomen. We definieren de relatie ε op de produktverzameling van de rij- en kolom-objekten (vgl. hoofdstuk 3.1):

$$r_i \varepsilon c_j \quad \text{als} \quad g_{ij} = 1 \tag{13}$$

Dit is geen ekwivalentie relatie meer. Als dissimilarities definieren we nu eenvoudigweg

$$\left.\begin{aligned} \delta_{ij} &= 0 \\ w_{ij} &= 1 \end{aligned}\right\} \quad \text{als} \quad r_i \varepsilon c_j$$

$$\left.\begin{aligned} \delta_{ij} &= 1 \\ w_{ij} &= 0 \end{aligned}\right\} \quad \text{als niet } r_i \varepsilon c_j \tag{14}$$

en als we hier de metriese MDS (ontvouwing) verliesfunktie uitwerken krijgen we

$$\begin{aligned} \sigma_3(X,Y) &= \sum_i \sum_j w_{ij}(\delta_{ij} - \gamma_{ij}(X,Y))^2 \\ &= m \text{ tr } X'X + \text{tr } Y'DY - 2 \text{ tr } X'GY \end{aligned} \tag{15}$$

Hierbij zijn dus de afstanden $\gamma_{ij}(X,Y)$ alleen gedefinieerd tussen de i-de rij van X en de j-de rij van Y. Als we bovendien definieren

$$\sigma_3(X,*) \triangleq \min \{ \sigma_3(X,Y) : Y \} \tag{16}$$

is het duidelijk dat het minimum bereikt wordt voor

$$Y = D^{-1}G'X \tag{17}$$

dus het bekende en plezierige resultaat dat de kolompunten moeten samenvallen met de zwaartepunten van hun bijbehorende rijpunten. We kunnen nu invullen

$$\sigma_3(X,*) = m \text{ tr } X'X + \text{tr } X'GD^{-1}DD^{-1}G'X - 2 \text{ tr } X'GD^{-1}G'X$$

$$= m \text{ tr } X'X - \text{tr } X'GD^{-1}G'X = \sigma_1(X) \qquad (18)$$

Op dezelfde manier kunnen we X 'eruit minimaliseren' en houden dan $\sigma_2(Y)$ over. De drie verschillende formuleringen zijn dus inderdaad onderling ekwivalent. Aan de laatste formulering is goed te zien wat de relaties zijn met de meer gebruikelijke niet-metriese ontvouwing. In rij-konditioneel niet-metries binair ontvouwen wordt geeist

$$\text{als } g_{ij} = 0 \text{ en } g_{il} = 1 \text{ dan } \gamma_{ij}(X,Y) \geq \gamma_{il}(X,Y) \qquad (19)$$

In het overeenkomstige kolom-konditionele geval wordt geeist

$$\text{als } g_{ij} = 0 \text{ en } g_{kj} = 1 \text{ dan } \gamma_{ij}(X,Y) \geq \gamma_{kj}(X,Y) \qquad (20)$$

In het niet-metriese onkonditionele geval hebben we

$$\text{als } g_{ij} = 0 \text{ en } g_{kl} = 1 \text{ dan } \gamma_{ij}(X,Y) \geq \gamma_{kl}(X,Y) \qquad (21)$$

We hebben hier gezien dat HOMALS in feite eist

$$\text{als } g_{ij} = 1 \text{ dan } \gamma_{ij}(X,Y) = 0 \qquad (22)$$

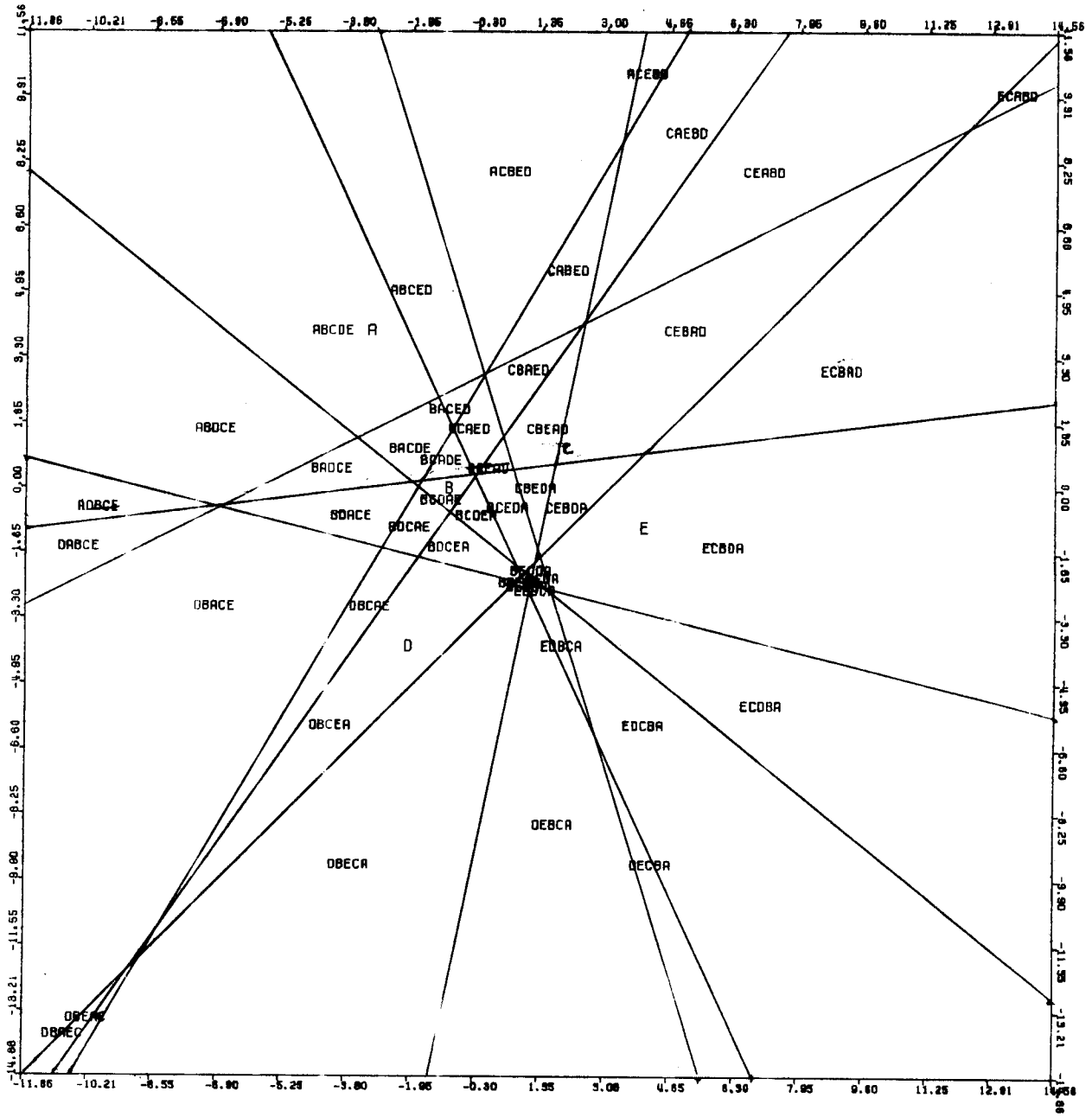
een veel sterkere eis dus dan alle andere gevallen. We hebben een praktiese oplossing van het binaire ontvouwingsprobleem 'gekocht' door de data metries op te vatten en een strengere normalisatie konditie aan één van de twee konfiguraties op te leggen. Daarbij zijn rijen en kolommen meer symmetries behandeld.

10.2. Algemene niet-metriese ontvouwing via HOMALS;

In het niet-binaire ontvouwings probleem is het de bedoeling een verzameling rangordeningen af te beelden (zgn. I-schalen, van individual) die n individuen hebben gegeven van m objekten. Gewoonlijk worden deze rangordeningen direkt geinterpreteerd als (monotoon verstoorde) afstanden tussen individu-punten (ideaalpunten) en objekt-punten (de zgn. J-schaal, van joint) in p dimensies. We noemen de rangnummers individuele utiliteiten u_{ij} en we zoeken een representatie X,Y zodanig dat

$$\text{als } u_{ij} < u_{il} \text{ dan } \gamma_{ij}(X,Y) \geq \gamma_{il}(X,Y) \qquad (23)$$

waarbij de afstanden $\gamma_{ij}(X,Y)$ weer euklidies zijn en we voor het



figuur 10.2. Alle isotone gebieden voor 5 punten in 2 dim.

gemak 'ties' uitsluiten (d.w.z. $u_{ij} \neq u_{il}$ voor iedere i). In onze inleiding hebben we al aangestipt dat het konstrueren van een goed algoritme voor dit probleem niet eenvoudig is en we bestuderen hier twee alternatieve formuleringen, waarbij we uitgaan van de HOMALS aanpak.

Allereerst bekijken we de komplete verzameling van isotone gebieden die in figuur 10.2 voor vijf willekeurig gesitueerde objekt punten zijn getekend. Een isotoon gebied is gedefinieerd als een gebied

waarvoor geldt dat alle punten die er binnen vallen dezelfde rangorde van afstanden tot de objekt-punten hebben (vgl. Coombs (1964), pag. 143). Elke middelloodlijn van twee punten verdeelt de ruimte in twee halfruimtes en omdat elk isotoon gebied de intersektie van ½m(m-1) halfruimtes is, krijgen we een partitie in <u>konvexe</u> gebieden.

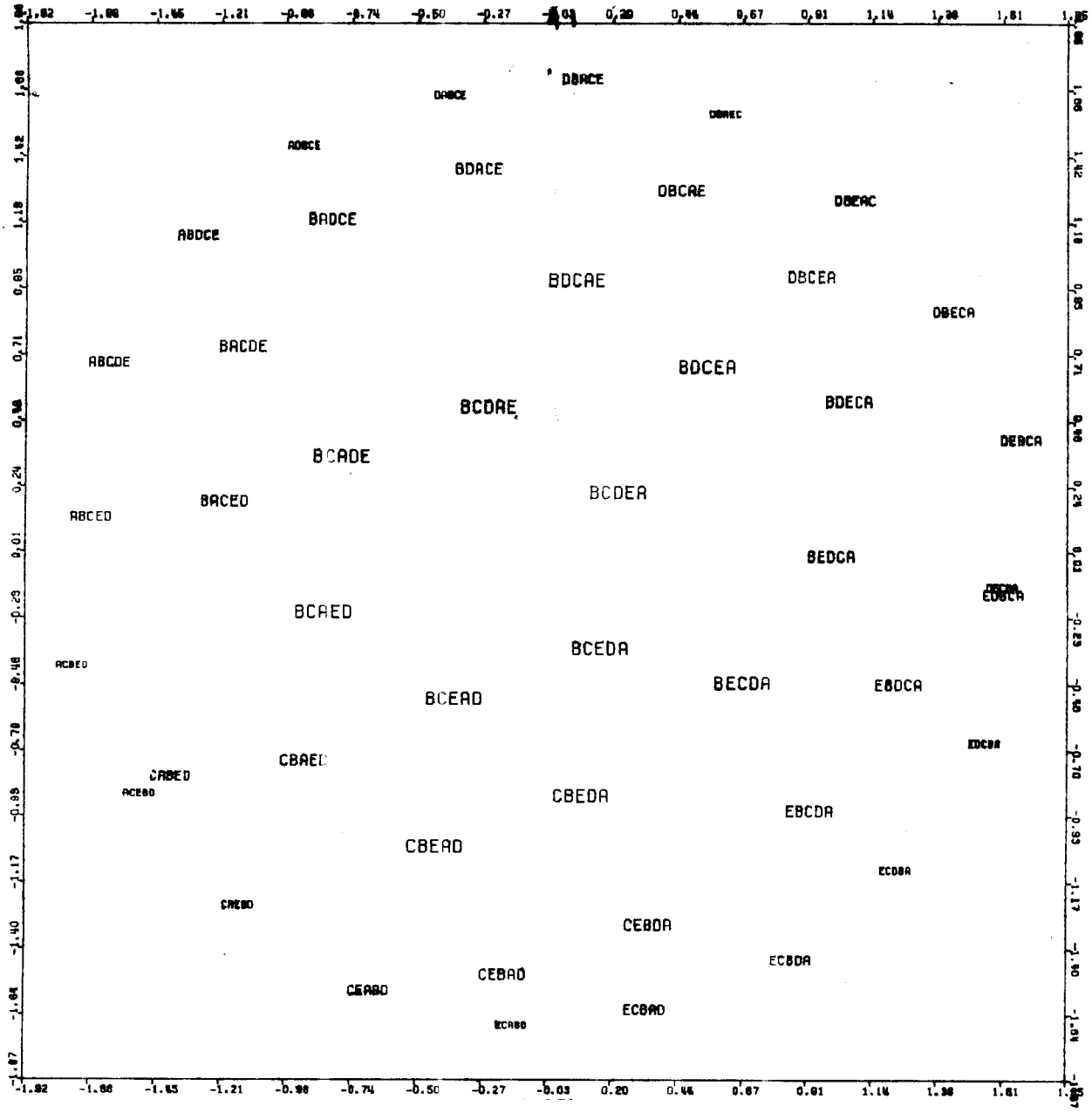
Figuur 10.2 laat al gelijk een fundamentele zwakte van de gebruikelijke aanpak van ontvouwing zien, die gebaseerd is op het idee dat de isotone gebieden inkrimpen tot punten (zoals in symmetriese niet-metriese MDS). Deze (ideaal)punten zijn dan de middelpunten van de iso-utiliteitsbollen/cirkels waarover in de inleiding gesproken werd. De 'inkrimpfilosofie' gaat in het centrum van de ruimte wel een beetje op (alleen als we meer objekt-punten toevoegen), maar een stuk minder drasties, omdat we alleen informatie aksepteren over de rij-konditionele rangorde van een deelverzameling van de afstanden (zie Heiser en de Leeuw (1979) voor een volledig metriese aanpak).

Een eerste, meest voor de hand liggende manier om het probleem te herformuleren in HOMALS vorm is door de opsplitsing van individuen bij elke paar-gewijze vergelijking van objekten te koderen. We krijgen dan ½m(m-1) binaire variabelen, geindiceerd met j,l, waarbij j,l alle paren met $j < l$ doorloopt:

$$\begin{aligned} &\text{als} \quad u_{ij} > u_{il} \quad \text{dan} \quad g_{i1}^{jl} = 1 \quad \text{en} \quad g_{i2}^{jl} = 0 \\ &\text{als} \quad u_{ij} < u_{il} \quad \text{dan} \quad g_{i1}^{jl} = 0 \quad \text{en} \quad g_{i2}^{jl} = 1 \end{aligned} \tag{24}$$

In het perfekte geval verdeelt dan elke variabele de ruimte in twee gescheiden gebieden, en met nominaal-kontinu verlies zouden de individuen perfekt representeerbaar zijn. Voorzover HOMALS (nominaal-diskreet) een goede benadering is van zo'n kontinue oplossing, zou HOMALS op kodering (24) ook een goede benadering van het ontvouwings probleem moeten geven (vlg. hoofdstuk 5.1.4).

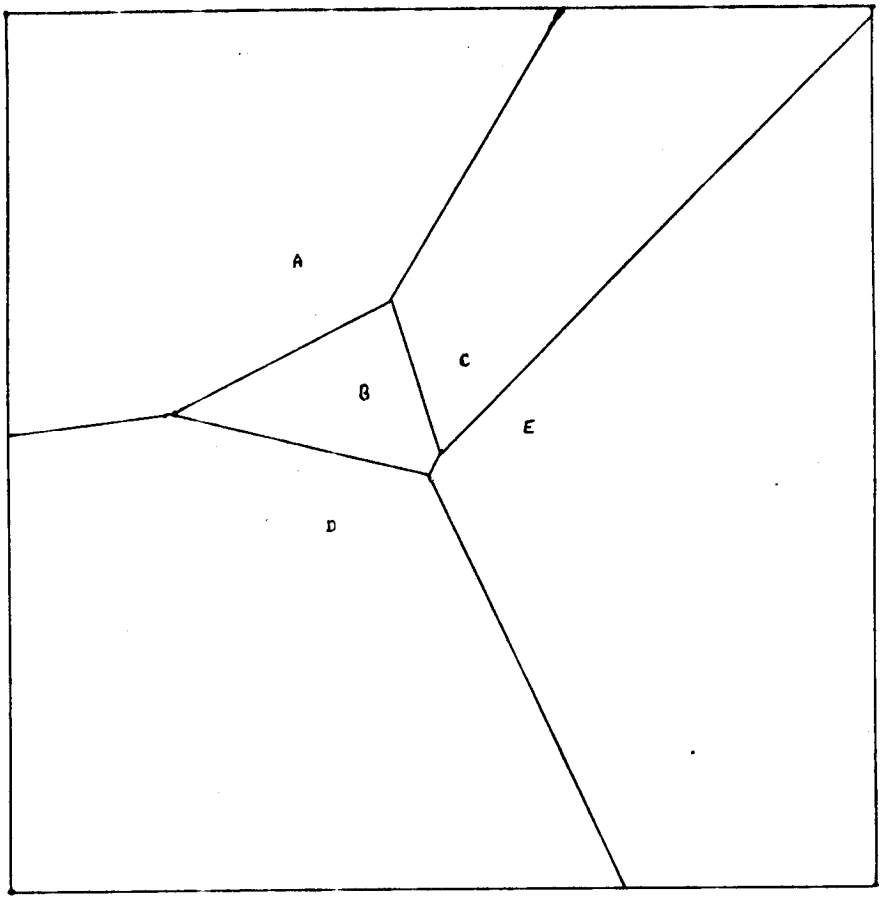
In figuur 10.3 geven we de HOMALS oplossing voor de data die uit figuur 10.2 kunnen worden gekonstrueerd. In de plot is de derde dimensie suggestief aangegeven door de labels groter te maken naar mate de koordinaatwaarden op de derde as groter zijn. De punten liggen dus op een bult. Opmerkelijk is, dat HOMALS de individupunten als het ware zoveel mogelijk op gelijke onderlinge afstand



figuur 10.3. HOMALS oplossing volgens binaire kodering.

heeft gelegd. Het kluitje punten BCADE, BCDAE, BCDEA, BCEDA, BCEAD en BCAED, dat in deze volgorde steeds één volgorde-wisseling heeft, is op een cirkel verspreid; meer perifere punten die onderling één volgorde-wisseling hebben zijn meer naar elkaar toegetrokken. Afgezien van deze effekten is de kwaliteit van de rekonstruktie frappant. De rechte lijnen van figuur 10.2 kunnen met kwadratiese lijnen benaderd worden. Als objekt-punten zouden we kunnen nemen het zwaartepunt van alle individuen die het betreffende objekt als eerste keuze hebben.

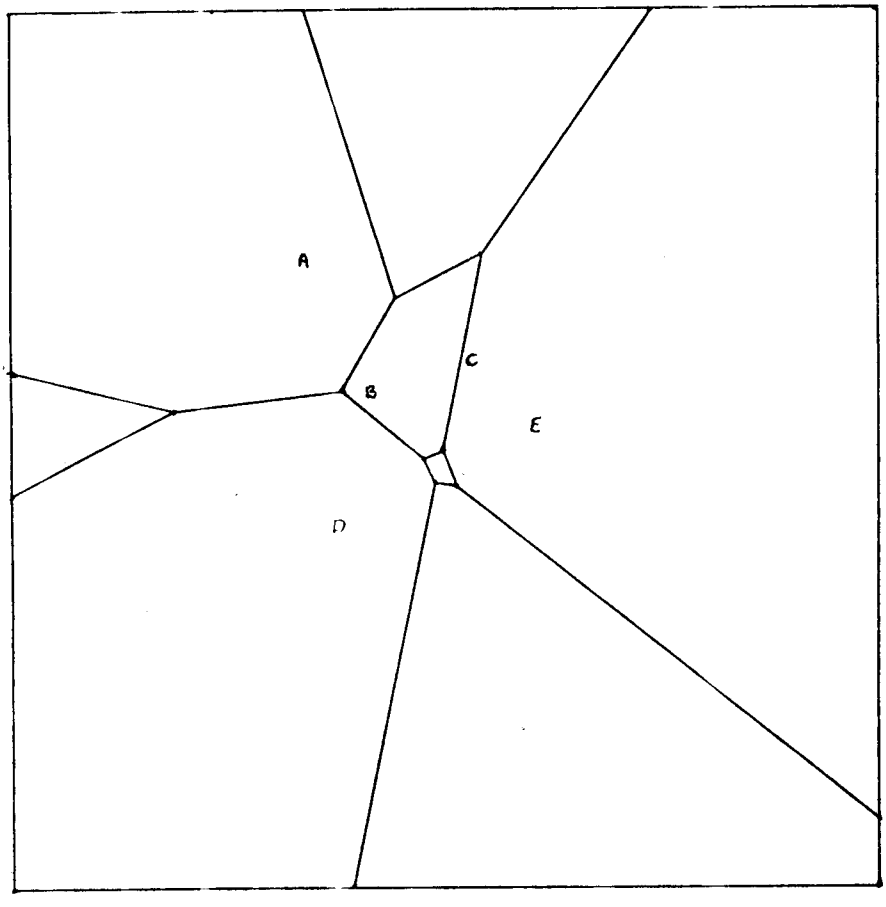
Een tweede aanpak is gebaseerd op het idee dat we m - 1 partitio-



figuur 10.4. Opsplitsing naar eerste keuze.

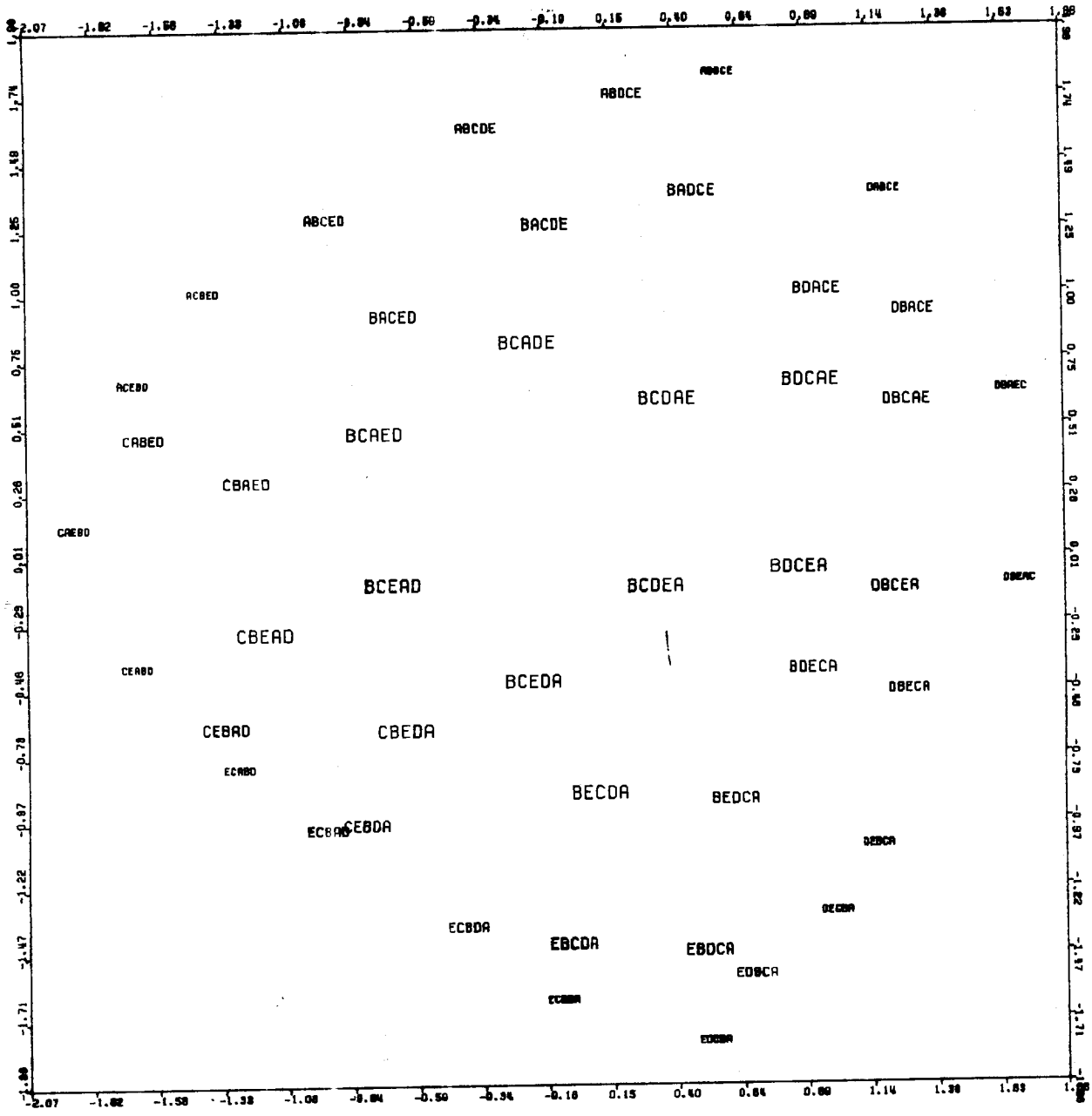
neringen van de individuen kunnen maken, die in het perfekte geval overeenkomen met een opsplitsing van de individu-punten in konvexe gebieden. De eerste opsplitsing, naar eerste keuze, is getekend in figuur 10.4. Voor de tweede partitionering moeten we niet nemen de variabele 'tweede keuze', want die resulteert niet in mooie konvexe gebieden. We moeten die individuen bij elkaar nemen, die steeds voor een bepaald paar van objekten ofwel de één als eerste en de ander als tweede, ofwel andersom gekozen hebben. Dit noteren we met (AB)(CDE). In het voorbeeld van figuur 10.2 krijgen we dus een tweede variabele met als kategorieen (AB)(CDE), (AC)(BDE), (AD)(BCE), (BC)(ADE), (BD)(ACE), (BE)(ACD), (CE)(ABD) en (DE)(ABC). Deze opsplitsing 'eerste en tweede keuze' is getekend in figuur 10.5.

Op deze manier kunnen we een derde variabele definieren met kategorieen van de vorm (ABC)(DE) en een vierde met kategorieen van de vorm (ABCD)(E). Alle opsplitsingen zijn konvex en vormen, over elkaar heen gelegd, precies de originele isotone gebieden. Alle



figuur 10.5. Opsplitsing naar 1ste en 2de keus.

variabelen zijn weer meervoudig kontinu-nominaal perfekt representeerbaar (alle kategorieen kunnen paargewijs worden gescheiden door een rechte lijn) en te benaderen met HOMALS diskreet-nominaal. Het is niet eenvoudig 'ties' toe te laten, behalve in het speciale geval dat zij bijv. ontstaan zijn door een data-verzamelingstechniek als "orden de k_1 eerste en k_2 laatste keuzen van m objekten". De objekt-punten definieren we weer als zwaartepunten van eerste keuzen, dit keer dus de kategorie-punten van de eerste variabele. Het HOMALS resultaat wanneer we de konvexe gebieden kodering gebruiken staat geplot in figuur 10.6. Opnieuw vinden we in grote lijnen de positie van de individuen op een bult terug, met hetzelfde type effekten (deze zijn trouwens o.a. toe te schrijven aan het feit dat we alle patronen precies één keer hebben opgenomen en niet hebben gewogen met een of andere enkeltoppige dichtheidsfunktie). Het lijkt erop, dat de 'lange as' ECABD versus DBAEC/DBEAC



figuur 10.6. HOMALS oplossing volgens konvexe kodering.

wat 'kromgetrokken' is; dit hoeft ons niet te verontrusten aangezien de positie van de patronen in de 'open' gebieden natuurlijk erg vrij is.

De aanpak van het algemene ontvouwingsprobleem via HOMALS ziet er veelbelovend uit. Er is nog maar weinig ervaring mee opgedaan, de invloed van weging is bijvoorbeeld nog niet systematies onderzocht. Deze sektie moge eens te meer bewijzen dat het soms nut heeft, even stil te staan bij de definitie van de kategorieen van een variabele. Niet altijd is de meest voor de hand liggende, op de

gedachte van diskretisering van een kontinue variabele gebaseerde kodering ook de meest gewenste. Bovendien hoeft rij-konditionaliteit van de 'originele' data nog niet noodzakelijkerwijs onvergelijkbaarheid van individuen te betekenen.

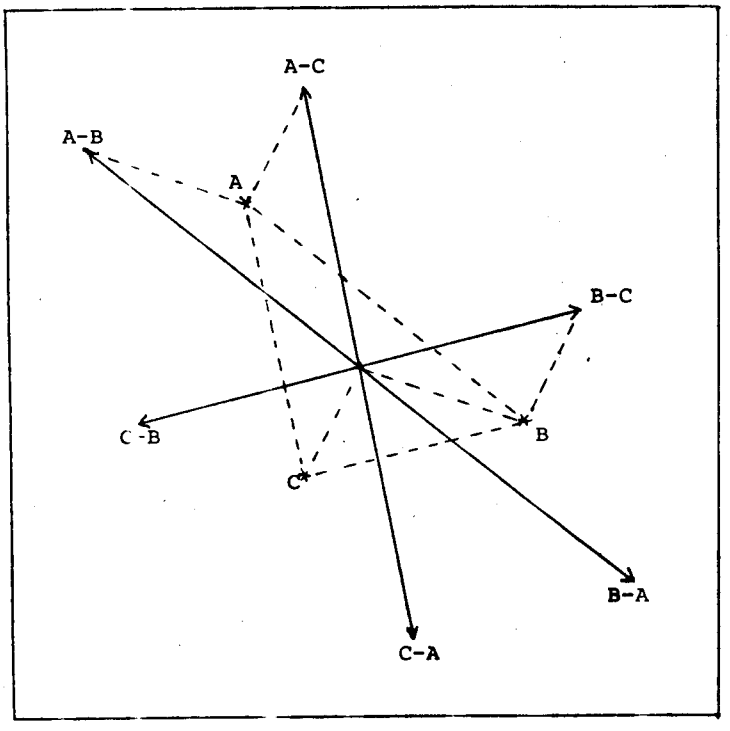
10.3. MDPREF als HOMALS met restrikties.

MDPREF is de naam van een techniek voor de analyse van preferentie gegevens die al een tijd bestaat (zie bijv. Guttman, 1946; Slater, 1960; Bechtel, 1969; Benzécri, 1967) maar vooral door toedoen van Carroll (1972) bekend is geraakt. Een andere naam is 'vector-model' of 'Q-analyse'. Uitgangspunt zijn weer de individuele utiliteiten u_{ij} uit de vorige sektie, maar dit keer zullen we de individuen niet weergeven als een punt, maar als een vektor. In z'n meest gebruikelijke vorm komt de techniek neer op een rechtstreekse Singuliere Waarden Dekompositie van U. Wat we hier willen laten zien is uitsluitend, hoe deze procedure valt af te leiden als een bijzonder geval van HOMALS. Een veel algemenere en tevens gedetailleerdere behandeling is te vinden in de Leeuw (1973).

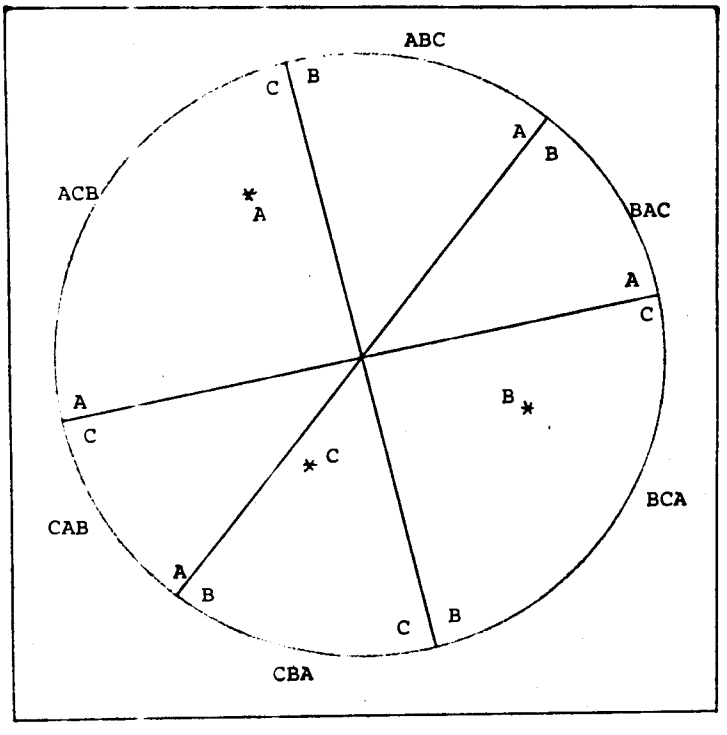
We gebruiken weer dezelfde kodering van (24), maar gaan restrikties op de kategorie-punten aanleggen, omdat we geen punten voor objekt paren, maar voor de objekten zelf willen vinden. Hoe? Te denken valt aan Thurstone- of Bradley-Terry-Luce modellen voor Pair Comparison gegevens, waar de waarschijnlijkheid dat j geprefereerd wordt boven l een funktie is van het verschil tussen hun schaalwaarden op een één-dimensionale schaal. In ons meer-dimensionale geval eisen we zoals altijd dat het kategorie-punt y_1^{jl} (het zwaartepunt van alle individuen die j boven l prefereren) zo ver mogelijk af ligt van y_2^{jl} (het zwaartepunt van de anderen) en dit bovendien in de richting waarin j en l het meest verschillen; d.w.z. we identificeren y_1^{jl} en y_2^{jl} met de verschilvektor van de te vinden punten z_j en z_l:

$$y_{1s}^{jl} = z_{js} - z_{ls}$$
$$y_{2s}^{jl} = z_{ls} - z_{js} \qquad (25)$$

Het idee is geillustreerd in figuur 10.7.a, voor drie punten A,B,C. Merk op dat A-B en B-A even lang zijn, d.w.z. in het algemeen zul-



figuur 10.7.a Drie punten met hun verschilvektoren.



figuur 10.7.b. Isotone gebieden t.o.v. de verschilvektoren.

de individu-punten niet meer gecentreerd kunnen zijn (tenzij in het geval van gelijke marginale frekwenties). Verder wordt in feite geeist, dat de kategorie-punten zó liggen, dat hun middelloodlijnen allemaal door de oorsprong gaan (zie figuur 10.7.b). We hebben als 't ware de dichte isotone gebieden (zoals in figuur 10.2) hier weggewerkt. Eén konsekwentie hiervan is, dat in een kompleet, perfekt geval van elk patroon ook z'n spiegelbeeld zal voorkomen; een andere, dat we zonder verlies aan algemeenheid kunnen eisen dat de individu-punten op dezelfde afstand van de oorsprong liggen (vektor-representatie).

Hoe verwerken we de reparametrisering (25) van Y in HOMALS? Definieer de m(m-1)xm design matrix A, die uit stukken A_{jl} bestaat, van de orde 2xm, met als elementen

$$a_{1j}^{jl} = 1 \qquad a_{1l}^{jl} = -1$$

$$a_{2j}^{jl} = -1 \qquad a_{2l}^{jl} = 1$$

en nul overal elders. Dan wordt (25)

$$Y_{jl} = A_{jl}Z \tag{26}$$

en in plaats van het gebruikelijke

$$\min \text{SSQ } (X - GY) \tag{27}$$

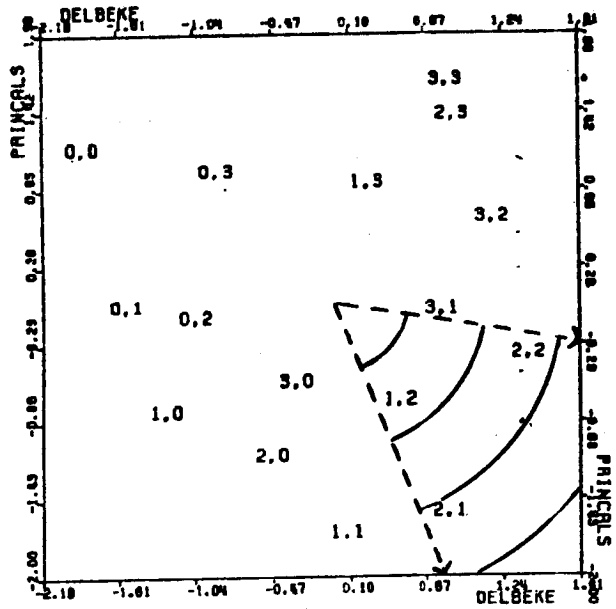
over X en Y, wordt het probleem

$$\min \text{SSQ } (X - GAZ) \tag{28}$$

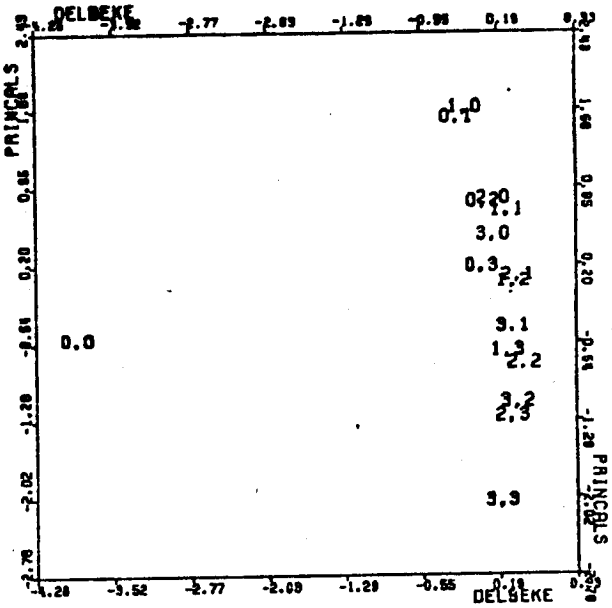
over X en Z. Het is niet moeilijk in te zien, dat GA op een konstante na de rij-gecentreerde rangnummers van U bevat en dat dus HOMALS met restrikties op de kategorie-punten overeenkomt met een SVD op de rij-gecentreerde rangnummers.

De techniek is 'niet-metries' in de zin dat de utiliteiten altijd worden behandeld in pair-comparison vorm (cq. rangnummer vorm), zodat de gevonden konfiguratie invariant is onder monotone transformaties van de data. Verder is de techniek natuurlijk metries; er wordt geen optimale schaling op de data uitgevoerd.

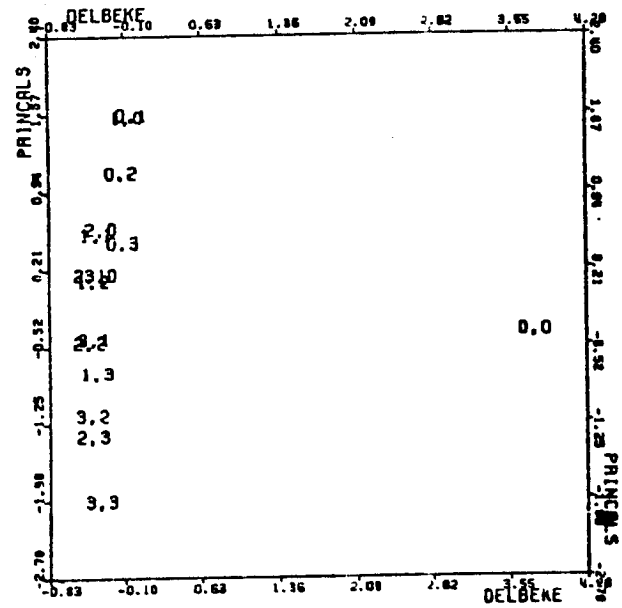
Wie dit laatste wel wil, moet PRINCALS gebruiken in gekantelde vorm (d.w.z. rijen objekten, kolommen individuen). Een voorbeeld



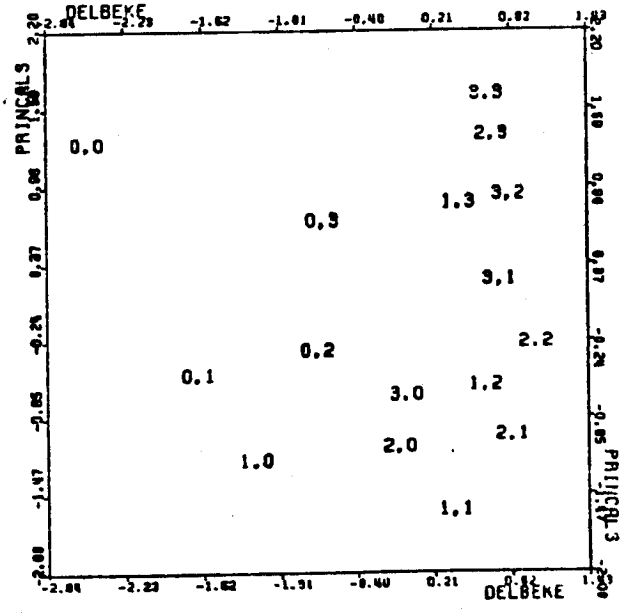
NUMERICAL TWO DIM.



ORDINAL FIRST TWO OF TWO DIM.



ORDINAL FIRST TWO OF THREE DIMENSIONS



ORDINAL FIRST TWO OF FOUR DIMENSIONS

figuur 10.8. PRINCALS oplossingen voor Delbeke data.

hiervan staat in figuur 10.8. Het betreft voorkeuren voor familiesamenstelling (Delbeke, 1978); de objekten zijn 16 verschillende kombinaties van (aantal zonen, aantal dochters), lopend van (0,0) naar (3,3). De individuen zijn 82 studenten van de universiteit van Leuven, die in 't algemeen een vrij grote 'number bias' hebben en een lichte 'boy bias'. In de PRINCALS analyse komt dit naar voren bij de numerieke oplossing (links boven), waar het merendeel van de individu-pijlen binnen de gearceerde kegel valt.

Het opmerkelijke van de ordinale oplossingen in 2 en 3 dimensies (rechts boven en links onder) is, dat naast de algemene afkeer van (0,0) alleen 'kinderaantal' nog differentieert tussen individuen. Het programma heeft geprofiteerd van z'n vrijheid, transformaties met zeer grote stappen te maken. Dit verschijnsel komt overeen met de in niet-metries ontvouwen zo bekende 'degeneratie', waarbij alle objekten op een cirkel terechtkomen, uitgezonderd het minst populaire (vgl Kruskal en Carroll, 1969). Het lijkt er op, dat deze tendentie meer kenmerkend is voor de intrinsieke zwakte van rij(kolom)-konditionele isotone regressie voor alle variabelen(individuen) dan voor 'het model'. Wanneer we PRINCALS in vier dimensies haar gang laten gaan (figuur 10.8 rechts onder), verdwijnt de degeneratie. Er zijn nu zoveel vrijheidsgraden, dat er geen grote stappen meer in de transformatie hoeven te worden gemaakt. De eerste twee assen van de aanvangschatting (links boven) blijven vrijwel ongemoeid.

10.4. HOMALS en MSA.

Multidimensionale Scalogram Analyse (MSA) is een serie technieken van Guttman en Lingoes (Lingoes, 1968; voor een geautoriseerde meer recente bespreking, zie Zvulun, 1979), die bedoeld is als een zeer algemene manier om indikator matrixen in de euklidiese ruimte af te beelden. Wij beschouwen het hier als een generalisatie van HOMALS, die kategorieën afbeeldt als zgn. aaneengesloten gebieden (contiguous regions) in plaats van als punten; deze aaneengesloten gebieden worden op iteratieve wijze op basis van een HOMALS-type konfiguratie gevonden en zijn zeer vrij van vorm.

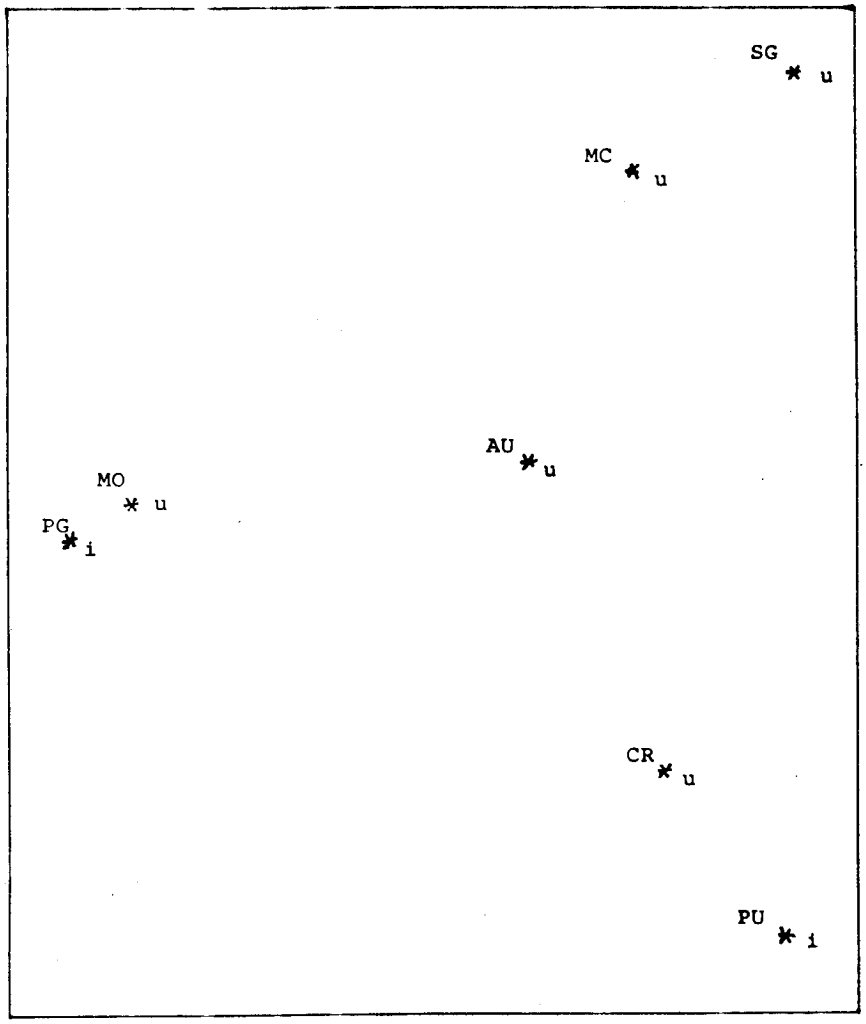
De grenzen van een aaneengesloten gebied in MSA-I zijn geen vlakken, bollen of kegels (zoals in de modellen genoemd in 10.0),

maar worden gedefinieerd in termen van inwendige punten (inner points) en uitwendige punten (outer points). De individu-punten die behoren bij een bepaalde kategorie van een variabele kunnen inwendig óf uitwendig zijn; het intuitieve idee van een aaneengesloten gebied is, dat inwendige punten worden omsloten door uitwendige punten van hun eigen kategorie. We zullen echter zien, dat de operationalisatie van aaneengeslotenheid niet noodzakelijkerwijs tot puntenwolken leidt die 'dicht' zijn.

Hoe worden de inwendige en uitwendige punten van een kategorie bepaald? We beginnen met de uitwendige punten van, zeg, kategorie a. Deze worden bepaald door achtereenvolgens voor alle individu-punten die niet tot a behoren te kijken wat het dichtstbij zijnde a-punt is. Deze a-punten zijn dus uitwendig. De andere a-punten zijn inwendig. De individu-punten behorend bij één kategorie van één variabele zijn aaneengesloten als en alleen als elk inwendig punt dichter bij een uitwendig punt van dezelfde kategorie ligt dan bij een uitwendig punt van een andere kategorie. De techniek streeft naar een zodanige verschuiving van individu-punten, dat de aaneengeslotenheid van alle kategoriewolken van alle variabelen zo groot mogelijk is.

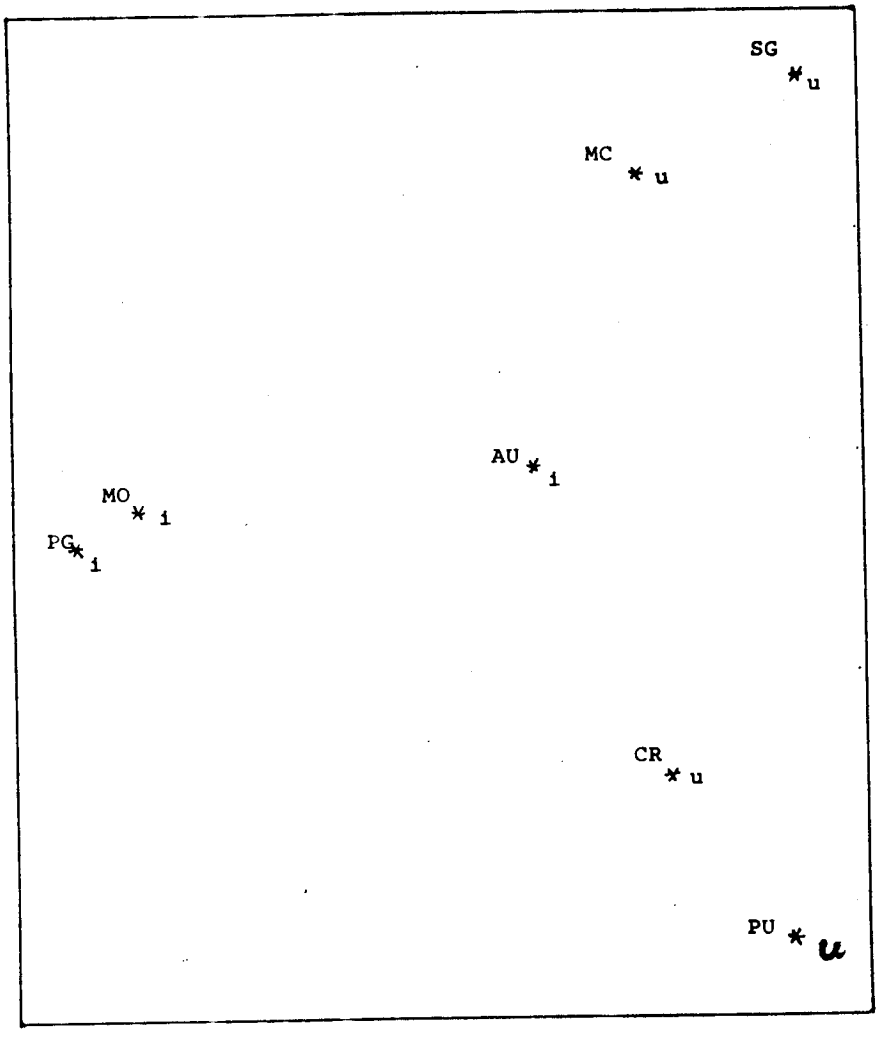
De start van MSA-I wordt geleverd door de individu-punten van HOMALS. Natuurlijk niet echt HOMALS, de procedure die gebruikt wordt heet MAC-II (van Multivariate Analysis of Contingencies). Het is een eigenwaarde implementatie van de formules uit Guttman (1941). We bekijken nu of de oplossing voor de GBS-data (zie hoofdstuk 3.1) MSA-I-aaneengesloten is. Voor variabele INTE is dit getekend in figuur 10.9. De kategorieën van INTE zijn duidelijk aaneengesloten. De enige twee inwendige punten liggen op de rand van de plot. Maar ook bijvoorbeeld PHYS, waarvan de kategorie a hoort bij PU en SG, die toch niet bepaald dicht bij elkaar liggen, is MSA-I-aaneengesloten (zie figuur 10.10). De lezer kan gemakkelijk zelf verifieren, dat de HOMALS oplossing voor alle variabelen MSA-I-aaneengesloten is.

Dit verschijnsel, dat de MSA-I kontiguiteitseisen zó vrij zijn dat in feite de HOMALS oplossing er al aan voldoet, zijn we in vrijwel alle gepubliceerde MSA voorbeelden tegengekomen. Guttman



figuur 10.9. Inwendige (i) en uitwendige (u) punten voor variabele INTE van GBS data.

en Lingoes definieren een λ-coefficient die de mate zou meten, over kategorieën en variabelen, waarin aan aaneengeslotenheid wordt voldaan. Hun algoritme zou λ maximaliseren over X, via een soort gradient-methode; hierbij telt niet alleen het aantal schendingen, maar ook hun ernst mee. Het is echter niet erg duidelijk, wat de eigenschappen zijn van dit algoritme, behalve dat het een prima aanvangsschatting heeft en dus meestal in één iteratie klaar is. Terzijde merken we nog op, dat de maximale afmetingen van een probleem in MSA-I niet groter mogen zijn dan omstreeks 150 individuen en 150 kategorieën ($\sum k_j$). Twee voorname redenen hiervoor zijn zonder twijfel de eigenwaarde aanpak van MAC-II en het feit dat voor het bepalen van de inwendig- dan wel uitwendigheid van punten steeds alle $\frac{1}{2}n(n-1)$ afstanden tussen



figuur 10.10. Inwendige (i) en uitwendige (u) punten voor variabele PHYS van GBS data.

individu-punten moeten worden uitgerekend.

De overige programma's uit de MSA serie leggen restrikties op aan de kategoriegrenzen. We zijn hier weer terug bij de eenvoudige scheidingsvlakken: in MSA-II zijn de kategoriegrenzen gedefinieerd als cirkels of bollen (binaire ontvouwing dus, met 'rank images' i.p.v. isotone regressie) en in MSA-III lijnen of (hyper)vlakken. We noemen ze hier alleen voor de volledigheid; we hebben geen ervaring met deze programmatuur.

10.5. De radex.

De g-faktor theorie van Spearman was gebaseerd op een heel simpel idee: als tests 'hetzelfde' meten, in meerdere of mindere mate, en voor de rest 'iets heel anders', moet er een denkbeeldige test g bestaan die alle partiele korrelaties $r_{jk.g}$ nul maakt. De hypothese van één gemeenschappelijke faktor is aantrekkelijk, falsifieerbaar en inderdaad, gaat voor de meeste testskores niet op. Mathematies gezien is er geen enkel bezwaar om loodrecht op de eerste nog een faktor te 'trekken' en dat was wat de amerikaanse psychologen onder leiding van Thurstone tegen het eind van het interbellum enthousiast gingen doen.

Na de oorlog kwam Guttman, geinspireerd door z'n eigen werk aan de perfekte schaal, terug op de Spearman-hierarchie met het argument dat Multipele Faktor Analyse konseptueel helemaal niet eenvoudig meer is, niet falsifieerbaar is (doorgaan met trekken tot er 'genoeg' variantie verklaard is) en alleen maar bestaat uit een domme rekenpartij. Zijn alternatief was de radextheorie, wat staat voor 'radial expansion of complexity'. Dit is dus een theorie over de struktuur van de korrelaties tussen testskores, die in zekere zin de notie van hierarchie weer in ere hersteld.

Het grondidee is, dat tests kunnen verschillen in soort en in mate van complexiteit. Tests van dezelfde soort, bijv. numerieke vaardigheid, kunnen alleen verschillen in mate van complexiteit en zijn daardoor geordend als een kumulatieve schaal. In volgorde van complexiteit gezet, vereist elke volgende test alle vaardigheden van de voorafgaande, plus iets meer. Een batterij tests die een dergelijk type ordening vertonen heet een simplex. Aan de andere kant kunnen tests van gelijke mate van complexiteit bekeken worden, die verschillen in soort; bijv. verbale vaardigheid, numerieke vaardigheid, motoriese vaardigheid etc. Hier stelt Guttman ook een ordeningsprincipe voor, maar niet van 'minst' naar 'meest'. De bedoelde ordening heeft geen begin of eind, is circulair. Aangrenzende tests op deze konseptuele cirkel hebben veel gemeen, verder van elkaar liggende minder. Een batterij tests die hieraan voldoen heet een circumplex.Meer in het algemeen zal een testbatterij uit tests bestaan die gelijktijdig variëren in soort en in mate van complexiteit, en deze vervlechting van simplexen en circumplexen heet de radex.

De radextheorie wordt op eenvoudige wijze uit de doeken gedaan in Guttman (1954); een meer op psychometrici toegespitste behandeling geeft Guttman (1955). Het onderscheid tussen het kumuleren van 'meer van hetzelfde' (genestheid van objekten) en het achtereenvolgens substitueren van 'steeds iets anders' (overlap tussen objekten) is vrijwel gelijk aan het in de experimentele psychologie bekende onderscheid tussen 'prothetiese' en 'metathetiese' kontinua (vgl. Stevens, 1957 en Restle, 1959); vrijwel, want het bijzondere van de circumplex is haar geslotenheid: er wordt uit een eindige verzameling met teruglegging en in een bepaalde volgorde geput.

Een recent interessant overzicht vanuit MDS gezichtshoek, met voorbeelden uit de perceptiepsychologie, is Shepard (1979). Deze bekent dat hij, zoals vele anderen, aanvankelijk de radex niet serieus wilde nemen. Maar de opkomst van MDS heeft het mogelijk gemaakt bevrijd te raken van de oogkleppen bestaande uit ortogonale gemeenschappelijke faktoren. En, zegt Shepard, plotseling zag ik ze ook op mijn eigen terrein: de kleurencirkel! het muziek-intervallen hoefijzer! de radex van de perspektiviese waarneming van roterende drie-dimensionale objekten!

Een van de redenen waarom wij ons hier met de radex bezighouden is weer z'n normatief karakter; het is een familie van strukturen waarvan je je kunt afvragen wat een techniek 'ermee doet'. Er zijn natuurlijk ingewikkelder strukturen te verzinnen (zie o.a. Foa, 1965, en Degerman, 1972, die onder meer een spherex en een ringex onderscheiden), maar het blijkt al niet eenvoudig te zijn, de radex goed te karakteriseren. En dit is nodig, want anders zitten we vóór dat we het weten met een soort Multipele Radex Analyse, waar alle datamatrixen van de wereld in passen. We beginnen daarom met de simplex en de circumplex, en bekijken daarna een aantal direkte generalisaties.

We gebruiken de volgende terminologie. De tests stellen we voor als een nx1 vektor van stochastiese variabelen $\underline{t}$, met $E(\underline{t})=0$. Verder onderscheiden we de mx1 vektor van 'elementaire' stochastiese variabelen of komponenten $\underline{c}$, met $E(\underline{c})=0$ en $E(\underline{cc}')=D$, waarbij D een diagonale matrix is met komponent varianties σ_j^2, en de nxm binaire selektiematrix S, die aangeeft welke tests door

door welke komponenten worden bepaald. Alle modellen die we zullen bespreken hebben de vorm

$$\underline{t} = S\underline{c}$$

We nemen hier dus geen 'unieke komponenten' of 'foutenmodel' aan. Het gaat ons eigenlijk alleen om S. Deze lijkt heel veel op een 'faktorpatroon-' of 'faktorstruktuur-matrix', maar is dat niet (hij bevat geen korrelaties of kovarianties, maar een soort geidealiseerde 0-1 versie daarvan; vandaar het neutralere woord selektiematrix). In het geval dat we ook nog aannemen dat de komponentvarianties gelijk zijn (het gelijk-gespreide geval, $D=\sigma I$, en dit zullen we verder ook aannemen want de essentie van de theorie is onafhankelijk van de formulering in termen van komponenten) zitten er helemaal geen parameters meer in het model. Er wordt alleen iets gezegd over de samenstelling van de variabelen, en, als S een bijzondere ordenende struktuur heeft, iets over hun ordening: "In scale analysis we were concerned with order among *people*. But how about a concept of order among *variables*? Is it possible to attach a meaning to a rank order among quantitative variables? If so, we can return to the notion of a hierarchy which has been abandoned by those seeking multiple common factors"(Guttman, 1954, pag.269).

De ordeningen van een simplex en een circumplex kunnen eenvoudig in S verpakt worden en de inter-test korrelaties staan dan volkomen vast. Immers,

$$C(\underline{t}\underline{t}') = E(S\underline{c}\underline{c}'S') = SS'$$

$$V(\underline{t}_i) = E(s_i'\underline{c}\underline{c}'s_i) = e_i$$

$$R(\underline{t}\underline{t}') = E^{-\frac{1}{2}}SS'E^{-\frac{1}{2}}$$

waarbij we $\underline{t}_i$ gebruiken voor de i-de variabele uit $\underline{t}$, s_i voor de i-de rij uit S, e_i voor de i-de rijsom van S, en het symbool E eerst voor verwachting, daarna voor de bekende diagonale matrix van randfrekwenties.

Bij elke S hoort dus een vaste korrelatie matrix. Als S een onderdriehoeksmatrix is (steeds wat meer bij wat er al was), vertoont R het bekende vanaf de diagonaal naar de hoeken af opende patroon (simplex). Als S weer vierkant (mxm) is en gedefinieerd als

$s_{ij} = 1$ als $i - j \leqslant k - m$

$s_{ij} = 1$ als $1 \leqslant i - j \leqslant k$

$s_{ij} = 0$ anders

dan is R een circumplex van de orde k (k is het aantal komponenten dat meedoet). Gegeven de rangorde van de komponenten, zijn er dus altijd m-1 circumplexen van verschillende orde (de tests overlappen elkaar meer of minder). Het korrelatie-patroon in R loopt weer vanaf de diagonaal, eerst af, dan op.

In de radex behoort elke test tegelijkertijd tot een simplex én een circumplex. Guttman (1954) formuleert dit niet expliciet in termen van S, maar merkt eenvoudigweg op:"Let us assume we have, ..., five different kinds of content and four different levels of complexity. This enables us to discern 5x4, or 20 elementary components" (pag.337). Dit zou betekenen dat we voor S moeten nemen zoiets als in tabel 10.1 staat. We hebben deze tabel over-

circumplex components	A				B				C				D			
levels of complexity	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
2	1	1	0	0	1	1	0	0	1	1	0	0	0	0	0	0
3	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
5	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
6	0	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0
test 7	0	0	0	0	1	1	1	0	1	1	1	0	1	1	1	0
8	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
9	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
10	1	1	0	0	0	0	0	0	1	1	0	0	1	1	0	0
11	1	1	1	0	0	0	0	0	1	1	1	0	1	1	1	0
12	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1
13	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
14	1	1	0	0	1	1	0	0	0	0	0	0	1	1	0	0
15	1	1	1	0	1	1	1	0	0	0	0	0	1	1	1	0
16	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1

tabel 10.1. Selektiematrix voor een radex met vier simplexen en vier circumplexen van de orde 3.

overgenomen uit van de Wollenberg (1974). Er zijn weer evenveel komponenten als mogelijke typen tests, maar de eenvoud, zowel in termen van S als van R, is diskutabel. Zijn de zo gegenereerde korrelaties nog wel eenvoudig weer te geven?

De gemakkelijkste manier om een radex af te beelden is d.m.v. een niet-metriese MDS techniek. Daarbij worden tests die hoog korreleren dicht bij elkaar in een euclidiese ruimte gezet, tests die lager korreleren verder van elkaar af. De numerieke waarden van de korrelaties doen dus niet ter zake, alleen hun onderlinge rangorde telt. Dit komt ons niet-parametries model goed van pas. Guttman (1964, 1966) behandelt zijn empiriese radexen ook op deze manier en zij zijn, ook tot z'n eigen verbazing, keurig twee-dimensionaal (simplexen worden stralen, circumplexen cirkels, de radex een soort van dart-board). Uit de studie van van de Wollenberg, die een groot aantal 'perfekte' gevallen met MDS heeft uitgetest, blijkt echter dat in 't algemeen voor korrelaties die gegenereerd zijn uit tabellen zoals 10.1 geen twee-dimensionale MDS representatie bestaat. Het verst in de richting komt nog een radex met vijf simplexen en drie circumplexen van de orde drie; de 15 tests komen dan op een cylinder te liggen. Er is wel het een en ander op zijn behandeling van ties aan te merken, maar dat doet vermoedelijk aan het globale resultaat niets af.

Wat we hier zullen doen, is een andere S nemen en die met HOMALS afbeelden in plaats van R met MDS. Voor de simplex weten we al wat er dan gebeurt: S is in dat geval een Guttman-schaal en zijn afbeelding een hoefijzer. Een aantal andere mogelijke selektiematrixen staan in tabel 10.2. Links staat de meest omvattende en toch nog eenvoudige struktuur: er is een (cirkulaire) rangorde van zes komponenten, en de tests verschillen zowel in het aantal komponenten wat bij hen hoort als in hun positie binnen de rangorde. Het is duidelijk dat de tabel is opgebouwd door alle mogelijke circumplexmatrixen van verschillende orde op elkaar te stapelen plus de patronen 'alles' en 'niets'. Maar het is ook niet moeilijk te zien, dat de tabel heel wat simplexen bevat: bijv. tests 1,2,8,14,20,26 vormen een simplex, evenals tests 1,6,12,18, 23,28, etc. In feite kunnen we, door steeds andere selekties uit de rijen te maken, $mx2^{m-2}$ simplexpatronen (hier: 96) in de tabel zien zitten die elkaar overlappen, of m afzonderlijke die alleen

	A	B	C	D	E	F	DIA	HAR	HOE	GUT
1	1	1	1	1	1	1	+			+
2	0	1	1	1	1	1	+			
3	1	0	1	1	1	1				
4	1	1	0	1	1	1				
5	1	1	1	0	1	1				
6	1	1	1	1	0	1				
7	1	1	1	1	1	0	+			+
8	0	0	1	1	1	1	+			
9	1	0	0	1	1	1				
10	1	1	0	0	1	1				
11	1	1	1	0	0	1				
12	1	1	1	1	0	0	+			+
13	0	1	1	1	1	0	+			
14	0	0	0	1	1	1	+	+	+	
15	1	0	0	0	1	1			+	
16	1	1	0	0	0	1			+	
17	1	1	1	0	0	0	+	+	+	+
18	0	1	1	1	0	0	+	+	+	
19	0	0	1	1	1	0	+	+	+	
20	0	0	0	0	1	1	+	+		
21	1	0	0	0	0	1				
22	1	1	0	0	0	0	+	+		+
23	0	1	1	0	0	0	+			
24	0	0	1	1	0	0	+			
25	0	0	0	1	1	0	+			
26	0	0	0	0	0	1	+	+		
27	1	0	0	0	0	0	+	+		+
28	0	1	0	0	0	0	+			
29	0	0	1	0	0	0	+			
30	0	0	0	1	0	0	+			
31	0	0	0	0	1	0	+			
32	0	0	0	0	0	0	+	+		

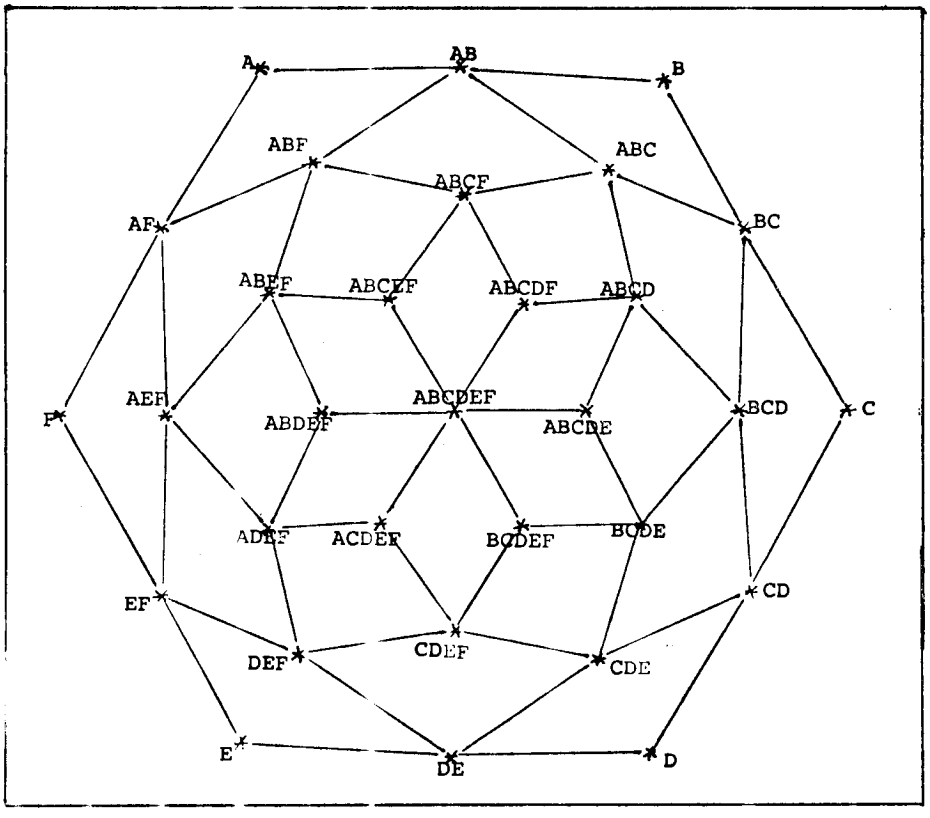
16 11 8 7 8 11
16 11 8 7 8
16 11 8 7
16 11 8
16 11
16

11 = 4+3+2+1 (+1)
8 = 4+2+1 (+1)
7 = 4+2 (+1)

tabel 10.2. Toegestane patronen van de ballon-schaal + 4 andere.

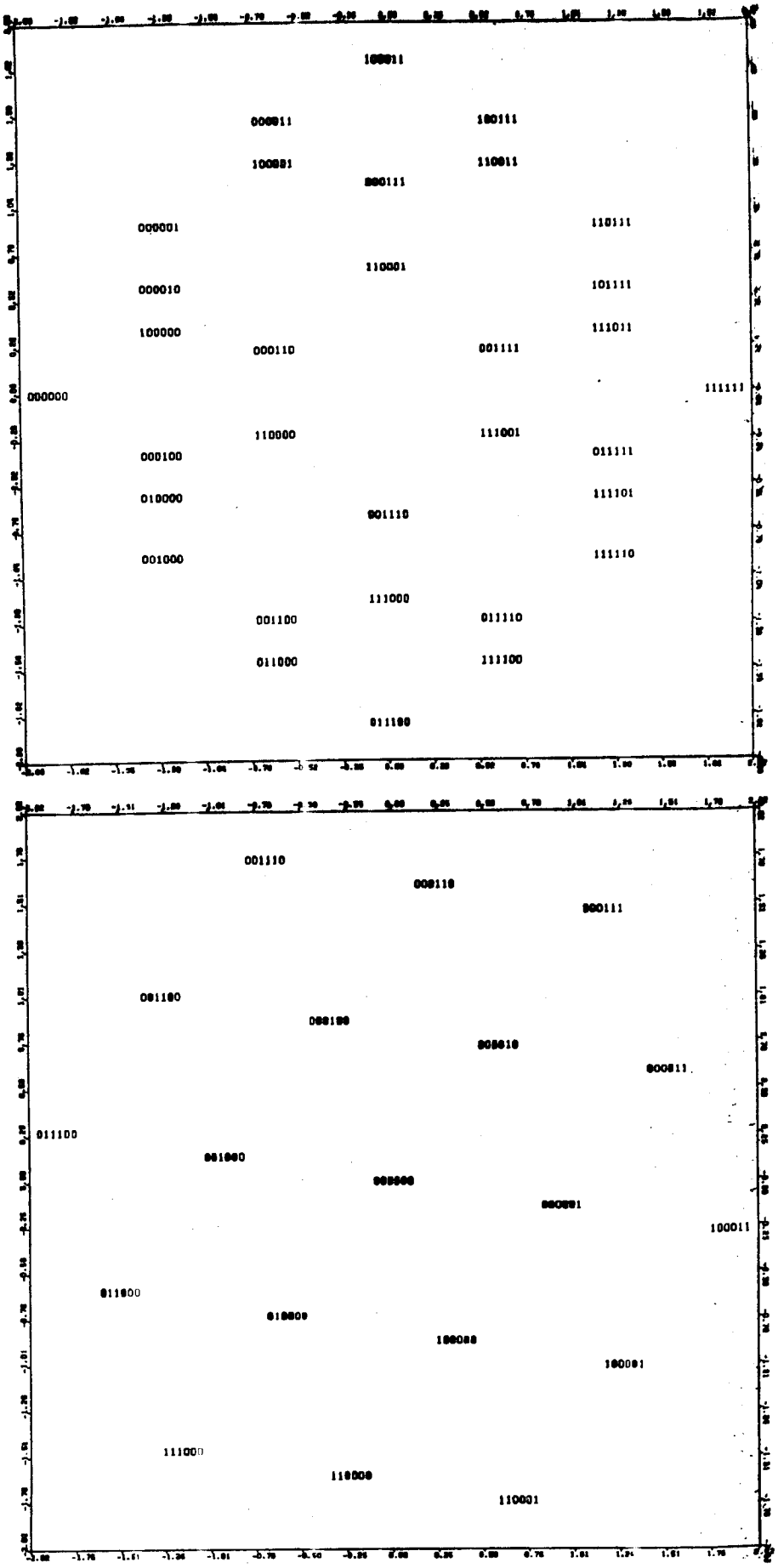
het eerste patroon gemeen hebben. Veel dus, maar nog altijd aanzienlijk minder dan het totaal aantal mogelijke rangordeningen m! (hier: 720). De tabel heeft nog een aantal andere interessante eigenschappen, zoals dat van elk rij-patroon ook diens komplement voorkomt, maar daar gaan we hier verder niet op in. Merk op dat genestheid en overlap nu gedefinieerd zijn in termen van dezelfde elementaire komponenten en niet, zoals in tabel 10.1, in termen van een produkt van twee soorten komponenten. Anders gezegd: er is één ordening van kolomobjekten en twee vervlochten typen ordening van rijobjekten.

We kunnen dit visualiseren door de graaf te tekenen die bij de

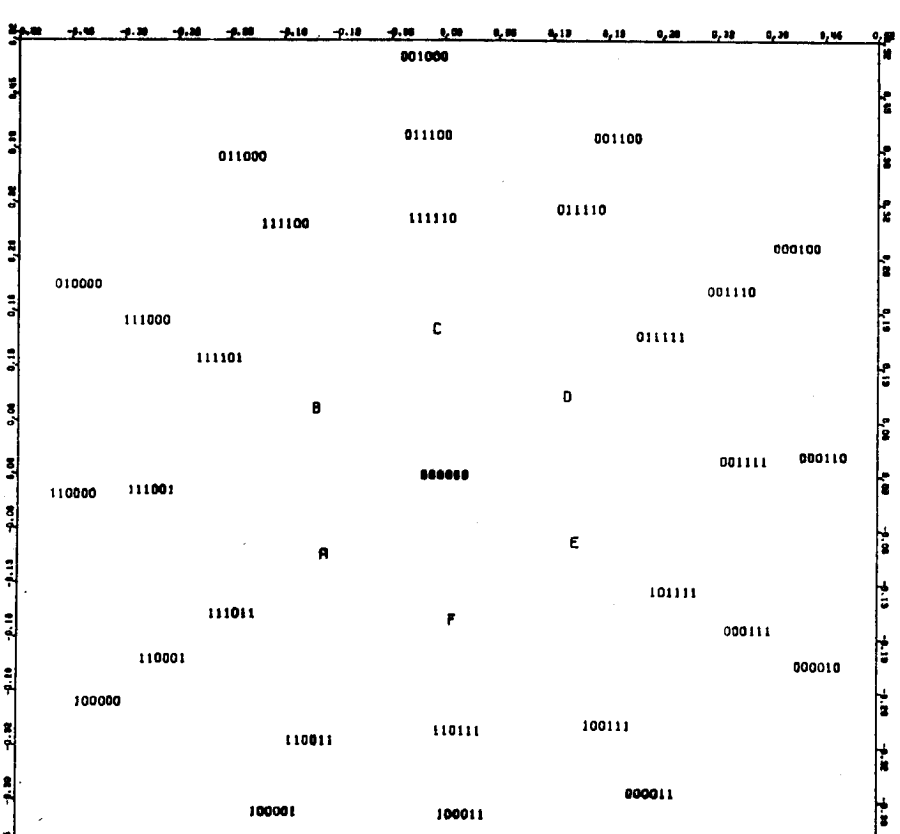


figuur 10.11. Graaf van de ballon-schaal.

rijen van tabel 10.2 hoort (zie figuur 10.11). We verbinden daarbij een patroon p met alle patronen q waarvoor geldt, dat q alle enen heeft van p plus precies één extra (het patroon 000000 laten we hier even weg; we hebben de patronen in de graaf gelabeld met de kolomsymbolen). Deze generalisatie van de Guttman-schaal noemen we - naar de vorm die hij in de euklidiese ruimte blijkt aan te nemen - de ballon-schaal. In figuur 10.12 zijn de individu-punten van de drie-dimensionale HOMALS oplossing voor de ballon-schaal geplot. De eerste as loopt van 'noordpool' (111111) naar 'zuidpool' (000000) en scheidt de vijf circumplexen van verschillende orde. De tweede en derde dimensie gunnen ons een kijkje op de ballon 'van boven'. De graaf is over een bol gespannen; simplex-patronen worden afgebeeld als 'meridianen', circumplex-patronen als 'parallellen'. Komplementaire patronen zijn elkaars 'tegenvoeters'. Binnen de ballon liggen de kategorie-punten (niet geplot); de 1-kategorieën op een cirkel binnen het noordelijk halfrond, de 0-kategorieën op een cirkel binnen het zuidelijk halfrond.

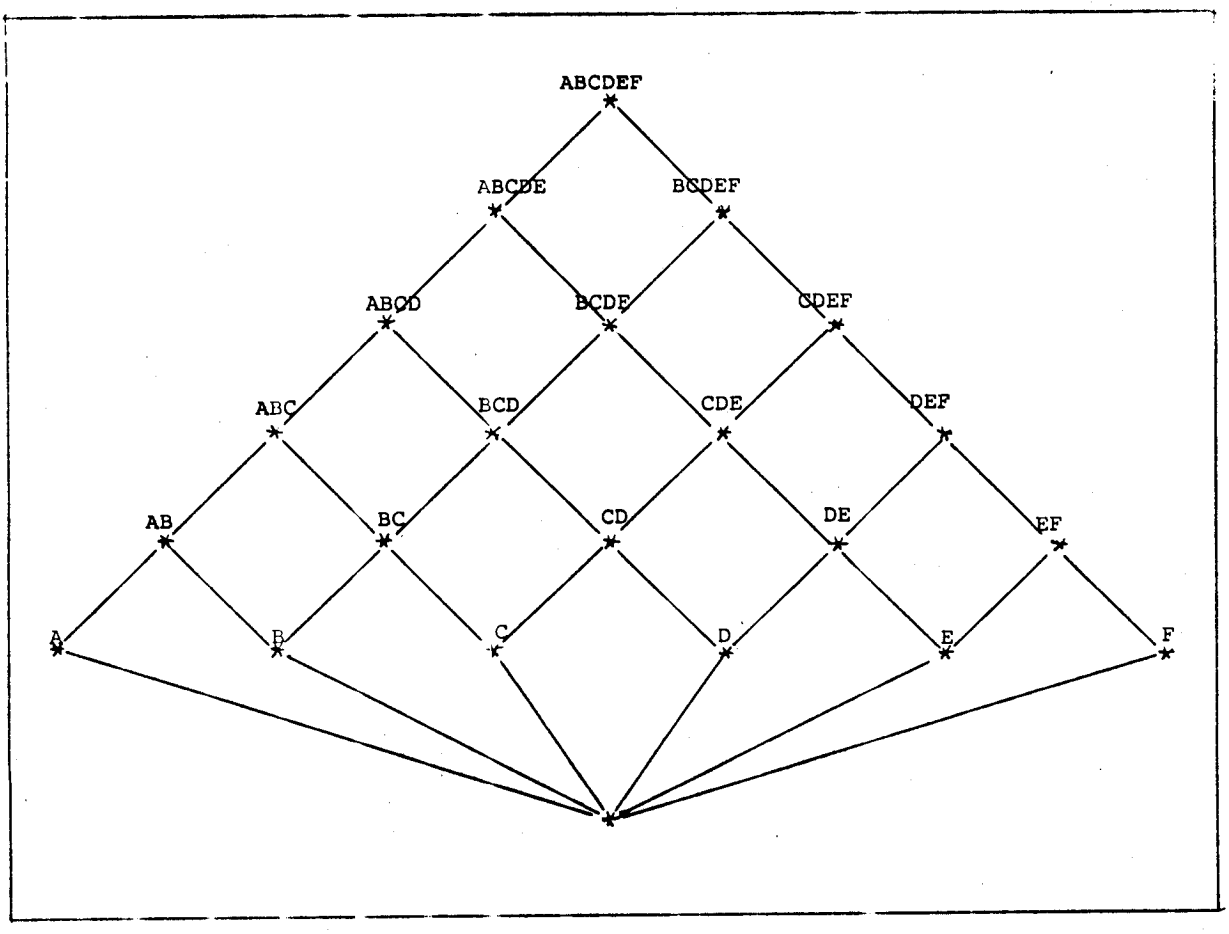


figuur 10.12. HOMALS individu-punten voor de ballon-schaal. Eerste 2 dim. boven, 2&3 onder.



figuur 10.13. SMACOF-3 oplossing voor de ballon-schaal. Rij- en kolompunten.

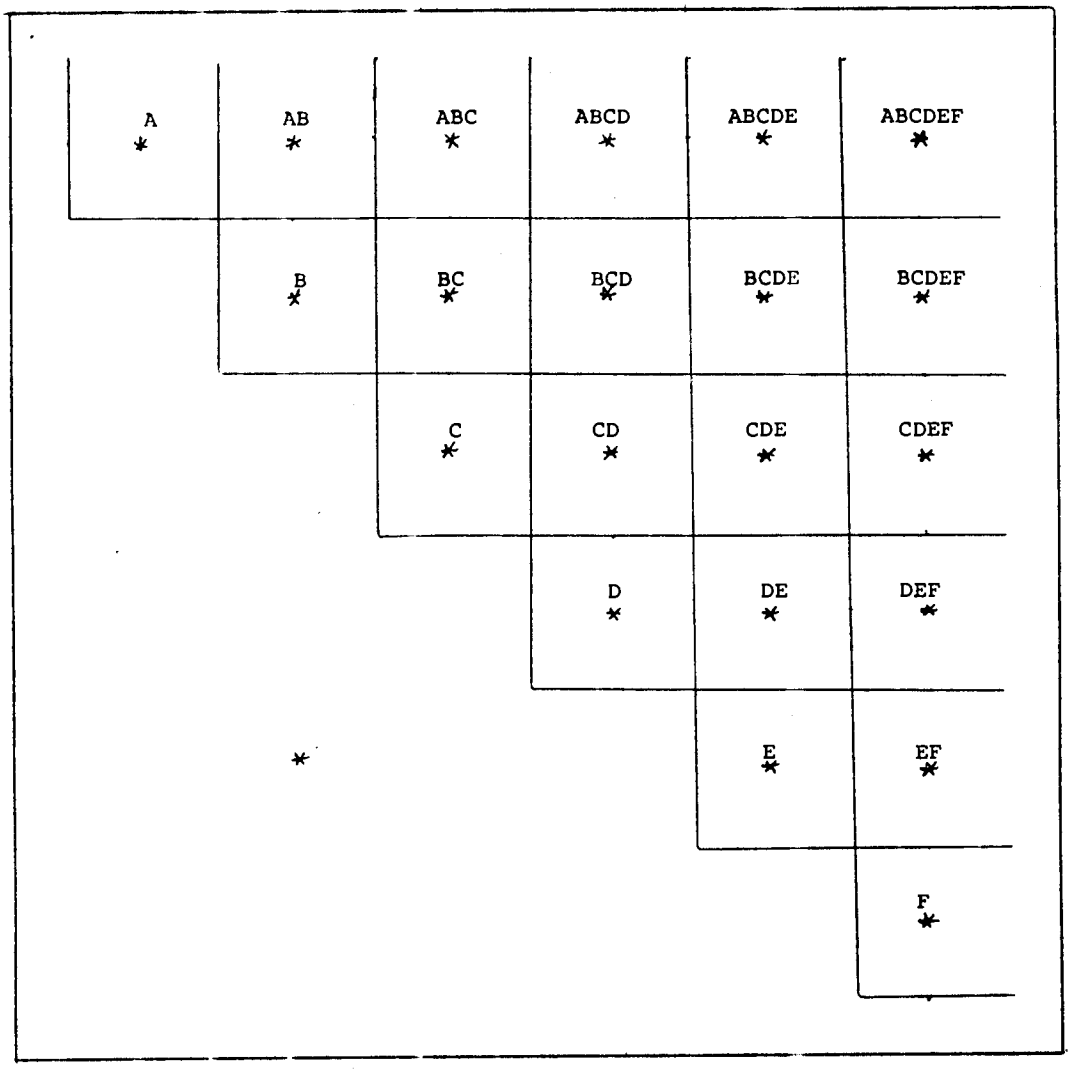
Wat doet nu een 'gewone' metriese ontvouwingstechniek met de ballonschaal? We hebben om dit uit te zoeken het programma SMACOF-3 (Heiser en de Leeuw, 1979) gebruikt; dit is ook een metries en onkonditioneel programma, maar geeft een enkele representatie van de kolom-objekten (dus niet één punt voor de 1-kategorie en één punt voor de 0-kategorie). Het lukt nu in twee dimensies (zie figuur 10.13), en de gelijkenis met de manier waarop we de graaf hebben getekend is frappant. De punten op het zuidelijk halfrond zijn naar de periferie verhuisd, terwijl de zuidpool als het ware is blijven liggen; hij veroorzaakt daardoor een groot gedeelte van het SMACOF-3 verlies. Niet-metries gezien is de representatie op de zuidpool na perfekt; d.w.z. het is mogelijk met een eenheidscirkel vanuit elk rij-punt precies de bijbehorende kolom-punten te omcirkelen. Zoals de HOMALS oplossing niet door MSA-I verbeterd zou kunnen worden, zo zal de SMACOF-3 oplossing



figuur 10.14. Graaf van de diamant-schaal.

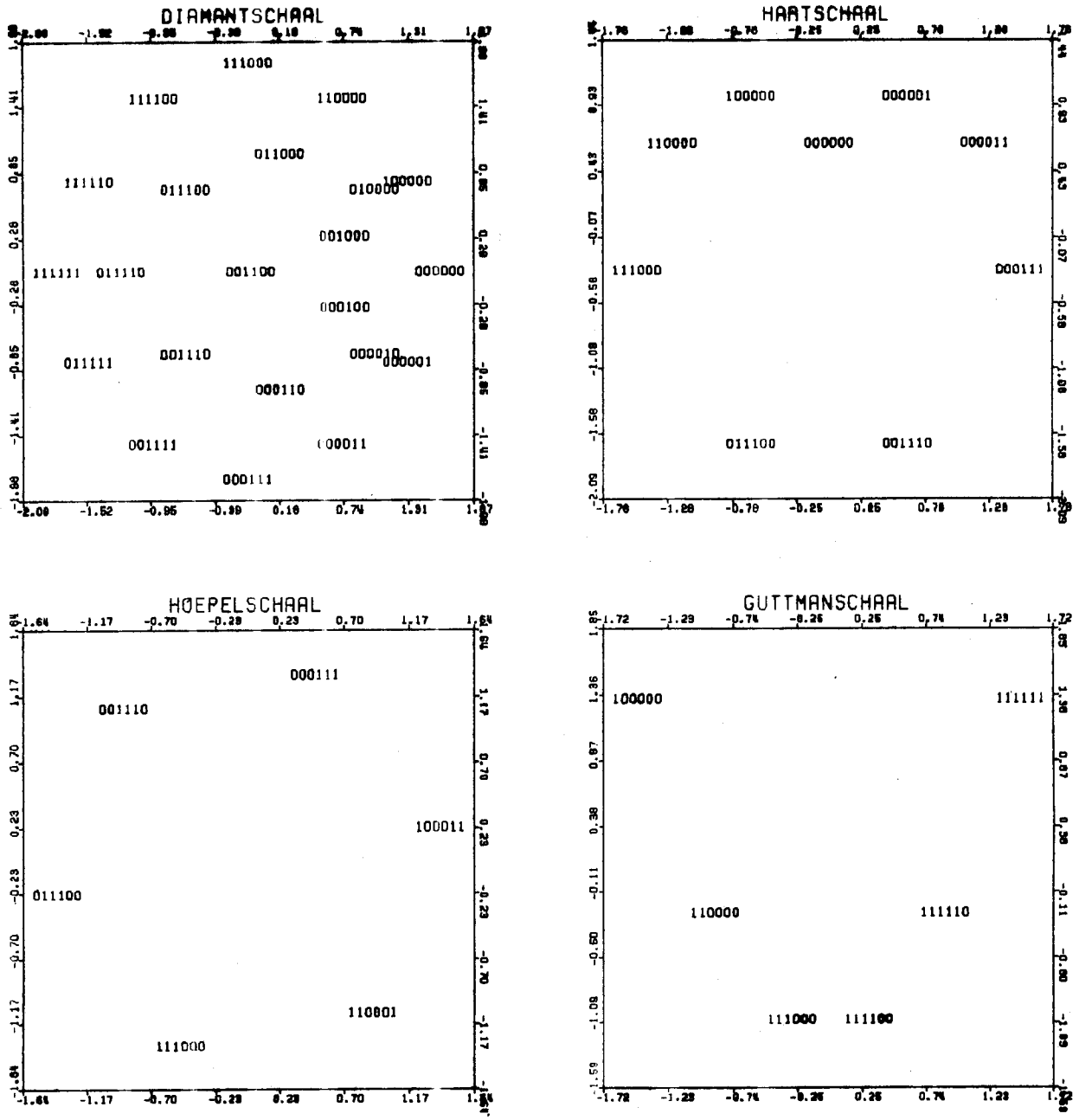
niet door MSA-II of een ander niet-metries programma verbeterd kunnen worden (behalve wellicht het verhuizen van de zuidpool naar oneindig).

We keren nu terug naar tabel 10.2, waar rechts een aantal deelverzamelingen van patronen zijn aangegeven, die we dus kunnen beschouwen als speciale gevallen van de ballon. De grootste is de diamant-schaal, een naam ontleend aan Shye (1978). De diamant heeft een aantal interessante eigenschappen van zichzelf. Z'n graaf staat getekend in figuur 10.14. In de eerste plaats is het een voorbeeld van een 'Guttman-vlecht' (Flament, 1971), d.w.z. de verzameling patronen is gesloten onder vereniging en intersektie (join en meet; de verzameling is een lattice). De gekozen patronen vertonen allemaal de zgn. 'opeenvolgende 1-en eigenschap' (Kendall, 1969; Tucker, 1972), die zo'n grote rol speelt bij seriatieproblemen in de archeologie (vgl. Kendall, 1963, 1969 en diverse artikelen in Hodson e.a., 1971). De kolomobjekten liggen



figuur 10.15. De diamant als konjunktief model.

nu niet meer op een cirkel, maar op een (uitgebogen, kromme) lijn. Dit suggereert onmiddelijk de interpretatie van een algemene één-dimensionale binaire Coombs-schaal, waarbij individuen verschillen zowel kwa ideaalpunt als kwa 'kritiese afstand' (vgl. hoofdstuk 2.2.7). In de ontwikkelingspsychologie kan de diamant model staan voor de opkomst, bloei en neergang van ontwikkelingsfasen (vgl. Coombs en Smith, 1973). Ook een interpretatie in termen van het konjunktieve model (vgl. hoofdstuk 10.0; een uitstekend overzicht is Coombs, 1964, hoofdstuk 12) is mogelijk. Dit staat getekend in figuur 10.15; we hebben gewoon de graaf een slag gekanteld en rechthoekige scheidingsvlakken aangebracht. Merk op dat de volg-orde van de scheidingsvlakken op de beide assen precies gespiegeld is; de diamant is dus een speciaal geval van het twee-dimensionale konjunktieve model. Hij bevat ook nog wel diverse Guttman-schalen,



figuur 10.16. De diamant en andere schalen volgens HOMALS.

waarvan twee tegengestelde in de volgorde van de kolom-objekten, maar geen volledige circumplexen meer.

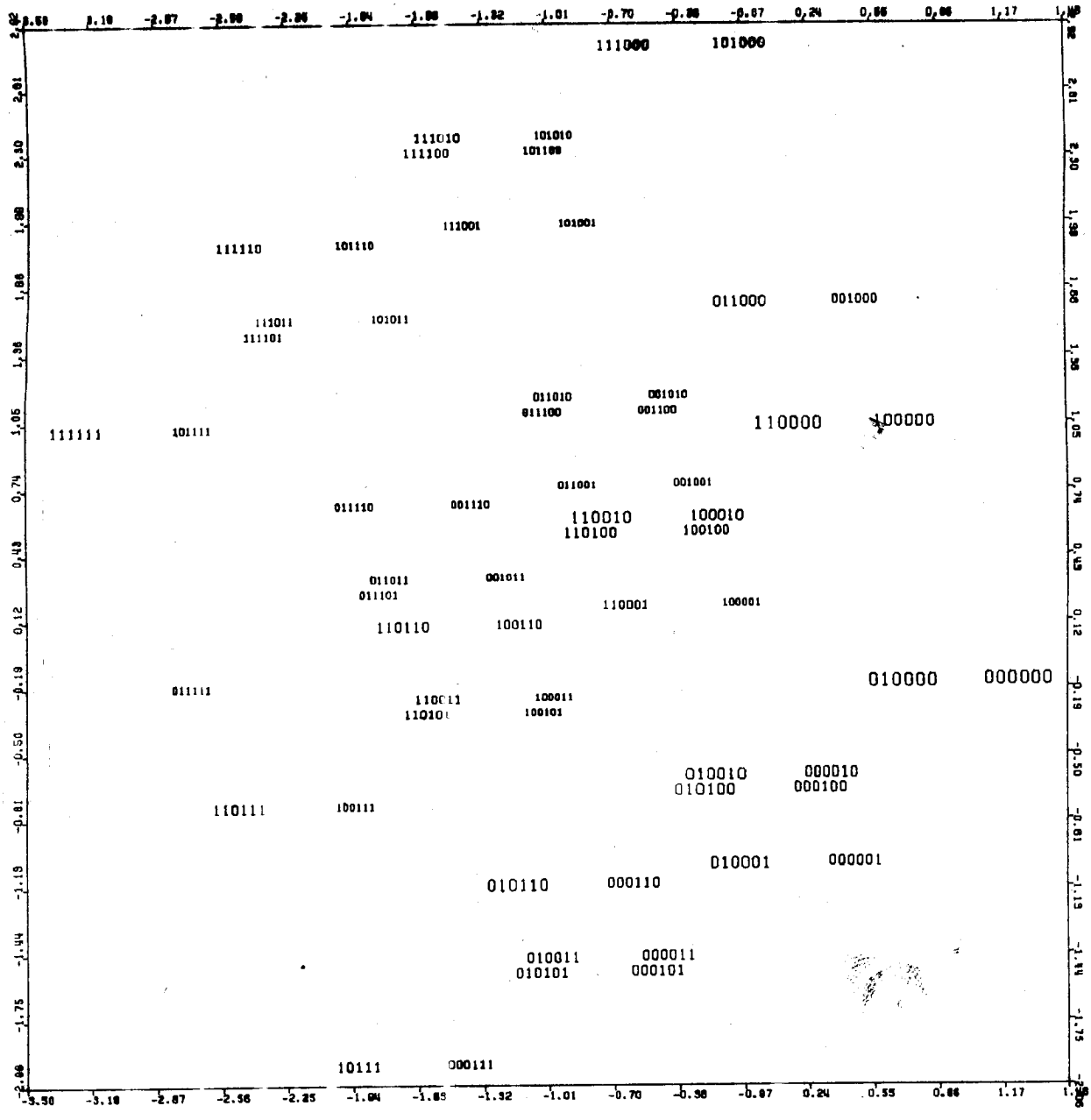
Er is nog een interessante deelverzameling die we de hart-schaal noemen, waarbij elk patroon k opeenvolgende 1-en heeft, of minder dan k indien er maar één 1-0 wisseling is. Deze struktuur kunnen we verwachten wanneer we de data opvatten als gedichotomiseerde afstanden uit een één-dimensionaal ontvouwingsmodel zonder idiosynkraties beslissingskriterium. We hebben voor de volledigheid ook nog één van de vijf hoepel-schalen en één van de 96 Guttman-schalen uit de ballon aangegeven. Wat HOMALS van dit alles maakt

staat in figuur 10.16. Alle schalen zijn stukken uit de ballon; de diamant is nu wel twee-dimensionaal, maar wordt als we het aantal komponenten zouden laten toenemen steeds boller, vooral bij z'n zuidpool-cirkel.

Alle tot nu toe besproken schalen opgevat als selektiematrix zijn toespitsingen van het simpele struktuur konsept van Thurstone: "The combination of a test configuration and the coordinate axes is called a structure. The coordinate axes determine the coordinate planes. If each test vector is in one or more of the coordinate planes, then the combination of the configuration and the coordinate axes is called a simple structure. The corresponding factor pattern will then have one or more zero entries in each row. If a test vector lies in one of the coordinate planes in a three-dimensional configuration, then it can be described as a linear combination of two coordinate vectors, so that one of its factor loadings is zero. If a test vector lies in two of the coordinate planes in a thr-dimensional problem, then it is collinear with one of the coordinate axes, and it will have two zero factor loadings" (Thurstone, 1947, pag. 181). De kruks van dit konsept is volgens Thurstone, dat het onthult dat de individuele tests minder komplex zijn dan de testbatterij als geheel. De kruks van de radex theorie is: multipele faktor analyse is niet simpel genoeg, simpele struktuur is niet simpel genoeg, schaalbare simpele struktuur is simpel. Dit betekent dat niet alleen de komplexiteit van de individuele tests, maar ook die van de testbatterij als geheel gering is.

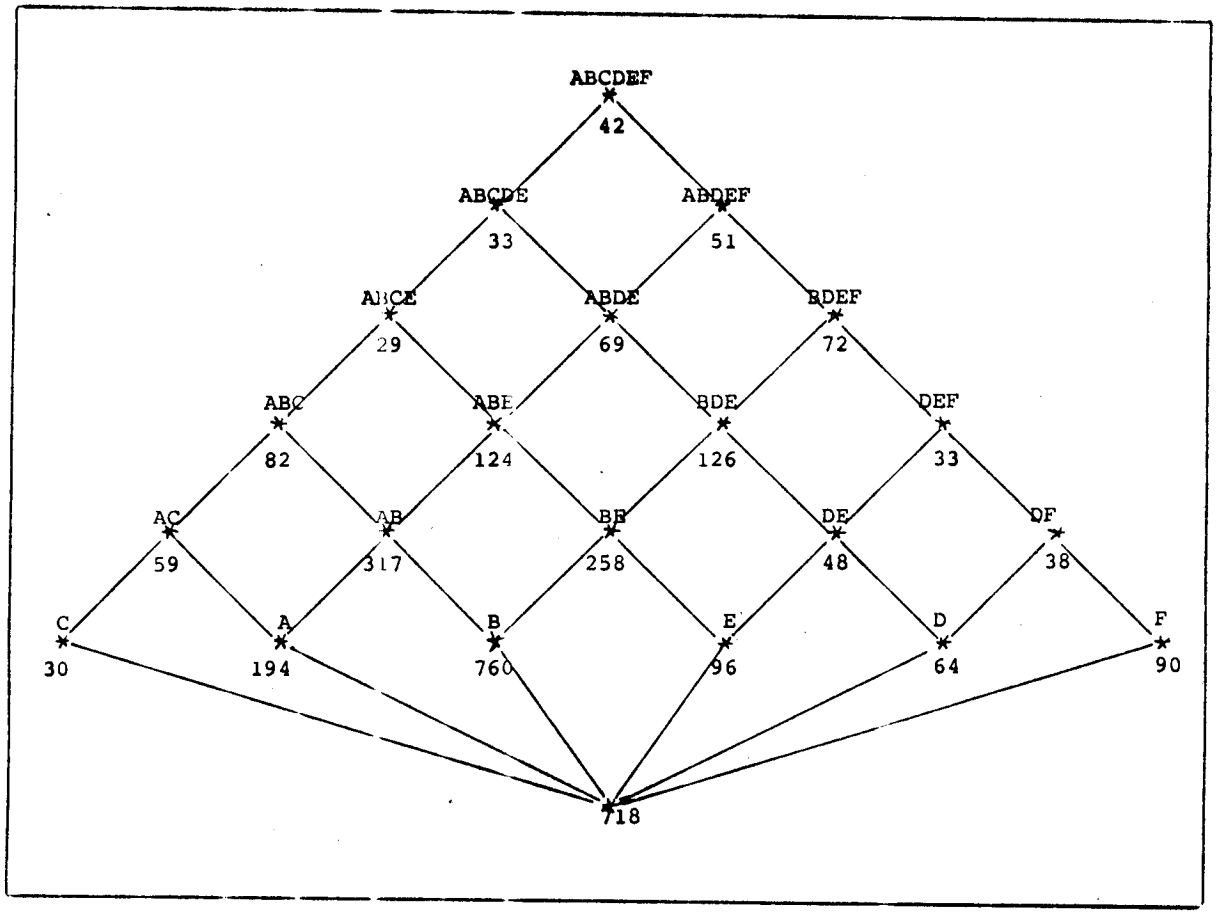
Een praktiese konsekwentie van het bovenstaande zou kunnen zijn, dat men niet een wezenloze arbitraire rotatie procedure zoals VARIMAX, maar een niet-lineaire analyse zoals ANACOR op de matrix van faktorladingen moet doen. Maar wie daartoe bereid is, zal wellicht gelijk ANACOR op de ruwe gegevens of MDS op de korrelaties toepassen.

Komt de ballon-schaal ook in de werkelijkheid voor? We geven twee voorbeelden. In figuur 10.17 zijn de door ANACOR gevonden profiel punten van de Japanners (vgl. hoofdstuk 3.2) nog eens geplot; we hebben hier echter de hoogte van de labels aangepast aan de frekwentie waarmee de profielen voorkomen. Dit om uit te laten komen,



figuur 10.17. Profiel-punten van de Sugiyama data (ANACOR).

dat we hier met een ruwe diamant te maken hebben, die als graaf is weergegeven in figuur 10.18. Als we, beginnend bij 111111, de lijnen van de graaf in de plot zouden tekenen vinden we niet alleen de raster struktuur terug, maar blijken we ook steeds alleen infrekwente patronen te hoeven overslaan. Het totaal aantal patronen dat we zo 'perfekt verklaren', is 3333 oftewel ruim 78%. Alleen aan de onderkant is het een beetje wringen geblazen, omdat daar een aantal 'foute' patronen liggen die vrij veel voorkomen. Aan de frekwenties die bij de knopen van de graaf vermeld zijn kunnen we zien, dat ze aan de randen een lichte neiging



figuur 10.18. Graaf van diamant patronen in Sugiyama data.

tot wiebelen hebben, maar verder doen denken aan een patroon uit de driehoek van Pascal.

De'enkele'patronen geven tevens de volgorde van de items aan: vanuit het midden naar links GRAVE, PRACT en BOOKS steeds toegewijdere praktijken; naar rechts MASCO, SUCCE en FORTU, steeds bijgeloviger praktijken. Dit patroon komt overeen met wat we vonden met PCA (vgl. hoofdstuk 3.2.2.) op de 2^e as.

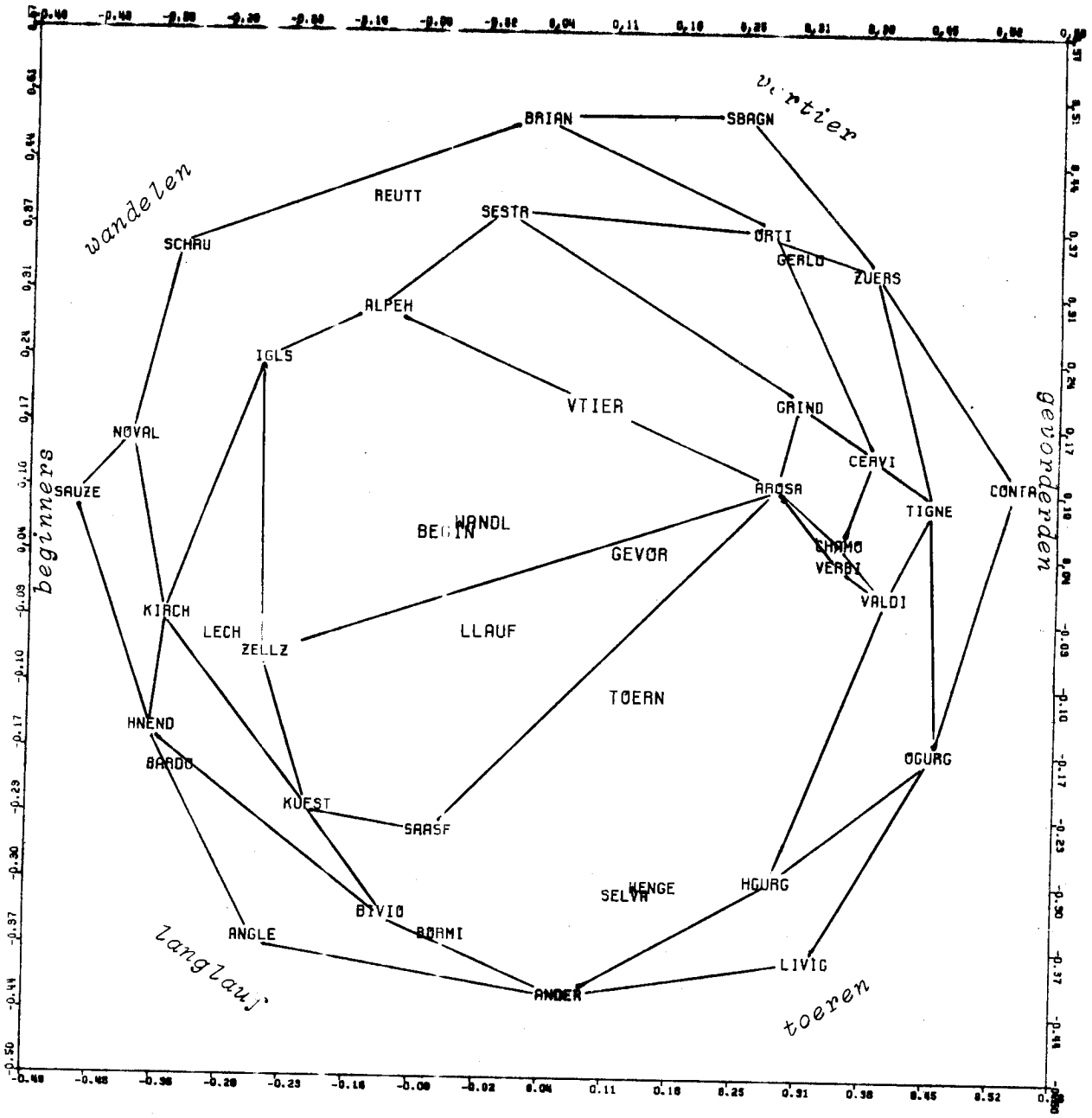
Voor een voorbeeld van de volledige ballon-schaal hebben we een beroep gedaan op de ANWB Wintersport gids uit 1978. Hierin staan, voor een groot aantal wintersportplaatsen, onder meer kwalifikaties op de volgende zes aspekten: geschikt voor beginners (BE, BEGIN), goede wandelmogelijkheden (WA, WANDL), veel vertier (VE, VTIER), geschikt voor gevorderden (GE, GEVOR), geschikt om te toeren (TO, TOERN) en veel langlauf mogelijkheden (LA, LLAUF). De gegevens van 115 middelmatig grote tot grote plaatsen staan in tabel 10.3. We hebben deze tabel zó geordend, dat de ballon struktuur gelijk

	BE	WA	VE	GE	TO	LA	Overige plaatsen
Arosa	1	1	1	1	1	1	Berchtesgarden, Canazei, Innsbruck, Kitzbühel, Seefeld, Davos, Engelberg, Sankt Moritz, Villars, Sölden
Chamonix	0	1	1	1	1	1	Garmisch, Sankt Anton, Adelboden, Zermatt, Cormaiore
Verbier	1	0	1	1	1	1	-
Saas Fee	1	1	0	1	1	1	Lenzerheide, Ponte di Legno, Flims
Zell am See	1	1	1	0	1	1	Oberstdorf, Saalbach
Alpe d'Huez	1	1	1	1	0	1	Font Romeu, Megève, Cortina d'Ampezzo, Crans, Montana, Gstaad
Grindelwald	1	1	1	1	1	0	Lanslevillard
Val d'Isère	0	0	1	1	1	1	-
Kufstein	1	1	0	0	1	1	Hinterglemm, Serfaus, Einsiedeln
Igls	1	1	1	0	0	1	Dobbiaco, Winterberg(Sauerland), Saint Gervais, Valberg, Lermoos
Sestrières	1	1	1	1	0	0	-
Cervinia	0	1	1	1	1	0	Monte Bodone
Hochgurgl	0	0	0	1	1	1	Riezlern, Les Diablerets
Bivio	1	0	0	0	1	1	Gressonei, Hintertux
Kirchberg	1	1	0	0	0	1	Sankt Johann, Chateau d'oex, Les Getz, Orcières Villard de Lans, Filzmoos, Champéry, San Barnedino
Ortisei	0	1	1	1	0	0	Spind Leruv Mlyn
Tignes	0	0	1	1	1	0	-
Andermatt	0	0	0	0	1	1	Disentis, les Orres, Gaschurn
Haute Nendaz	1	0	0	0	0	1	Baqueira-Beret
Nova Levante	1	1	0	0	0	0	Combloux, Saint Lary, Amden
Briançon	0	1	1	0	0	0	Bayrischzell
Zürs	0	0	1	1	0	0	-
Obergurgl	0	0	0	1	1	0	Hirschegg, Mittelberg
Les Angles	0	0	0	0	0	1	Calavese, Bovec
le Sauze	1	0	0	0	0	0	Schladming, Fontcouverte
Schruns	0	1	0	0	0	0	Bressanone, Dornbirn, Wörgl
Superbagnères	0	0	1	0	0	0	Mellau
les Contamines	0	0	0	1	0	0	Saline d'Ulzio
Livigno	0	0	0	0	1	0	Macugnaga, Galtür, Gargellen
Bardonnecchia	0	1	0	0	0	1	Brunico, Vigo di Fassa
Bormio	0	1	0	0	1	1	Kandersteg, Mayerhofen, Lienz
Lech	1	1	0	1	0	1	Merano 2000
Reutte	0	1	1	0	0	1	-
Selva	0	1	0	1	1	1	Ischgl, Kaprun
Wengen	1	1	0	1	1	0	-
Gerlos	1	0	1	0	1	0	-

tabel 10.3. Karakterisering van 115 wintersportplaatsen.

duidelijk is; twee van de 31 (plaatsen met 000000 hebben we niet opgenomen) toegestane patronen bij deze volgorde van aspekten komen niet voor, en er zijn zeven 'verboden' patronen. Dit maakt dat 100 van de 115 (87%) plaatsen in de ballon passen.

We hebben deze tabel met het unfolding programma SMACOF3 geanalyseerd, om een zo plat mogelijke representatie te krijgen (zie figuur 10.19). De lijnen van de dominantiestruktuur zijn in de plot aangegeven, waardoor het regelmatige patroon goed herkenbaar is. De plaatsen waar 'van alles te doen is' (Arosa etc) zijn niet zoals we uit figuur 10.13 zouden verwachten midden tussen de kolom-punten terechtgekomen, maar iets naar rechts. De verboden patronen komen natuurlijk terecht in de buurt waar ze het meest thuishoren (bijv. Lech bij Kirchberg). De circulaire



figuur 10.19 Ontvouwing van 115 wintersportplaatsen ANWB.

ordening van de aspekten (BE-WA-VE-GE-TO-LA-BE) doet heel natuurlijk aan: LLAUF zit tussen BEGIN en TOERN in, TOERN is iets waar je LLAUF interesse/gebied en GEVOR techniek voor moet hebben, waar het geschikt is voor GEVOR en WANDL moet wel VTIER zijn (met vertier wordt hier kennelijk bedoeld bars, disco's en nachtclubs en niet het sportieve samenzijn in berghutten en onder de bomen naast het langlauf-spoor), etc. Naarmate de plaatsen meer naar de periferie liggen zijn ze meer gespecialiseerd of eenzijdiger. Dit voorbeeld van radiale expansie loopt dus van complex in het midden naar eenvoudig aan de rand; het is Guttman's oorspronkelijke radex idee binnenste buiten.

A P P E N D I X A

1: Eigenwaarden en eigenvektoren

1:1 Invariante richtingen

Stel A is een vierkante matriks van de orde m. Iedere vektor x in $\mathbb{R}^m$ heeft een beeld Ax, ook in $\mathbb{R}^m$. Over het algemeen zijn de vektor en zijn beeld natuurlijk verschillend. Het is interessant om te weten of A invariante vektoren heeft, dat wil zeggen vektoren zodanig dat Ax = x. Natuurlijk is x = 0 altijd triviaal invariant. Omdat Ax = x ook geschreven kan worden als (A - I)x = 0 zien we dat niet-triviale invariante vektoren bestaan als en alleen als A - I singulier is.

Omdat dit nogal speciale omstandigheden zijn, krijgen we een interessantere theorie als we zoeken naar invariante richtingen. Richtingen zijn ekwivalentieklassen van vektoren, waarbij twee vektoren tot dezelfde richting behoren als ze evenredig zijn. Geometrisch is een richting dus een lijn door de oorsprong. Bij een richting $r = \{y : y = \mu x\}$ behoort ook een beeld $Ar = \{y : y = \mu Ax\}$, wat in het algemeen weer een andere lijn door de oorsprong definieert. We zien dat Ar = r als en alleen als $Ax = \lambda x$ voor het een of andere getal λ. We zeggen in dat geval dat x een eigenvektor is van A, en dat λ de bijbehorende eigenwaarde is.

1.2 Oplossen van de vergelijking

Voor welke waarden van λ heeft de vergelijking $Ax = \lambda x$ een niet-triviale oplossing ? Door te herschrijven als $(A - \lambda I)x = 0$ vinden we weer dat er niet-triviale oplossingen bestaan als $A - \lambda I$ singulier is, oftewel dat $\det(A - \lambda I) = 0$. Deze laatste vergelijking is de karakteristieke vergelijking van A, we nemen aan dat het begrip determinant bekend is bij de lezer. Als dit niet het geval is, dan is het voldoende om maar gewoon even te geloven dat $\det(A - \lambda I)$ een polynoom van de graad m is, de karakteristieke polynoom van A, en dat deze polynoom dus m reele of komplekse wortels heeft. Die wortels zijn per definitie de eigenwaarden van A. Als λ_o een eigenwaarde is, dan vinden we door oplossen van de vergelijking $(A - \lambda_o I)x = 0$ de bijbehorende eigenvektor x_o. Berekenen van eigenvektoren en eigenwaarden kan men dus doen door de wortels te berekenen van een polynoom van de graad m, en door vervolgens m keer een stel homogene lineaire vergelijkingen op te lossen. Dat wil echter niet zeggen dat het in de rekenkundige praktijk ook zo gebeurt !

We beperken ons nu tot het speciale geval waarin A symmetrisch is (door de hele klapper heen zijn alle matriksen reeel). In een beroemd artikel uit 1829 bewijst Cauchy de stelling dat in dat geval alle eigenwaarden reeel zijn, en daardoor zijn ook alle eigenvektoren reeel. Een matriks kan natuurlijk gelijke eigenwaarden hebben, omdat het mogelijk is dat de karakteristieke polynoom meervoudige wortels heeft. Veronderstel dat A p verschillende eigenwaarden $\lambda_1 > \ldots > \lambda_p$ heeft, waarbij

λ_1 in totaal m_1 maal voorkomt, ... , λ_p in totaal m_p maal voorkomt. Dus $m_1 + \ldots + m_p$ is gelijk aan m, de orde van A. Als x_1 en x_2 eigenvektoren zijn behorende bij verschillende λ_1 en λ_2, dan geldt dat $x_1'x_2 = 0$. Eigenvektoren behorend bij verschillende eigenwaarden zijn ortogonaal. Als een eigenwaarde multipliciteit m_1 heeft, dan horen er bij die eigenwaarde m_1 lineair onafhankelijke eigenvektoren, die ook ortogonaal gekozen kunnen worden. Als we deze opmerkingen kombineren vinden we dat we A kunnen schrijven als $A = X\Lambda X'$, waarbij X de eigenvektoren en waarbij Λ de eigenwaarden van A bevat. Bovendien geldt $XX' = X'X = I$, en Λ is diagonaal. Een andere manier om dit te schrijven is $A = \lambda_1 Q_1 + \ldots + \lambda_p Q_p$, waarbij de matriksen Q_s symmetrisch, idempotent ($Q_s^2 = Q_s$), en van de rang m_s zijn. De relatie tussen de twee representaties is $Q_s = X_s X_s'$, waarbij X_s een ortogonale basis is voor de ruimte van m_s eigenvektoren behorend bij λ_s.

Een matriks is positief definiet als alle eigenwaarden positief zijn, en positief semi-definiet als alle eigenwaarden niet-negatief zijn. Een matriks is singulier als er ten minste één eigenwaarde gelijk aan nul is, en regulier als alle eigenwaarden ongelijk aan nul zijn. Als $A = X\Lambda X'$ regulier is, dan is $A^{-1} = X\Lambda^{-1}X'$, dus heeft de inverse van A dezelfde eigenvektoren als A, terwijl de eigenwaarden van de inverse van A de reciprokes van de eigenwaarden van A zijn. Op dezelfde manier kunnen we, als A positief semi-definiet is, de symmetrische matriks $A^{\frac{1}{2}} = X\Lambda^{\frac{1}{2}}X'$ definieren.

Als A positief semidefiniet is, dan is $\{y : y'Ay = 1\}$ een ellips (met middelpunt in de oorsprong). In het algemeen is $\{y : y'Ay = \mu\}$ een ellips voor alle $\mu > 0$, en de verschillende waarden van μ definieren koncentrische ellipsen. Stel $A = X\Lambda X'$ is de eigenwaarden-eigenvektoren ontbinding van A, dan schrijven we door de rotatie $z = X'y$ toe te passen de ellipsen als $\{z : z'\Lambda z = \mu\}$, een eenvoudiger vorm waarbij de hoofdasssen van de ellips langs de koordinaatassen vallen. We kunnen dus eigenvektoren en eigenwaarden gebruiken om een ellips naar hoofdassen te transformeren. Als we de gebruikelijke meerdimensionale puntenwolk als een stel konsentrische ellipsen zien (denk aan Galton's ontdekking van de bivariate normaalverdeling), dan is dat transformeren naar hoofdassen niets anders dan principale komponentenanalyse.

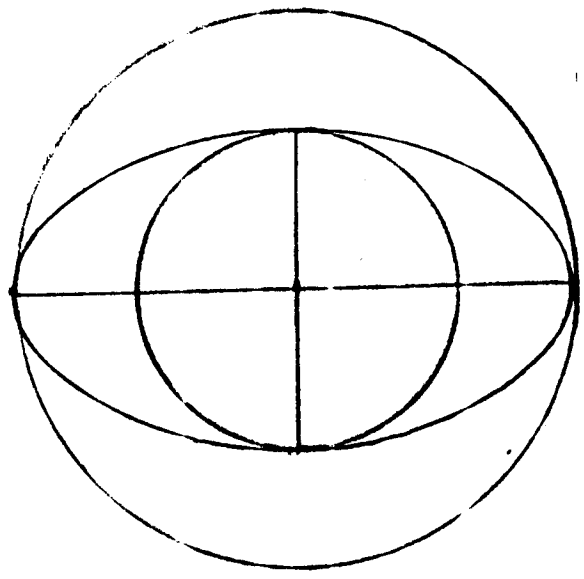
1.3 Het Rayleigh quotient

Er zijn nog wel andere manieren waarop men bij eigenwaarden en eigenvektoren belandt. Een van de eenvoudigste is het Rayleigh quotient, dat is een funktie $\lambda(x)$ van de vorm

$$\lambda(x) = \frac{x'Ax}{x'x},$$

waarbij A symmetrisch is. Door te differentieren vinden we de volgende stationaire vergelijkingen: $Ax = \lambda(x)x$, en dat zijn dezelfde vergelijkingen die eigenwaarden en eigenvektoren definieren. Dus: de grootste eigenwaarde van A is het maksimum van het Rayleigh quotient, de kleinste eigenwaarde het minimum. Veel van de

variantieverhoudingen uit de MVA zijn van dezelfde vorm als het Rayleigh quotient, optimalisatie van dat soort verhoudingen leidt daarom over het algemeen tot eigenwaarde-eigenvektor problemen.



We laten dit nog even met een plaatje zien. We zijn geinteresseerd in het maksimaliseren en minimaliseren van x'x op voorwaarde x'Ax = 1. Dus x moet op een ellips liggen (zie figuur). We willen dus de kortste en de langste vektor in die ellips vinden. We doen dat door een cirkel rond de oorsprong te trekken, en die cirkel vervolgens langzaam op te blazen. Waar hij voor het eerst de ellips snijdt, daar vinden we de kortste vektor. De kortste vektor korrespondeert met de grootste eigenwaarde, die immers gelijk is aan de reciproke van de lengte van de kortste vektor. Waar onze steeds groter wordende cirkel voor het laatst de ellips snijdt daar vinden we de langste vektor, en dus de kleinste eigenwaarde. Natuurlijk kunnen we het probleem ook aanpakken door de cirkel vast te nemen en de ellips op te blazen. Dan korrespondeert de langste vektor inderdaad met de grootste eigenwaarde, alleen wordt lengte gedefinieerd in termen van A.

1.4 Kleinste kwadratenbenadering

Stel A is een gegeven matriks, we willen nu een B vinden die in de eerste plaats van de rang p is, en die in de tweede plaats zo veel mogelijk op A lijkt. We vertalen dat door te zeggen dat we SSQ(A - B) willen minimaliseren over alle B van de rang p. Als $A = X\Lambda X'$ de eigenwaarde-eigenvektoren ontbinding is van A, dan blijkt de oplossing voor B te zijn $B = X_p \Lambda_p X_p'$, de afgekapte eigen-ontbinding die alleen de p grootste eigenwaarden en de bijbehorende eigenvektoren gebruikt.

1.5 Gegeneraliseerde inverse

Als Λ diagonaal is, dan definieren we Λ^+ als een diagonale matriks die nullen heeft op dezelfde plaatsen als Λ, maar die de niet-nul elementen van Λ vervangt door hun reciprokes. Als $A = X\Lambda X'$, dan definieren we de Moore-Penrose inverse van A als $A^+ \triangleq X\Lambda^+X'$. Deze definitie is overigens zeker niet de beste manier om A^+ te berekenen. De volgende eigenschappen zijn van belang: AA^+ is symmetrisch en idempotent, $AA^+A = A$ en $A^+AA^+ = A^+$.

2: Singuliere waarden en singuliere vektoren

2:1 Invariante richtingen

Stel nu dat A een $n \times m$ matriks is, zonder verlies van algemeenheid veronderstellen we dat $m \leq n$. A beeldt R^m in R^n af, met iedere y in R^m korrespondeert een beeld $x = Ay$ in R^n. Omgekeerd beeldt A', de getransponeerde van A, R^n in R^m af. Met iedere x in R^n korrespondeert dus een beeld $y = A'x$ in R^m. Dit suggereert de volgende definitie van invariante vektoren: een paar x in R^n en y in R^m is invariant als $Ay = x$ en $A'x = y$. Evenals in 1.1 bestaan invariante vektoren alleen onder zeer speciale omstandigheden, het is noodzakelijk en voldoende dan $A'A - I$ singulier is (een ekwievalente konditie is dat $AA' - I$ singulier is).

Evenals in 1.1 gaan we daarom (op dezelfde manier) over op invariante richtingen. We eisen daarvoor dat $Ay = \lambda x$ en $A'x = \mu y$, terwijl we zonder verlies van algemeenheid aan mogen nemen dat $x'x = y'y = 1$. Dit impliceert overigens direkt dat $\lambda = \mu$. We zeggen in dit geval dat y een rechtse singuliere vektor is van A, dat x een linkse singuliere vektor is van A (en een rechtse van A'), en dat λ een singuliere waarde is van A (en van A').

2:2 Oplossen van de vergelijking

Dit kan heel kort, gegeven 1:2. Uit $Ay = \psi x$ en $A'x = \psi y$ volgt immers onmiddellijk $AA'x = \psi^2 x$ en $A'Ay = \psi^2 y$. Dus de oplossingen voor x zijn de eigenvektoren van AA' en die voor y zijn de eigenvektoren van $A'A$. De singuliere waarden zijn de wortels van de eigenwaarden van AA', of, wat hetzelfde is, de wortels van de eigenwaarden van $A'A$.

Dit lijkt te impliceren dat we twee eigenwaardenproblemen op moeten lossen om singuliere waarden en vektoren te vinden, maar dat is natuurlijk niet zo. Als Y de eigenvektoren van $A'A$ zijn, met $YY' = Y'Y = I$, en $A'AY = Y\Lambda$, dan zijn de linkse singuliere vektoren van A gewoon $X = AY\Lambda^{-\frac{1}{2}}$. Als we linkse en rechtse singuliere vektoren en singuliere waarden kombineren tot $A = X\Psi Y'$, dan vinden we de singuliere waarden dekompositie van A. Sommige mensen denken dat die SVD (singular value decompositiion) iets heel recents is, maar die mensen hebben zoals gewoonlijk ongelijk. De SVD werd ontdekt door Beltrami in 1873, en onafhankelijk van hem nog door Jordan (1974) en Sylvester (1889).

Een alternatieve manier om de SVD te schrijven is $A = \sum \psi_s Q_s$, waarbij $Q_s'Q_s$ en Q_sQ_s' allebei symmetrisch en idempotent zijn, met rang gelijk aan de multipliciteit van de bijbehorende singuliere waarde. In termen van X en Y kunnen we schrijven $C_s = X_sY_s'$.

2:3 Zoiets als het Rayleigh quotient

Definieer

$$\rho(x,y) \triangleq \frac{x'Ay}{(x'x)^{\frac{1}{2}}(y'y)^{\frac{1}{2}}}.$$

Door te differentieren vinden we nu de stationaire vergelijkingen

$Ay = \rho(x,y)x,$

$A'x = \rho(x,y)y,$

en dat zijn precies de vergelijkingen die singuliere waarden en singuliere vektoren definieren. Dus: singuliere waarden en vektoren komen overeen met de stationaire waarden van $\rho(x,y)$.

Het is lastig om een plaatje te tekenen in dit geval omdat zelfs voor een 2×2 matriks A de ruimte van x en y al vier dimensionaal is.

2:4 Kleinste kwadratenbenadering

Stel A is een gegeven matriks, we willen weer een B vinden van de rang p die zo veel mogelijk op A lijkt. Dus we willen SSQ(A - B) minimaliseren. De oplossing is $B = X_p \Psi_p Y_p'$, de afgekapte SVD. Veel mensen schrijven deze stelling toe aan de heren Eckart en Young, die hem in 1937 in Psychometrika publiceerden. Ze vergeten daarbij dat Erhard Schmidt al in 1906 een veel algemenere stelling publiceerde.

2:5 Gegeneraliseerde inverse

Als $A = X\Psi Y'$, dan $A^+ = Y\Psi^+X'$. Het is duidelijk dat ook hier $AA^+A = A$ en $A^+AA^+ = A^+$. Bovendien zijn zowel A^+A als AA^+ symmetrisch en idempotent. Evenals in 1:5 noemen we A^+ de Moore-Penrose inverse.

A P P E N D I X B.1 VAN JAAR TOT JAAR

In 1965 is het Centraal Bureau voor de Statistiek begonnen met een doorlopende registratie van de schoolloopbaan van ruim 11.000 jongens en meisjes die in dat schooljaar het gewoon lager onderwijs hebben verlaten (een nationale steekproef uit de totale populatie van 217.000 leerlingen). Het CBS heeft, ten behoeve van een generatiestatistiek, van deze leerlingen de schoolgeschiedenis geregistreerd tot op het moment dat zij het volledig dagonderwijs verlieten. Voor al deze leerlingen geldt, dat zij in 1965 betrokken zijn geweest bij het zgn 'Generatieonderzoek 1965' ; in het kader van dit onderzoek is van de deelnemende leerlingen een aantal antecedenten verzameld, waaronder hun testgegevens op grond van de Nederlandse Onderwijs Differentiatie Testserie (NDT) .

Men heeft zich gerealiseerd dat een dergelijke uitvoerige registratie een unieke mogelijkheid bood tot verdergaand onderzoek, waarbij men dacht aan :

- een zich over een aantal jaren uitstrekkend onderzoek naar de achtergronden van de schoolkarrière, waarbij met name aandacht geschonken kon worden aan faktoren buiten en zo mogelijk ook binnen de school die op deze karrière van invloed zijn ;
- een overeenkomstig onderzoek naar het feitelijk verloop en de achtergronden van het beroepskeuzeproces dat wordt voortgezet nadat de leerling het volledig dagonderwijs heeft verlaten en dus niet meer door het CBS wordt geregistreerd.

In 1967 is een voorstel voor een dergelijk onderzoek uitgewerkt. Dit resulteerde in een gedetailleerd onderzoeksplan, dat is beschreven in een rapport getiteld : 'Van Jaar tot Jaar, Onderzoek naar de school- en beroepskarrière van jongens en meisjes die in 1965 het lager onderwijs verlieten , Onderzoeksvoorstel', (ITS 1968).

De oorspronkelijke opzet van het onderzoek was een 'panel-study' in vijf fasen : een proportionele steekproef uit bovengenoemde nationale steekproef zou gedurende tien jaar - gerekend vanaf 1965 - vijf maal ondervraagd worden, telkens met een tussenpoos van vijf jaar. De lange voorbereidingstijd maakte het onmogelijk deze onderzoeksopzet te realiseren. De eerste ondervraging van de nieuwe steekproef (die oorspronkelijk 2000 leerlingen en - zoals hier geanalyseerd - uiteindelijk 1845 leerlingen bevatte) vond plaats in 1970. In deze zgn 1^{e} fase van het 'Van Jaar tot Jaar' onderzoek werd deze gehele groep leerlingen samen met hun ouders mondeling geïnterviewd. De vragen hadden betrekking op de school- en beroepsloopbaan, op de verwachtingen en wensen van leerlingen en ouders, enz. Op dat moment bleek 34% van de proportionele steekproef nog secundair dagonderwijs te volgen. Deze groep werd in 1974 opnieuw benaderd (de zgn 2^{e} fase). Van 85% van deze groep werd een bruikbare set gegevens (vergelijkbaar met die uit 1970) verzameld. De gegevens uit de eerste en de tweede fase werden vervolgens door het ITS samengevoegd tot een geheel, zodat men van de gehele groep de manier waarop men het secundair dagonderwijs had doorlopen, kon analyseren [1)].

1) In de derde fase van het 'Van Jaar tot Jaar' onderzoek komt het beroepskeuze proces aan de orde. De resultaten van deze fase zullen pas in 1980 verschijnen.

Het feit, dat men in 1974 slechts een gedeelte van de proportionele steekproef heeft ondervraagd, kan tot vertekeningen leiden. Personen, die in 1970 niet meer aan het secundair dagonderwijs deelnamen, kunnen hier immers na 1970 weer aan begonnen zijn. De nadruk op dagonderwijs brengt ook met zich mee, dat allen die na 1970 suksesvol aan het secundair avondonderwijs deelnamen buiten beschouwing zijn gelaten. Ook degenen die in 1970 middelbaar of hoger beroepsonderwijs volgden zijn in 1974 niet meer geinterviewd (221 personen). Vooral voor diegenen die vanaf het lager beroepsonderwijs via het middelbaar beroepsonderwijs het hoger beroepsonderwijs bereikten en zo een relatief hoog onderwijsnivo haalden, brengt de gevolgde procedure een te lage 'inschaling' met zich mee.

Het gaat hier bij eindnivo dus uitsluitend om het hoogst bereikte nivo in het secundair dagonderwijs (met verwaarlozing van degenen die na 1970 hier suksesvol naar terugkeerden).

Een tweede beperking van het materiaal, die het misschien minder relevant kan maken voor de huidige situatie, is het feit dat de onderzoeksgegevens voornamelijk betrekking hebben op het lager en secundair dagonderwijs voor de invoering van de Mammoetwet. Ook werden de leerlingen voor de keuze gesteld verder onderwijs te gaan volgen of te gaan werken aanvankelijk bij een hoog-conjunctuur in de zestiger jaren en later bij een zich verdiepende economische crisis en een daarmee samenhangende toenemende (jeugd)werkloosheid in de jaren zeventig. Deze veranderingen van de situatie in de tijd zijn echter inherent aan een longitudinaal onderzoek : in de laatste fasen loopt men het risiko dat door strukturele en sociaal-economische veranderingen in de omstandigheden gegevens uit de beginperiode van het onderzoek niet meer relevant zijn.

Ondanks deze beperkingen zou het haast onvergeeflijk zijn dergelijk interessant en uniek materiaal niet verder te analyseren.

Tenslotte nog een overzicht van het tijdstip waarop de verschillende gegevens uit het VJTJ onderzoek verzameld zijn (wat niet hetzelfde hoeft te zijn als het tijdstip waarop de variabele betrekking heeft) .

1965 - BVA , URB , BIL , BIM , INT , DLO , LL6 , ADV , KGS , PRE , SEX

1970 - OPV , OPM , AKG , ASO , ASL , OOA , DWO , BMB , INS , KLS , TON , AOS , LLS , EXT

1970 }
1974 } - EIN

Deze beknopte beschrijving van het 'Van Jaar tot Jaar' onderzoek berust op het onderzoeksvoorstel (1968) , het verslag van de eerste fase (Kropman en Collaris 1974) en het verslag van de tweede fase (Collaris en Kropman 1978). Tevens is dankbaar gebruik gemaakt van de beschrijving van het onderzoek en van de beperkingen ervan alsmede van de kritiek op de gevolgde procedure door Dronkers (1978) .

OVERZICHT VARIABELEN UIT HET VAN JAAR TOT JAAR ONDERZOEK

BVA - beroepsnivo vader (CBS - indeling)

1 - landarbeid en ongeschoolde arbeid
2 - geschoolde handarbeid
3 - uitvoerende hoofdarbeid
4 - zelfstandige middenstand
5 - boeren en tuinders
6 - middenkader
7 - akademische vrije beroepen en leidinggevenden

OPV - opleidingsnivo vader

1 - alleen l.o. of v.g.l.o.
2 - alleen vakkursussen
3 - l.b.o.
4 - u.l.o. / m.u.l.o.
5 - m.b.o.
6 - v.h.m.o.
7 - h.b.o. of w.o.

OPM - opleidingsnivo moeder (zie OPV)

AKG - aantal kinderen in het gezin

1 - 1 , 2 - 2 , ... , 8 - 8 , 9 - meer dan 8

URB - urbanisatiegraad ouderlijke woongemeente (CBS , 1968)

1 - plattelandsgemeenten
2 - verstedelijkte plattelandsgemeenten
3 - plattelandsstadjes , kleine steden en middelgrote steden
4 - grote steden

ASO - aspiratienivo ouders (Reissman 1953 , zie Kropman en Collaris 1974 : D9)

Gevraagd werd of men vond dat zoon of dochter een 'fantastisch goede baan' zou moeten aksepteren als deze een van de volgende bezwaren met zich mee bracht :

1 - nachtdienst
2 - zware verantwoordelijkheid
3 - lange tijden van huis
4 - opnieuw een vak leren
5 - zeer onregelmatige werktijden
6 - altijd heel hard werken
7 - uit woonplaats weg
8 - vergt veel van gezondheid
9 - verlies van vriend(inn)en
10 - verlies veel vrije tijd
11 - verlies hobbies

Men kon steeds met een van de volgende kategorieën antwoorden :

1 - zeker niet nemen
2 - een groot bezwaar
3 - niet zo erg
4 - niet kunnen schelen

De skores voor de 11 vragen werden opgeteld. Zo ontstonden skores lopend van 11 tot 44 , deze werden ingedikt tot 6 klassen.

ASL - aspiratienivo leerling (zie ASO) . De items hadden hier betrekking op de ondervraagde zelf.

BIL - B.I.T.L.- schaal

BIM - B.I.T.M.- schaal

De B.I.T. is een bewerking van de Berufs Interessen Test van Irle door Wiegersma (1959). De test bestaat uit 162 items (beroepsmatige aktiviteiten). Op grond van deze items zijn 9 belangstellingsgebiedschalen gekonstrueerd en 2 nivoschalen , de L-schaal en de M-schaal, die gebaseerd zijn op aktiviteiten die respektievelijk geprefereerd worden door leerlingen die lager en middelbaar voortgezet onderwijs kiezen (Kropman en Collaris 1974 : 4.166). Deze laatste twee schalen zijn hier gebruikt. De B.I.T.L.-schaal is vooral samengesteld uit komponenten van de schalen Technische Handvaardigheid, Ambachtelijke Vormgeving, Voedselbereiding, Agrarische Arbeid en Handel. De B.I.T.M. schaal bestaat vooral uit komponenten van de schalen Techniek en Natuurwetenschappen en Literaire en Geesteswetenschappelijke Arbeid. De skores op alle schalen zijn voor jongens en meisjes apart omgezet in stanine-skores (dwz getransformeerd tot een 'normaalverdeling'). Dit maakt vergelijking van de skores van jongens en meisjes op zijn minst problematisch, Kropman en Collaris vinden het geen zin hebben (1974 : 4.167). Dit probleem zal echter vooral spelen bij schalen, waarbij er mbt de niet-getransformeerde skores een groot verschil was tussen jongens en meisjes. Wiegersma (1959) vond bij vergelijking van de oorspronkelijke skores van jongens en meisjes op de beide nivoschalen (itt tot andere schalen) weinig verschil, dwz jongens skoorden niet duidelijk lager dan meisjes of omgekeerd. Hoewel dit vergelijking van jongens en meisjes wb deze twee variabelen (schalen) minder 'zinloos' maakt, blijft de aparte normering problematisch bij het interpreteren van verschillen (vgl De Leeuw en Stoop 1979 , p. 16) .

INT - interesse ouders in de vorderingen van het kind volgens de onderwijzer (6^e klas)

1 - sterke belangstelling 2 - geen sterke belangstelling

OOA - openstaan ouders voor advies

DWO - doorzetten wens ouders

BMB - belang beroepskeuze meisje

Om de opvattingen van de ouders over school- en beroepskeuze te achterhalen werden aan een van hen tien uitspraken voorgelegd, die betrekking hebben op de mate waarin bij de school- en beroepskeuze rekening moet worden gehouden met de wensen van het kind zelf, de wensen van de ouders, het advies var de onderwijzer en de uitslag van een eventuele test (Kropman en Collaris 1974 : 4.83). Een analyse van de reakties op deze items bracht drie klusters aan het licht, die respektievelijk te beschouwen zijn als indikaties voor :

- de mate waarin de ouders aan het kind keuzevrijheid willen toestaan ten aanzien van school- en beroepskeuze;
- de mate waarin ouders hun eigen wensen met betrekking tot de school- en beroepskeuze van het kind willen doorzetten, desnoods tegen de wensen van het kind in (DWO) ;
- de mate waarin ouders van mening zijn dat school- en beroepskeuze gebaseerd moet zijn op, of zelfs wordt overgelaten aan het oordeel van de onderwijzer en eventuele testresultaten (OOA) .

Een naar inhoud hiervan losstaand item is buiten de kluster-analyse gehouden. De betreffende uitspraak luidt :

Het kiezen van een beroep is voor meisjes niet zo belangrijk, omdat ze na een korte tijd toch gaan trouwen.

1 - helemaal mee eens ... 4 - helemaal niet mee eens

In Dronkers (1978) en Dronkers en Jungbluth (1979) zijn als variabelen slechts het eerste en het derde kluster opgenomen, omdat het tweede niet signifikant samenhing met doorstroming naar het v.h.m.o. Kropman en Collaris vinden ook, dat wb dit kluster milieuverschillen niet tot uiting komen (1974 : 4.87). In het door het Steinmetz Archief aan ons ter beschikking gestelde materiaal was echter wel het tweede, maar niet het eerste kluster opgenomen. Dit werd pas in latere instantie door ons ontdekt.

De twee variabelen (klusters) zijn respektievelijk geskoord op een 4-puntsschaal (DWO : 1 - eens ... 4 - oneens met het doorzetten van de wens van de ouders) en 5-puntschaal (OOA : 1 - eens ... 5 - oneens met het moeten baseren van school- en beroepskeuze op het oordeel van de onderwijzer en het resultaat van een eventuele test).

INS - instemming van de ouders met de schoolkeuze van het kind

1 - instemming 2 - had het anders moeten doen

KLS - kleuterschool gevolgd

1 - gevolgd 2 - niet gevolgd

DLO - doubleerstatus op de lagere school

1 - niet gedoubleerd 2 - een of meer keer gedoubleerd

LL6 - aantal leerlingen in de zesde klas van de lagere school [1)]

1 - minder dan 10 3 - 20 - 30
2 - 10 - 19 4 - meer dan 30

ADV - advies onderwijzer voor schoolkeuze na lager onderwijs

1 - v.g.l.o. 3 - u.l.o.
2 - l.b.o. of l.t.b. 4 - m.m.s. of v.h.m.o.

1) Of van de opleidingsgroep binnen de zesde klas (vgl KGS)

KGS - gemiddelde zesde klas van betreffende leerling op enkele schoolvorderingentests, die in het kader van de Nederlandse Onderwijs Differentiatie Testserie zijn afgenomen. Wanneer de onderzoekspersoon heeft behoord tot een speciale 'opleidingsgroep' binnen een klas, heeft het gemiddelde van de schoolvorderingentests betrekking op die groep en niet op de hele klas.

1 - laag ..... 4 - hoog

PRE - prestatieskore v.h.m.o. (Zie Kropman en Collaris 1974 : 4.24)

Als indikatie voor de schoolkapaciteiten (uitgaand van het vigerend onderwijs systeem) van de ondervraagde leerlingen worden hier zgn. 'prestatieskores' gebruikt. Deze skores zijn gebaseerd op de Nederlandse Onderwijs Differentiatie Testserie (NDT), die aan alle onderzoekspersonen tijdens hun verblijf in de zesde klas van de lagere school is afgenomen. Deze testserie, waarin ook enkele school- en milieugegevens zijn opgenomen resulteert in 'prediktieskores', die de kans van slagen aangeven van de betreffende leerling op de verschillende soorten voortgezet onderwijs (hier alleen v.h.m.o.) . Omdat in deze prediktieskores verschillende soorten gegevens (kapaciteiten, schoolgegevens, gezinsgegevens) gekontamineerd zijn, zijn ze als indikatie voor kapaciteiten minder bruikbaar (vgl Kropman en Collaris 1974 : 4.107 ; Van Kemenade en Kropman 1972). Om aan dit bezwaar tegemoet te komen worden hier prestatieskores gebruikt, die gebaseerd zijn op enkele onderdelen van de NDT, waarbij expliciete school en milieugegevens zijn weggelaten. Opgenomen zijn testgegevens over 1) rekenen en cijferen, 2) geschiedenis, 3) snelheid en nauwkeurigheid en 4) progressieve matrices en de som van de rapportcijfers voor rekenen, taal, geschiedenis en aardrijkskunde.
Desondanks zijn ook de prestatieskores geen zuivere indikatie voor kapaciteiten in de zin van aanlegfaktoren. Op de testresultaten en rapportcijfers zijn school- en gezinskenmerken ongetwijfeld van invloed geweest.
De prestatieskores zijn net als de B.I.T.-schalen genormeerd, nl omgezet in stanine skores ; hier is echter niet apart genormeerd voor jongens en meisjes.

TON - tussentijds onderwijs nivo, dwz de eerste school, die de leerling na de lagere school heeft bezocht. Dit kan eventueel slechts voor een korte tijd geweest zijn.

1 - geen dagonderwijs 3 - l.b.o. / huishoudschool 5 - v.h.m.o.
2 - v.g.l.o. 4 - u.l.o.

AOS - andere opleidingen verbonden aan secundair dagonderwijs (1e school na l.o.)

1 - geen andere opl. 2 - wel andere opl.

LLS - aantal leerlingen school secundair onderwijs (1^e school na l.o.)

1 - minder dan 100 2 - 100-200 .... 7 - 600-700 8 - meer dan 700

EXT - aantal extracurriculaire aktiviteiten op school secundair onderwijs (1e school na l.o.). Het gaat hier om de aanwezigheid volgens de leerling van de volgende aktiviteiten / faciliteiten :

- schoolbibliotheek
- schoolklubs of verenigingen
- exkursies
- schoolkrant
- leerlingenraad of schoolparlement

Deze variabele geeft het aantal extracurriculaire aktiviteiten aan. Als er géén aanwezig zijn wordt dit gekodeerd als :

0 - als de onderzoekspersoon na het lager onderwijs geen verder dagonderwijs heeft gevolgd (dwz behandeld als missing) ;

6 - als de onderzoekspersoon wel aan het secundair dagonderwijs heeft deelgenomen, maar van geen enkele aktiviteit de aanwezigheid te kennen geeft.

In de overige gevallen kan deze variabele dus de waarde 1 t/m 5 aannemen.

EIN - hoogst bereikt onderwijsnivo (dwz eindnivo secundair dagonderwijs)

1 - alleen lo
2 - vglo zonder diploma
3 - vglo met diploma
lbo brugklas 1e jaar
4 - lbo 2e jaar
ulo/mavo en vhmo , brugklas of klas 1 of 2
5 - lbo verlaten uit klas 3 of 4
ulo/mavo/havo/mms klas 3
6 - lbo verlaten met diploma
7 - ulo/mavo/havo/havo/mms verlaten uit klas 4
overig vhmo uit klas 3
8 - ulo/mavo verlaten met diploma
9 - havo/mms zonder diploma uit klas 5
overig vhmo uit klas 4, 5 of 6
10 - havo/mms verlaten met diploma
11 - hbs verlaten met diploma
12 - gymnasium/atheneum verlaten met diploma

SEX - geslacht leerling

1 - man 2 - vrouw

A P P E N D I X B.2 ABORTUS

Inhoud : 37 variabelen , 575 personen :

CAP - 3 vragen over de doodstraf
AB - 16 vragen over abortus
EUT - 5 vragen over euthanasie
SF - 5 vragen over sexuele vrijheid
- 8 achtergrondsvariabelen

naam	nkat	missing	betekenis kat. (inhoud vragen , zie pag 2)
CP1 t/m CP3	5	0,9	(1) volkomen mee eens (5) helemaal niet mee eens
A01 t/m A08	2	0,9	(1) mee eens (2) mee oneens
A09 t/m A10	6	0,9	(1) tot 3 maanden (4) tot 6 maanden (2) tot 4 maanden (5) langer dan 6 maanden (3) tot 5 maanden (6) niet gerechtvaardigd
A11 t/m A14	5	0,9	(1) volkomen mee eens (5) helemaal niet mee eens
A15	3	0,4,9	(1) abortuswet, uitsluitend bijzondere gevallen (2) abortuswet, maakt a. moeilijk (3) geen wet, arts beslist of hij vrouw wil helpen
A16	3	0,4,9	(1) na 12e week absoluut verboden (2) na 12e week alleen in bijzondere gevallen (3) geen tijdsbeperking
EU1 t/m EU5	2	0,9	(1) geoorloofd (2) niet geoorloofd
SF1 t/m SF5	5	0,9	(1) volkomen mee eens (5) helemaal niet mee eens
SEX (geslacht)	2	0,9	(1) man (2) vrouw
AGE (leeftijd)	6	0,9	(1) LT 20 (2) 20-30 . . . . (6) GT 60
SOC (soc.klas)	8	0,9	(1) hoogste ..... (8) laagste
REL (gods - dienst)	4	0,9	(1) N.H. (2) Gereformeerd (3) RK (4) geen
POL (pol. partij)	5	0,9	(1) links (PvdA, PPR, D'66, PSP) (2) CDA (CDA, KVP, AR, CHU) (3) liberaal (VVD, DS'70) (4) rechts (GVP, SGP, BP) (5) geen
EDU (opleiding)	4	0,9	(1) lager + voortgezet lager (2) uitgebreid lager + uitgebreid lager en voortgezet (3) middelbaar + middelbaar en voortgezet (4) HBO , universiteit ed

FUN (funktie) 11 0

(1) direkteur, meer dan 10 pers.
(2) " minder dan 10 pers.
(3) zelfstandige vrije beroepen
(4) " boeren en tuinders
(5) hogere employees, ambtenaren
(6) midd. " , "
(7) lagere " , "
(8) geschoolde arbeiders
(9) ongesch. arbeiders + landarbeiders
(10) studenten
(11) huisvrouwen

URB (urbanisatie) 7 0,9

(1) A'dam + aggl.
(2) R'dam + aggl.
(3) Den Haag + aggl.
(4) middelgrote stad
(5) kleine stad
(6) geindustrialiseerd platteland
(7) agrarisch platteland

Inhoud vragen :

CP1 : Op het gijzelen van mensen zou de doodstraf moeten staan
CP2 : Moord zou met de dood moeten worden bestraft
CP3 : Het doden van mensen in tijd van oorlog is gerechtvaardigd

A01 : Ik vind abortus geoorloofd als het verder voldragen van de zwangerschap gevaar oplevert voor de gezondheid of het leven van de moeder
A02 : Ik vind abortus geoorloofd als een vrouw om een of andere reden het wil en er medisch gezien geen bezwaren zijn
A03 : Ik vind abortus geoorloofd als de kans groot is dat een misvormd of gehandicapt kind ter wereld wordt gebracht
A04 : Ik vind abortus geoorloofd als de aanstaande moeder niet gehuwd is en niet wil huwen met de betrokken man
A05 : Ik vind abortus geoorloofd als de vrouw ten gevolge van een aanranding of verkrachting zwanger is geworden
A06 : Ik vind abortus geoorloofd als de vrouw al een groot gezin heeft en meer kinderen niet gewenst zijn
A07 : Ik vind abortus geoorloofd als de aanstaande moeder niet gehuwd is en niet kan huwen met de betrokken man
A08 : Ik vind abortus geoorloofd als er een kans bestaat, dat het kind een ongelukkige jeugd tegemoet zal gaan, omdat zijn ouders niet echt van hem kunnen houden

A09 : Een vrouw van 45 jaar denkt bij het uitblijven van de menstruatie aan de overgang en maakt zich geen zorgen. Pas later blijkt dat zij zwanger is. Zij heeft een gezin met reeds volwassen kinderen. Tot welke maand van de zwangerschap - de 3^e , 4^e , 5^e , 6^e of langer - acht u abortus in dit bijzondere geval nog gerechtvaardigd? Of bent u van mening dat abortus in dit geval niet gerechtvaardigd is.

A10 : Een meisje van 15 jaar -ongehuwd- heeft het vermoeden dat zij zwanger is. Zij durft hierover noch met haar huisarts, noch met haar ouders te praten. Daardoor duurt het veel langer dan noodzakelijk is voordat medische hulp wordt ingeroepen. Tot welke maand .... (zie A09)

A11 : De vrouw heeft recht op abortus indien zij dat wil.

A12 : Artsen die aborteren zijn niet beter dan moordenaars.

A13 : Mensen die instemmen met abortus hebben weinig eerbied voor het leven.

A14 : Abortus is onder geen enkele voorwaarde geoorloofd.

A15 : Politici van verschillende partijen zijn de laatste jaren -en ook nu nog- bezig met het indienen van voorstellen met betrekking tot de abortus. Bent u van mening, dat er een abortuswet moet komen, die het laten plegen van abortus alleen in bijzondere gevallen toelaat, of bent u van mening, dat er een wet moet komen, die het laten plegen van abortus moeilijk maakt of zegt u, dat er geen abortuswet hoeft te zijn, maar dat de arts vrij moet zijn om te beslissen of hij al dan niet de vrouw wil helpen.

A16 : Het gesprek over abortus heeft zich in de laatste helft van dit jaar toegespitst op het wel of niet mogen toelaten van abortus na de 12^e week van de zwangerschap. Bent u van mening, dat in de wet een bepaling zal moeten worden opgenomen, dat abortus na de 12^e week absoluut verboden moet worden, of dat abortus na 12 weken uitsluitend in bijzondere gevallen mag plaats hebben of zegt u dat er in de wet geen tijdsgrens moet worden gesteld voor abortus.

EU1 : Ik vind euthanasie geoorloofd als een zieke er uiteindelijk zelf om vraagt omdat hij/zij van zichzelf weet dat hij/zij ongeneeslijk ziek is.

EU2 : ... als de naaste familie er om vraagt ingeval de zieke helemaal niet meer bij kennis is en de hoop op herstel niet meer bestaat.

EU3 : ... als bij de geboorte van een kind komt vast te staan dat het kind nog wel zuiver medisch technisch in leven kan worden gehouden, maar dat menselijke herkenning nooit meer mogelijk zal zijn.

EU4 : ... als stervenden met ongelovelijke pijnen hierdoor uit hun lijden kunnen worden verlost.

EU5 : ... als oudere mensen niet meer voor zichzelf kunnen zorgen en de wens te kennen geven liever te sterven.

SF1 : Ik vind het goed dat kinderen tot 10 jaar naakt aan het strand lopen.

SF2 : Als men geslachtsgemeenschap helemaal loslaat van het kinderen krijgen zou het spoedig alleen maar egoïsme worden.

SF3 : Ouders moeten sexuele spelletjes van jonge kinderen verbieden.

SF4 : Als jonge mensen voor hun huwelijk sexuele omgang met elkaar hebben, dan hebben zij geen respekt voor elkaar.

SF5 : Ouders moeten hun oudere kinderen voorhouden dat het beter is zich te beheersen en niet aan zelfbevrediging te doen.

dataverzameling : 1974

referentie : Veenhoven, R. en F. Hentenaar , 1975 , Nederlanders over abortus , Stimezo-onderzoek no. 3

tabel marginale frekwenties en diskriminatiematen HOMALS één-dimensionaal

NAAM	NR	M	1	2	3	4	5	6	7	8	9	10	11	DIM.
CP1	1	1	188	129	61	113	83							.026
CP2	2	1	167	112	77	108	110							.019
CP3	3	3	86	131	89	108	158							.014
AO1	4	1	528	46										.082
AO2	5	7	256	312										.505
AO3	6	8	460	107										.372
AO4	7	6	217	352										.467
AO5	8	6	485	84										.307
AO6	9	4	275	296										.515
AO7	10	8	244	323										.539
AO8	11	10	259	306										.461
AO9	12	4	217	48	18	8	21	259						.544
A10	13	8	205	62	31	12	33	224						.562
A11	14	0	178	115	36	93	153							.537
A12	15	0	41	32	77	111	314							.558
A13	16	0	114	60	69	117	215							.638
A14	17	1	43	54	62	110	305							.584
A15	18	26	249	50	250									.225
A16	19	33	158	266	118									.219
EU1	20	6	396	173										.316
EU2	21	4	299	272										.247
EU3	22	14	405	156										.346
EU4	23	4	435	136										.333
EU5	24	7	104	464										.083
SF1	25	1	130	85	56	98	205							.168
SF2	26	2	84	67	85	115	222							.161
SF3	27	1	124	109	100	114	127							.156
SF4	28	1	49	42	56	126	301							.288
SF5	29	4	124	97	88	88	174							.221
SEX	30	1	248	326										.002
AGE	31	1	39	100	130	95	98	112						.046
SOC	32	0	17	23	65	120	201	75	70	4				.012
REL	33	19	103	62	168	223								.259
POL	34	0	176	128	92	20	159							.350
EDU	35	20	311	134	43	67								.041
FUN	36	3	3	27	8	13	12	70	84	56	33	21	245	.076
URB	37	1	44	44	40	166	65	64	151					.062
													hom	.2734

N.B. Op grond van de informatie uit de één-dimensionale HOMALS is besloten in verdere analyses niet alle variabelen op te nemen. Hier gepresenteerde resultaten hebben betrekking op (een subset van) een databestand, waarin de vragen over de doodstraf (Capital Punishment , CP1 t/m CP5) en de achtergrondsvariabelen (muv REL en POL) niet opgenomen zijn.

A P P E N D I X C.1

DECK SETUP HOMALS-GS

CARD 1: JOB NUMBER CARD

Column	Format	Information
1-5	I5	number of jobs

CARD 2: TITLE CARD

Column	Format	Information
1-80	20A4	Any alphameric code to name the analysis

CARD 3: PROBLEM SIZE PARAMETER CARD

Column	Format	Information
1-5	I5	number of objects or individuals
6-10	I5	number of variables in the datamatrix
11-15	I5	number of variables in the analysis
16-20	I5	number of dimensions
21-25	I5	greatest possible number of categories in the dataset
26-30	I5	total number of categories

CARD 4: ANALYSIS PARAMETER CARD

Column	Format	Information
1-5	I5	maximum number of iterations to compute the final solution
6-15	E10.8	stop criterion for the final solution

CARD 5: I/O PARAMETER CARD

Column	Format	Information
1-5	I5	unit number of the input medium for the data
6-10	I5	input data listing 0 = no 1 = yes
11-15	I5	print options for quantifications 0 = no print 1 = print individual-scores only 2 = print individual-scores and category-scores 3 = print category-scores only

16-20	I5	plot options for quantifications 0 = no plot 1 = plot individual-scores and discrimination measures 2 = like 1 and partitioned plots of individual-scores and plot category-scores according to variables, specified in IPARTI
21-25	I5	computation of standardized category-scores 0 = no 1 = yes
26-30	I5	print of discrimination measures 0 = no 1 = yes
31-35	I5	number of categories per variable 0 = variables have different numbers of categories, specified in ICATGO k = all variables have k categories
36-40	I5	unit number for output of individual-scores to other media than line printer 0 = no extra output required k = output to nr. k
41-45	I5	unit number for output of category-scores to other media than line printer 0 = no extra output required k = output to nr. k
46-50	I5	unit number for output of standardized category-scores to other media than line printer 0 = no extra output required k = output to nr. k

Only if number of categories = 0 (Cols. 31-35) extra card(s)

Column	Format		Information
1-80	16I5	ICATGO	maximum numbers of categories of all variables (16 vars per card)

Only if plot options = 2 (Cols 16-20) : extra card(s) specifying plot partitioning and category-scores-plot selecting variables (max.= number of variables in datamatrix, 80 vars per card)

Column	Format		Information
1-80	80I1	IPARTI	columns specify variables in the same order as in the datamatrix 0 = no extra plot options for this variable 1 = individual-scores plot, labeled with categories of this variable 2 = like 1 and the plot of the category-scores of this variable 3 = plot of category-scores of this variable

Three variable-format cards (always required); these are the last cards before the datamatrix

Column	Format	Information
1-80	20A4	variable integer format (FORTRAN)

De datamatrix can follow now (depending on card 5 cols 1-5)

Depending on the value of number of jobs (card 1 cols 1-5) more jobs can be done; all cards, except for the jobnumber card, have to be repeated hereafther. This has to be done for every extra job.

A P P E N D I X C.2

DECK SETUP PRINCALS

CARD 1: JOB NUMBER CARD

Column	Format	Information
1-5	I5	number of jobs

CARD 2: TITLE CARD

Column	Format	Information
1-80	20A4	any alphameric code to name the analysis

CARD 3: PROBLEM SIZE PARAMETER CARD

Column	Format	Information
1-5	I5	number of objects or individuals
6-10	I5	number of variables in the datamatrix
11-15	I5	number of variables in the analysis
16-20	I5	number of dimensions
21-25	I5	greatest possible number of categories in the dataset within one variable
26-30	I5	total number of categories

CARD 4: ANALYSIS PARAMETER CARD

Column	Format	Information
1-5	I5	maximum number of iterations to compute the final solution
6-15	E10.8	stop criterion for the final solution
16-20	I5	maximum number of iterations to compute the initial configuration
21-30	E10.8	stop criterion for the initial configuration

The parameters labeled * can be used for the initial configuration as well as for the final solution. These parameters can be read with two formats, depending on the value of the parameter IPRIN(Card 5,cols 46-50): either IPRIN $\leq$ 1 and the parameters are read with the format I5 or IPRIN = 2 and they are read with the format 3X,2I1. If IPRIN = 2 the first column read,applies to the initial configuration and the second column applies to the final solution.This implies that no unit numbers greater than 9 are permitted if IPRIN = 2.

CARD 5: I/O PARAMETER CARD

Column	Format		Information
1-5	I5		unit number of the input medium for the data
6-10	I5		input data listing 0 = no 1 = yes
11-15	I5	*	print options for quantifications
	3X,2I1	(if IPRIN=2)	
		initial	final
		0	0 = no print
		1	1 = print individual-scores only
		2	2 = print individual-scores and variable-information
		3	3 = print variable-information only
16-20	I5	*	print history of iterations
	3X,2I1	(if IPRIN=2)	
		initial	final
		0	0 = no
		1	1 = yes
21-25	I5	*	plot options for quantification
	3X,2I1	(if IPRIN=2)	
		initial	final
		0	0 = no plot
		1	1 = plot individual-scores and variable-correlations
		2	2 = plot individual-scores,variable-correlations and partitioned plots of individual-scores and plot category scores according to variables, specified in IPARTI
26-30	I5	*	unit number for output of individual-
	3X,2I1	(if IPRIN=2)	scores to other media than line printer
		initial	final
		0	0 = no extra output required
		k	j = output to nr. k and/or j

31-35	I5	*	unit number for output of category-
	3X,2I1	(if IPRIN=2)	scores to other media than line printer
		initial	final
		0	0 = no extra output required
		k	j = output to nr. k and/or j
36-40	I5	*	unit number for output of rescaled data
	3X,2I1	(if IPRIN=2)	vectors to other media than line printer
		initial	final
		0	0 = no extra output required
		k	j = output to nr. k and/or j
41-45	I5		measurement level of variables
			0 = mixed levels which are specified in ITYP
			1 = multiple nominal variables only
			2 = ordinal variables only
			3 = numerical variables only
46-50	I5	IPRIN	i/o options for initial configuration(i.c.
			0 = no output of i.c.
			1 = identical options as for final solution
			2 = options specified in first column of the relevant parameters (*)
51-55	I5		number of categories per variable
			0 = variables have different numbers of categories (specified in ICATGO)
			k = all variables have k categories
56-60	I5	*	unit number for output of correla-
	3X,2I1	(if IPRIN=2)	tions to other media than line printer
		initial	final
		0	0 = no extra output
		k	k = output to nr. k

Only if number of categorie per variable = 0 (cols 51-55) extra card(s):

Column	Format		Information
1-80	16I5	ICATGO	maximum numbers of categories of all variables (16 vars per card)

Only if plot options = 2 (Cols 21-25):extra card(s) specifying plot partitioning and category-scores-plot selecting variables (max.= number of variables in datamatrix, 80 vars per card)

Column	Format		Information
1-80	80I1	IPARTI	colums specify variables in the same order as in the datamatrix 0 = no extra plot options for this variable 1 = individual-scores plot, labeled with categories of this variable 2 = like 1 and the plot of the category-scores of this variable 3 = plot of category-scores of this variable

Only if measurement level = 0 (Cols 41-45) : extra card(s) specifying measurement level per variable (20 vars per card)

Column	Format		Information
1-80	16I5	ITYP	0 = variable is multiple nominal 1 = variable is single nominal 2 = variable is single ordinal 3 = variable is single numerical

Three variable format cards(always required); these are the last cards before the datamatrix

Column	Format	Information
1-80	20A4	variable integer format (FORTRAN)

The datamatrix can follow now (depending on card 5 cols 1-5)

Depending on the value of number of jobs (card 1 cols 1-5) more jobs can be done; all cards, except for the jobnumber card, have to be repeated hereafter. This has to be done for every extra job.

A P P E N D I X C.3

DECK SETUP CANALS

CARD 1: JOB NUMBER CARD

Column	Format	Information
1-5	I5	number of jobs

CARD 2: TITLE CARD

Column	Format	Information
1-80	20A4	any alphameric information to title the printout

CARD 3: DATA SPECIFICATION CARD

Column	Format	Information
1-5	I5	number of objects or individuals
6-10	I5	number of variables in the first set
11-15	I5	number of variables in the second set
16-20	I5	highest category score (max.of card 6)

CARD 4: ANALYSIS SPECIFICATION CARD

Column	Format	Information
1-15	I5	number of dimensions
6-10	I5	maximum number of iterations (0=default=20)
11-25	F15.10 '	minimum stress difference (0=default=0.001)
26-40	F15.10 '	maximum weight increase (0=default=0.0001)

CARD 5: I/O SPECIFICATION CARD

Column	Format	Information
1-5	I5	unit number for input data matrix (5=card, 8,... 16=tape or disk)
6-10	I5	print output data matrix 0 = no 1 = yes
11-15	I5	unit number for output optimal individual-scores (0=no output, 7=card, 8,... 16=tape or disk)
16-20	I5	print history of iterations 0 = no 1 = yes

21-25	I5	unit number for output original and optimal scores and for output weights (0=no output, 7=card, 8,... 16=tape or disk)

CARD 6: CATEGORY CARD "

Column	Format	Information
1-80	16I5	maximum category for every variable (without missing scores)

CARD 7: VARIABLE TYPE CARD "

Column	Format	Information
1-80	16I5	variable type for every variable 0 = nominal 1 = ordinal 2 = numerical

CARD 8: DATA FORMAT CARD

Column	Format	Information
1-80	20A4	Fortran I-format

CARD 9: idem

CARD 10: idem

According to the I/O specification a data matrix may follow
Additional jobs, specified on card 1,may follow by repeating card 2-10

') If the decimal point is present, its position overrides the position indicated by the format field descriptor

") If necessary use more than one card

A P P E N D I X C.4

DECK SETUP ANACOR

	kolom	format	naam	beschrijving
Kaart 1	1- 5	I5	IOPTI	keuze uit optie 1, 2 en 3 1: de invoer is een datamatrix 2: de invoer is een (ev. gereduceerde) profiel-frequentie matrix 3: de invoer is een willekeurige matrix
	6-10	I5	NROW	aantal rijen van de invoermatrix
	11-15	I5	NCTOT	optie 1,2: som van het aantal categorieën optie 3: aantal kolommen van de invoermatrix
	16-20	I5	NVAR	optie 1,2: aantal variabelen optie 3: dummy
Kaart 2	1- 5	I5	ICAR	nummer van het medium waarop de invoermatrix staat
	6-10	I5	IPRI	nummer van het uitvoermedium
	11-15	I5	IFIG	gewenste plotuitvoer 0: geen plot-uitvoer 1: optie 1,2: plotje van categoriescores optie 3: plotje van kolomscores 2: optie 1,2: plotje van individuscores optie 3: plotje van rijscores 3: beide plotjes
	16-21	2I5	IE1,IE2	geselecteerde dimensies. Default-waarde IE1=1, IE2=2.
Kaart 3	1-80	20A4	FMT	format van de rijen van de invoermatrix optie 1,2: altijd een I-format optie 3: altijd een F-format
Kaart 4	1-80	20I5	NCAT	aantal categorieën per variabele (bij optie 1,2)
Kaart 4	1-10	I10	NOBS	aantal observaties (bij optie 3).

L I T E R A T U U R

In de volgende literatuurlijst worden een aantal afkortingen gebruikt, die we hier even op een rijtje zetten.

AJS: Australian Journal of Statistics.
AMS: Annals of Mathematical Statistics.
BK: Biometrika.
BJSP: British Journal of Statistical Psychology.
= British Journal of Psychology, Statistical section.
= British Journal of mathematical and statistical psychology.
BJP: British Journal of Psychology.
EPM: Educational and psychological measurement.
IEEE/IT: Proceedings IEEE, Information theory.
ISI: International statistical institute.
ISUP: Institute de statististique de l'université de Paris.
ITS: Instituut voor toegepaste statistiek.
JAMS: Journal of the Australian Mathematical Society.
CRAS: Comptes Rendues de l'Academie des Sciences (Paris).
DAN/SSSR: Doklady Akademie Nauk.
PNAS: Proceedings National Academy of Sciences (Washington).
PCPS: Proceedings Cambrdige Philosophical Society.
PRSL: Proceedings of the Royal Society (London).
PRSE: Proceedings of the Royal Society (Edinburgh).
SIAM: Society for industrial and applied mathematics.
JRSS(A): Journal of the royal statistical society, series A(general).
JRSS(B): idem, series B(Methodological)
JRSS(C): idem, series C(Applied). Also known als "Applied Statistics".
JMAA: Journal of mathematical analysis and applications.
JMP: Journal of mathematical psychology.
JMV: Journal of multivariate analysis.
PM: Psychometrika.
ZAMM: Zeitschrift fur angenwandte Mathematik und Mechanik.
JASA: Journal of the American Statistical Association.
MBR: Multivariate behavioural reseach.

M. Abramowitz & I.A. Segun: Handbook of mathematical functions. New York, Dover. 1965

T.W. Anderson: An introduction to multivariate statistical analysis. New York, Wiley. 1958

F.J. Anscombe: Topics in the investigation of linear relations fitted by the method of least squares. JRSS(B), 29, 1967, 1-52.

P. Appell & J. Kampé de Fériet: Fonctions hypergéometriques et hyperspheriques. Polynomes de Hermite. Paris, Gauthier-Villars. 1926.

R.R. Bahadur: A representation of the joint distribution of responses to n dichotomous items. In: H. Solomon (ed): Studies in item analysis and prediction. Stanford, Stanford University Press, 1961.

J.F. Barrett & D.G. Lampard: An expansion for some second-order probability distributions. IEEE/IT, 1, 1955, 10-15.

M.S. Bartlett: When is inference statistical inference ? In: V.P. Godambe & D.A. Sprott (eds): Foundations of statistical inference. Toronto, Holt, Reinhart, & Winston Canada, 1971.

G. Bechtel: Individual differences in the linear multidimensional scaling of choice. Psychometric Society Meeting, Princeton, April 1969.

E.H. Bell & J. Sirjamaki: Social foundations of human behaviour. New York, Harper. 1967.

E. Beltrami: Sulle funzioni bilineari. Giorn. Math. Battaglin, 11, 1873, 98-106.

J.P. Benzécri: Sur l'analyse des préférences. Mimeo, ISUP, 1967.

J.P. Benzécri: Lois de probabilité sur un ensemble produit. Les diverses notions de independence et le critere d'entropie maximale. Mimeo, ISUP, 196

J.P. Benzécri e.a.: Analyse des données (2 vols). Paris, Dunod. 1973.

J.G. Bethlehem, H. Elffers, R.D. Gill, J. Rijvordt: Methoden, voetangels, en klemmen in de faktor-analyse. Amsterdam, Mathematisch Centrum, Rapport SN 7/77, 1977.

Y.M.M. Bishop, S.E. Fienberg, P.W. Holland: Discrete multivariate analysis, theory and practice. Cambridge, MIT Press, 1975.

A. Björck, G.H. Golub: Numerical methods for computing angles between linear subspaces. Math. Comput., 27, 1973, 579-594.

H.M. Blalock Jr.: Causal inference in nonexperimental research. Chapell Hill, University of North Carolina Press, 1964.

R.D. Bock: Methods and applications of optimal scaling. University of North Carolina, L.L. Thurstone Lab. Report 25, 1960.

R. Boudon: L'analyse mathématique des faits sociaux. Paris, Plan. 1967.

J.L. Brown Jr.: A criterion for the diagonal expansion of a second-order probability distribution in orthogonal polynomials. IEEE/IT, 4, 1958, 172.

J.R. Bunch, C.P. Nielsen, D.C. Sorenson: Rank-one omdification of the symmetric eigen-problem. Num. Math. 31, 1978, 31-48.

E. van der Burg & J. de Leeuw: How to use CANALS. Afd Datatheorie, RUL, 1978.

C. Burt: A comparison of factor analysis and analysis of variance. BJSP, 1, 1948, 3-27.

C. Burt: The influence of differential weighting. BJSP, 3, 1950, 105-128.

C. Burt: The factorial analysis of qualitative data. BJSP, 3, 1950, 166-185.

C. Burt: Test construction and the scaling of items. BJSP, 4, 19-1, 95-129.

C. Burt: Scale analysis and factor analysis. BJSP, 6, 1953, 5-23.

S. Cambanis & B. Liu: On the expansion of a bivariate distribution and its relationship to the output of a nonlinearity. IEEE/IT, 17, 1971, 17-25.

F. Caillez & J.P. Pagès: Introduction à l'analyse des données. Paris, SMASH. 1976.

J.D. Carroll: Individual differences and multidimensional scaling. In: R.N. Shepard, A.K. Romney, S.B. Nerlove (eds): Multidimensional scaling: theory and application in the behavioural sciences. New York, Seminar Press, 1972.

J.D. Carroll & P. Arabie: Multidimensional scaling. Ann. Rev. Psychol., 1980, in aantocht.

J.D. Carroll & J.J. Chang: Nonmetric multidimensional analysis of paired comparison data. Psychometric Society Meeting, 1964.

A.L. Cauchy: Sur l'équation à l'aide de laquelle on détermine les inégalités séculaires des mouvements des planètes. Exec. de Math., 4, 1829, 174-195.

J.C. Chang & R.E. Bargmann: Internal multi-dimensional scaling of categorical variables. Dept. of Statist., University Of Georgia, Report 108, 1974.

P.L. Chesson: The canonical decomposition of bivariate distributions. JMV, 6, 1976, 526-537.

A. Christofferson: Factor analysis of dichotomized variables. PM 40, 1975, 5-32.

A. Christofferson: Two-step weighted least squares factor analysis of dichotomized variables. PM, 42, 1977, 433-438.

N. Cliff: Orthogonal rotation to congruence. PM, 31, 1966, 33-42.

J.W.M. Collaris & J.A. Kropman: Van Jaar tot Jaar, tweede fase. Den Haag, Staatsuitgeverij, 1978.

W.W. Cooley & P.R. Lohnes: Multivariate procedures for the behavioural sciences. New York, Wiley, 1962.

W.W. Cooley & P.R. Lohnes: Multivariate data analysis. New York, Wiley, 1971.

C.H. Coombs: A theory of data. New York, Wiley, 1964.

C.H. Coombs & J.E. Smith: On the detection of structure in attitudes and developmental processes. Psychol. Rev., 80, 1973, 337-351.

R.D. Cooper, M.R. Hoare, M. Rahman: Stochastic processes and special functions: on the probabilistic origin of some positive kernels associated with classical orthogonal polynomials. JMAA, 61, 1977, 262-291.

R.M. Cormack: A review of classification. JRSS(A), 134, 1971, 321-367.

D.R. Cox: The analysis of multivariate binary data. JRSS(C), 21, 1972, 113-120

D.R. Cox: Note on grouping. JASA, 52, 1957, 543-547.

H. Daalder & J.G. Rusk: Perceptions of party in the Dutch Parliament. In: S.C. Patterson & J.C. Wahlke (eds): Comparative legislative behaviour:frontiers of research. New York, Wiley, 1972.

H. Daalder & J.P. van de Geer: Partijafstanden in de Tweede Kamer. Acta Politica, 12, 1977, 289-345

P. Dagnelie: Analyse statistique à plusieurs variables. Gembloux, Presses Agronomiques, 1975.

J.N. Darroch: Multiplicative and additive interaction in contingency tables. BK, 61, 1974, 207-214.

J. Dauxois & A. Pousse: Les analyses factorielles en calcul des probabilités et en statistique: essai d'étude synthétique. Dissertatie, Université Paul Sabatier, Toulouse, 1976.

A.W. Davis: Asymptotic theory for principal component analysis: nonnormal case. AJS, 19, 1977, 206-212.

R.L. Degerman: The geometric representation of some simple structures. In: R.N. Shepard, A.K. Romney, S.B. Nerlove (eds): Multidimensional scaling: Theory and application in the behavioural sciences. New York, Seminar Press, 1972.

L. Delbeke: Enkele analyses op voorkeursoordelen voor gezinssamenstellingen. Mimeo, Centrum voor Mathematische Psychologie, Leuven , 1978.

A.P. Dempster: Elements of continuous multivariate analysis. Reading, Addison-Wesley, 1969.

A.P. Dempster: An overview of multivariate data analysis. JMV, 1, 1971, 316-346.

D.R. Divgi: Calculation of the tetrachoric correlation coefficient. PM, 44, 1979, 169-172.

J. Dronkers: Manipuleerbare variabelen in de schoolloopbaan. 9th World congress of sociology. Uppsala, 1978.

J. Dronkers & J.J.M. Jungbluth: Schoolloopbaan en geslacht. In: J. Peschar (ed): Van achteren naar voren. Den Haag, Staatsuitgeverij, 1979.

G.K. Eagleson: Polynomial expansions of bivariate distributions. AMS, 35, 1964, 1208-1215.

G.K. Eagleson: Canonical expansions of birth and death processes. Theory Prob. Appl. 14, 1969, 209-218.

G.K. Eagleson: A characterization theorem for positive definite sequences on the Krawtchouk polynomials. AJS, 11, 1969, 29-38.

G.K. Eagleson & H.O. Lancaster: The regression system of sums with random elements in common. AJS, 9, 1967, 119-125.

C. Eckart & G. Young: The approxiamtion of one matrix by another of lower rank. PM, 1, 1936, 211-218.

F.Y. Edgeworth: The statistics of examinations. JRSS(A), 51, 1888, 599-635.

H.A. Edgerton & L.E. Kolbe: The method of minimum variation for the combination of criteria. PM, 1, 1936, 183-187.

A.L. Edwards: Techniques of attitude scale construction. New York, Appleton-Century-Crofts, 1957.

B. Efron: Bootstrap methods: another look at the jackknife. Ann. Statist., 7, 1979, 1-26.

P. Elias: Bounds on performance of optimum quantizers. IEEE/IT, 16, 1970, 172-184.

A. Erdelyi: Higher transcendental functions. New York, McGraw Hill, 1953.

B. Escofier & B. Le Roux: Etude de trois problemes de stabilité en analyse factorielle. Publ. ISUP, 21, 1972, 2-48.

B. Escofier & B. Le Roux: Mesure de l'influence d'un descripteur sur une analyse en composantes principales. Publ. ISUP, 22, 1977, 25-44.

Ky Fan: Maximum properties and inequalities for the eigenvalues of completely continuous operators. PNAS, 37, 1951, 760-766.

M. Fischbein & I. Ajzen: Belief, attitude, and behaviour: an introduction to theory and research. Reading, Addison-Wesley, 1975.

R.A. Fisher: The precision of discriminant functions. Ann. Eug., 10, 1940, 422-429

C. Flament: Tresse de Guttman. In: Ordres totaux finis. Paris, Gauthier-Villars en Mouton, 1971.

U.G. Foa: New developments in facet design and analysis.Psych. Rev. 72, 1965, 262-274.

F. Galton: Co-relations and their measurement, chiefly from anthropometric PRSL , 45, 1888, 135-145.

F.R. Gantmacher & M.G. Krein: Oszillationsmatrizen, oszillationskerne, und kleine Schwingungen mechanischer Systeme. Berlin, Akademie Verlag, 1960.

H. Gebelein: Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusamenhang mit der Ausgleichsrechnung. ZAMM, 21, 1941, 364-379.

J.P. van de Geer: Inleiding in de multivariate analyse. Arhem, Van Loghem Slaterus, 1967.

J.P. van de Geer: Introduction to multivariate analysis for the social sciences. San Francisco, Freeman, 1971

A. Gersho: Asymptotically optimal block quantization. IEEE/IT, 25, 1979, 373-380.

N.C. Giri: Multivariate statistical inference. New Yrok, Academic Press, 1977.

H. Gish & J.N. Pierce: Asymptotically optimal quantizing. IEEE/IT, 14, 1968, 676-683.

D.V. Glass (ed): Social mobility in Britain. Glencoe, Free Press, 1954.

R. Gnanadesikan: Methods for the statistical data analysis of multivariate observations. New York, Wiley, 1977.

D.V. Gokhale & S. Kullback: The information in contingency tables. New York, Dekker, 1978.

I.J. Good: Maximum entropy for hypotheses formulation, especially for multidimensional contingency tables. AMS, 34, 1963, 911-934.

L.A. Goodman: On the statistical analysis of mobility tables. Am J. Sociol., 70, 1965, 564-585.

L.A. Goodman: On the measurement of social mobility: an index of status persistence Amer. Sociol. Rev., 34, 1969, 832-850.

L.A. Goodman & W.H. Kruskal: Measures of association for cross classification.
I : JASA 49, 1954, 732-764.
II : JASA 54, 1959, 123-163.
III: JASA 58, 1963, 310-364.
IV : JASA 67, 1972, 415-421.

P.E. Green & J.D. Carroll: Mathematical tools for applied multivariate analysis. New York, Wiley, 1976.

R.C. Griffiths: Positive definite sequences and canonical correlation coefficients. AJS, 12, 1970, 162-165.

R.C. Groffiths: The canonical correlation coefficients of bivariate gamma distributions. AMS, 40, 1969, 1401-1408.

H. Gulliksen: Theory of mental tests. New York, Wiley, 1950.

L. Guttman: The quantification of a class of attributes: a theory and method of scale construction. In: P. Horst (ed): The prediction of personal adjustment. New York, SSRC, 1941.

L. Guttman: A basis for scaling qualitative data. Amer. Sociol. Rev., 9, 1944, 139-150.

L. Guttman: An approach for quantifying paired comparisons and rank order. AMS, 17, 1946, 144-163.

L. Guttman: The principal components of scale analysis. In S.A. Stouffer e.a.: Measurement and prediction, Princeton, Princeton University Press, 1950

L. Guttman: The basis for scalogram analysis. In: S.A. Stouffer e.a.: Measurement and prediction, Princeton, Princeton University Press, 1950.

L. Guttman: A note on Sir Cyril Burts "Factorial analysis of qualitative data". BJSP, 6, 1953, 1-4.

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

L. Guttman: A new approach to factor analysis: the radex. In P.F. Lazarsfeld (ed): Mathematical thinking in the social sciences. Glencoe, Free Press, 1954.

L. Guttman: A generalized simplex for factor analysis. PM, 20, 1955, 173-192.

L. Guttman: The structure of interrelations among intelligence tests. Procee-

dings of the 1964 Invitational Conference on testing problems. Princeton, ETS, 1964.

L. Guttman: The non-metric breakthrough for the behavioural sciences. Automatic Data Processing Conference of the Information Processing Association of Israel. Jerusalem, 1966.

L. Guttman: Order analysis of correlation matrices. In: R.B. Cattell (ed): Handbook of multivariate experimental psychology. Chicago, Rand-McNally, 1966.

L. Guttman: A general nonmetric technique for finding the smallest coordinate space for a configuration of points. PM, 33, 1968, 469-506.

S.J. Haberman: The analysis of frequency data. Chicago, University of Chicago Press, 1974.

J.M. Hammersley: Some general reflections on statistical practice. In: W.J. Ziegler (ed): Contributions to applied statistics. Basel, Birkhauser, 1976.

E.J. Hannan: The general theory of canonical correlation and its relation to functional analysis. JAMS, 2, 1961, 229-242.

R.J. Harris: A primer of multivariate statistics. New York, Academic Press, 1975

W.J. Heiser & J. de Leeuw: Metric multidimensional unfolding. MDN, 4, 1979, 26-50.

W.J. Heiser & J. de Leeuw: How to use SMACOF-3. Department of Datatheory, University of Leiden 1979.

J. Hemelrijk: Underlining random variables. Statistica Neerlandica, 20, 1966, 1-8.

M.O. Hill: Correspondence analysis: a neglected multivariate method. JRSS(C), 23, 1974, 340-354.

H.O. Hirschfeld: A connection between correlation and contingency. PCPS, 31, 1935, 520-524.

T. Hirshi & H.C. Selvin: Principles of survey analysis. Glencoe, Free Press, 1973.

F.R. Hodson, D.G. Kendall, P. Tautu: Mathematics in the archeological and historical sciences. Edinburgh, Edinburgh University Press, 1971.

L. Hogben: Statistical theory. New York, Norton, 1957.

P. Horst: Measuring complex attitudes. J. Soc. Psychol., 6, 1935, 369-374.

P. Horst: Obtaining a composite measure from a number of different measures of the same attribute. PM, 1, 1936, 53-60.

H. Hotelling: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol, 24, 1933, 417-441, 498-520.

H. Hotelling: Relations between two sets of variables. BK, 28, 1936, 321-377.

ITS: Onderzoeksvoorstel "Van Jaar tot Jaar". Nijmegen, ITS, 1968.

D.R. Jensen: A note on positive dependence and the structure of bivariate distributions. SIAM J. Appl. Math., 20, 1971, 749-753.

P.O. Johnson: The quantification of qualitative data in discriminant analysis. JASA, 45, 1950, 65-76.

C. Jordan: Mémoire sur les formes bilinéaires. J. Math. Pures. Appl., 19, 1874, 35-54.

S. Karlin: Oscillation properties of eigenvectors of strictly totally positive matrices. J. Anal. Math. Jerusalem, 9. 1964, 247-266.

S. Karlin: Total positivity. Stanford, Stanford University Press, 1968.

T. Kato: Perturbation theory for linear operators. Berlin, Springer, 1966.

J.A. van Kemenade & J.A. Kropman: Verborgen talenten ? Kritische kanttekeningen bij een onjuiste interpretatie. Sociologische Gids, 19, 1972, 219-228.

O. Kempthorne: Probability, statistics, and the knowledge business. In: V.P. Godambe & D.A. Sprott (eds): Foundations of statistical inference. Toronto, Holt, Reinhart & Winston Canada, 1971.

O. Kempthorne: Theories of inference and data analysis. In: T.A. Bancroft (ed): Statistical papers in honour of George Snedecor. Ames, Iowa State University Press, 1972.

D.G. Kendall: A statistical approach to Flinders Petrie's sequence dating. Bull. ISI, 40, 1963, 657-680.

D.G. Kendall: Incidence matrices, interval graphs, and seriation in archeology. Pacific J. Math., 28, 1969, 565-570.

D.G. Kendall: Some problems and methods in statistical archeology. World Archeology, 1, 1969, 61-76.

M.G. Kendall: A course in multivariate analysis. London, Griffin, 1957/1975.

M.G. Kendall: The history and future of statistics. In: T.A. Bancroft (ed): Statistical papers in honour of George Snedecor. Ames, Iowa State University Press, 1972.

J.R. Kettenring: Canonical analysis of several sets of variables. BK, 58, 1971, 433-46

J. Kiefer: Review of M.G. Kendall & A Stuart: The advanced theory of statistics, volume 2. AMS, 35, 1964, 1371-1380.

J.A. Kropman & J.W.M. Collaris: Van Jaar to Jaar, eerste fase. Nijmegen, ITS, 1974.

J.B. Kruskal: Multidimensional scaling bij optimizing goodness of fit to a nonmetric hypothesis. PM, 29, 1964, 1-27.

J.B. Kruskal: Multidimensional scaling and other methods for discovering structure. In: K. Enslein, A.J. Ralston, H.S. Wilf (eds): Statistical methods for digital computers , vol III. New York, Wiley, 1977.

J.B. Kruskal & J.D. Carroll: Geometric models and badness-of-fit fucntions. In: P.R. Krishnaiah (ed): Multivariate Analysis, vol II. New York, Academic Press, 1969

J.B. Kruskal & R.N. Shepard: A nonmetric variety of linear factor analysis. PM, 39, 1974, 123-157.

A.M. Kshirsagar: Multivariate analysis. New York, Dekker, 1978.

S. Kullback: Information theory and statistics. New York, Wiley, 1959.

D. Lafaye de Michaux: Approximation d'analyse canoniques non-linéaires de variables aléatoires. Dissertatie, Université de Nice, 1978.

C.J. Lammers: Is de universiteit een politieke leerschool? Universiteit en Hogeschool, 15, 1969, 1-43.

H.O. Lancaster: Some properties of the bivariate normal distribution considered in the form of a contingency table. BK, 44, 1957, 289-292.

H.O. Lancaster: The structure of bivariate distributions. AMS, 29, 1958, 719-736.

H.O. Lancaster: Zero correlation and independence. AJS, 1, 1959, 53-56.

H.O. Lancaster: On tests of independence in several dimensions. JAMS, 1, 1960, 241-254.

H.O. Lancaster: On statistical independence and zero correlation in several dimensions. JAMS, 1, 1960, 492-496.

H.O. Lancaster: Correlations and canonical forms of bivariate distributions. AMS, 34, 1963, 532-538.

H.O. Lancaster: The chi-squared distribution. New York, Wiley, 1969.

H.O. Lancaster: The multiplicative definition of interaction. AJS, 13, 1971, 36-44.

H.O. Lancaster: The multiplicative definition of interaction: an addendum. AJS, 17, 1975, 34-35.

H.O. Lancaster: Joint probability distributions in the Meixner classes. JRSS(B), 37, 1975, 434-443.

H.O. Lancaster & M.A. Hamdan: Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characteristics. PM, 29, 1964, 383-391.

D.N. Lawley: The factorial analysis of multiple item tests. PRSE, 62, 1944, 74-82.

L. Lebart: The significance of eigenvalues issued from correspondence analysis of contingency tables. In: Proceedings COMPSTAT 1976, Wien, Physika Verlag, 1976.

P.A. Lee: A diagonal expansion for the 2-variate Dirichlet probability density function. SIAM J. Appl. Math., 21, 1971, 155-165.

J. de Leeuw: Some contributions to the analysis of categorical data. Report RN 004-69, Dept. Datatheory, Univ. Leiden, 1969.

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

J. de Leeuw, F.W. Young, Y. Takane: Additive structure in qualitative data: an alternating least square method with optimal scaling features. PM, 41, 1976, 471-503.

J. de Leeuw: Correctness of Kruskal's algorithms for monotone regression with ties. PM, 42, 1977, 141-144.

J. de Leeuw: Normalized cone regression. Mimeo, Afd. Datatheorie, Univ. Leiden, 1977.

J. de Leeuw & I. Stoop: Sekundaire analyse "Van Jaar tot Jaar" met behulp van niet-lineaire multivariate technieken. In: J. Peschar (ed): Van achteren naar voren. Den Haag, Staatsuitgeverij, 1979.

J. de Leeuw & W.J. Heiser: Theory of multidimensional scaling. In: P.R. Krishnaiah & L. Kanal (eds): Handbook of Statistics. Amsterdam, North Holland, 1980.

J. de Leeuw & J.J.L. van Rijckevorsel: HOMALS EN PRINCALS. Tweede internationale symposium Data Analyse en Informatika, Versailles, Oktober 1979.

M.V. Levine: Transformations that render curves parallel. JMP, 7, 1970, 410-443.

M.V. Levine: Transforming curves into curves with the same shape. JMP, 9, 1972, 1-16.

M.V. Levine: Additive measurement with short segments of curves. JMP, 12, 1975, 212-224.

R. Likert: A technique for the measurement of attitudes. Archives of Psychol. 140, 1932, 5.53.

J.C. Lingoes: The multivariate analysis of qualitative data. MBR, 3, 1968, 61-94.

J. Loevinger: A systematic approach to the construction and evaluation of tests of ability. Psychol. Monograph, 61, 1947, no 4.

J. Loevinger: The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. Psych. Bull., 45, 1948, 507-530.

F.M. Lord: On the statistical treatment of football numbers. Amer. Psychologist 8, 1953, 750-751.

F.M. Lord: Some relations between Guttman's principal components of scale analysis and other psychometric theory. PM, 23, 1958, 291-296.

A. Lubin: Linear and nonlinear discriminating functions. BJSP, 3, 1950, 90-104.

W.R. MacDonell: On criminal anthropometry and the identification of criminals. BK, 1, 1901/1902, 177-227.

J.A. McFadden: A diagonal expansion in Gegenbauer polynomials for a class of second-order probability densities. SIAM J. Appl. Math., 14, 1966, 1433-1436.

D.K. McGraw & J.T. Wganer: Elliptically symmetric distributions. IEEE/IT, 14, 1968, 110-120.

D. MacKenzie: Statistical theory and social interests: a case study. Social studies of science, 8, 1978, 35-83.

M. Marcus: Finite-dimensional multilinear algebra. New York, Dekker, 1973.

M. Masson: Analyse non-linéaires de données. CRAS, 278, 1974, 803-806.

K. Maung: Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children. Ann. Eug., 11, 1941, 189-223.

K. Maung: Discriminant analysis of Tocher's eye colour data. Ann. Eug., 11, 1941, 64-76.

J. Max: Quantizing for minimum distortion. IEEE/IT, 6, 1960, 7-12.

R.G. Miller: The jackknife: a review. BK, 61, 1974, 1-15.

L. Mirsky: Symmetric gauge functions and unitarily invariant norms. Quart. J. Math. Oxford (2), 11, 1960, 50-59.

R.J. Mokken: A theory and a procedure of scale analysis. Mouton, Den Haag, 1970.

W. Molenaar: Ik word ziek van de statistiek. Heymans Bull HB-165-EX, Rijksuniversiteit Groningen, 1974.

D.F. Morrison: Multivariate statistical methods. New York, McGraw Hill, 1967/1976.

B. Muthèn: Contributions to factor analysis of dichotomous variables. PM, 43, 1978, 551-560.

J.C. Naouri: Analyse factorielle des correspondences continues. Publ. ISUP, 19, 1970, 1-100.

J. von Neumann: Some matrix inequalities and a metrization of matric space. Tomsk Univ. Rev., 1, 1937, 286-299.

J. Neveu: Bases mathématiques du calcul des probabilités. Parijs, Masson, 1964.

B. Niemöller: Schaalanalyse volgens Mokken. Amsterdam, Technisch centrum FSW, 1976.

S. Nishisato: Analysis of categorical data:dual scaling and its applications. Manuskript van boek, 1978.

B.J. Norton: Biology and philosophy: the methodological foundations of biometry. J. Hist. Biology, 8, 1975, 85-93.

B.J. Norton: Karl Pearson and statistics: the social origin of scientific innovation. Social studies of science, 8, 1978, 3-34.

U. Olsson: Maximum likelihood estimation of the polychoric correlation coefficient. PM, 44, 1979, 443-460.

M.E. O'Neill: Asymptotic distribution of the canonical correlation coefficients from contingency tables. AJS, 20, 1928, 75-82.

M.E. O'Neill: Distributional expansion for canonical correlation from contingency tables. JRSS(B), 40, 1978, 303-312.

K. Pearson: On lines and planes of closest fit to points in space. Phil. Mag., 2, 1901, 559-572.

K. Pearson: On the theory of contingency and its relation to association and normal correlation. Drapers Co Research Mem, Biometric series, 1, 1904.

K. Pearson: On the measurement of the influence of "Broad categories" on correlation. BK, 9, 1913, 116-139.

K. Pearson: On the general theory of multiple contingency with special reference to partial contingency. BK, 11, 1916, 145-158.

K. Pearson & D. Heron: On theories of association. BK, 9, 1913, 159-315.

M.L. Puri & P.K. Sen: Nonparametric methods in multivariate analysis. New York, Wiley, 19[illegible].

C.R. Rao: Multivariate analysis: an indispensable statistical aid in applied research. Sankhya, 22, 1960, 317-338.

G. Rasch: An individual-centered approach to item analysis with two categorie[illegible]

of answers. Proc. Nuffic Symp. Psych. Measurement Theory. 1966.

G. Rasch: An informal report on a theory of objectivity in comparisons. Proc. Nuffic Symp. Psych. Measurement Theory, 1966.

L. Reisman: Levels of aspiration and social class. Amer. Sociol. Rev., 18, 1953 233-

A. Rényi: On measures of dependence. Acta Math. Acad. Sc. Hungar., 10, 1959, 441-451.

F. Restle: A metric and an ordening on sets. PM, 24, 1959, 207-220.

G.M. Roe: Quantizing for minimum distortion. IEEE/IT, 10, 1964, 384-385.

E. Roskam: Metric analysis of ordinal data in psychology. Voorschoten, VAM, 1968.

S.N. Roy: Some aspects of multivariate analysis. New York, Wiley, 1957.

W.W. Rozeboom: Sensitivity of a linear composite of predictor items to differential item weighting. PM, 44, 1979, 289-296.

H. Rubin: Occam's razor needs new blades. In: V.P. Godambe & D.A. Sprott (eds) Foundations of statistical inference. Toronto, Holt, Reinhart, & Winston Canada, 1971.

B.M. Russett: Inequality and instability: the relation of land tenure to politics In: Rowney & Graham (eds): Quantitative history, 1969.

J.J.L. van Rijckevorsel & J. de Leeuw: An outline of HOMALS. Afdeling Datatheorie FSW, Rijksuniversiteit Leiden, Rapport RB002-78, 1978.

J.J.L. van Rijckevorsel & J. de Leeuw: An outline of PRINCALS. Afdeling Datatheorie FSW, Rijksuniversiteit Leiden, Rapport RB002-79, 1979.

O.V. Sarmanov: Maximum correlation coefficient (symmetric case) DAN/SSSR, 120, 1958, 715-718.

O.V. Sarmanov: Maximum correlation coefficient (asymmetric case) DAN/SSSR, 121, 1958, 52-55.

O.V. Sarmanov: Investigation of stationary Markov processes by the method of eigenfunction expansion. Selected translations in Math. Statist. en Prob, 4, 1963, 245-269.

O.V. Sarmanov & Z.N. Bratoeva: Probabilistic properties of bilinear expansions of Hermite polynomials. Theory Prob. Appl., 12, 1967, 470-481.

O.V. Sarmanov & V.K. Zacharov: Maximum coefficients of multiple correlation. DAN/SSSR, 130, 1960, 269-271

E. Schmidt: Zur theorie der lineairen und nichtlinearen Integralgleichungen. Math. Ann., 63, 1906, 433-476.

D.K. Sharma: Design of absolutely optimal quantizers for a wide class of distortion measures. IEEE/IT, 24, 1978, 693-702.

R.N. Shepard: The analysis of proximities: multidimensional scaling with an unknown distance function. PM, 27, 1962, 125-140, 219-245.

R.N. Shepard: The circumplex and related topological manifolds in the study of perception. In S. Shye (ed): Theory construction and data analysis in the behavioural sciences. San Francisco, Jossey-Bass, 1978.

W.F. Shepard: On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. Proc. London Math. Soc., 29, 1898, 253-280.

S.Shye: Partial order scalogram analysis. In: S. Shye (ed): Theory construction and data analysis in the behavioural sciences. San Francisco, Jossy-Bass, 1978.

R. Sibson: Order invariant methods for data analysis. JRSS(B), 34, 1972, 311-349.

P. Slater: The analysis of personal preferences. BJSP, 13, 1960, 119-135.

C. Spearman: Correlation of sums and differences. BJP, 5, 1913, 417-423.

R. Stammeyer & L. Staallekker: Van goeden bloede. Vakgroep Sociale Psychologie, Rijksuniversiteit Leiden, 1977.

S.S. Stevens: On the psychophysical law. Psych. Rev. , 64, 1957, 153-181.

D.K. Stewart & W.A. Love: A general canonical correlation index. Pych. Bull., 70, 1968, 160-163.

G.W. Stewart: Introduction to matrix computation. New Yrok, Academic Press, 1973.

M. Stone: Cross-validatory choice and assessment of statistical predictions. JRSS(B), 30, 1974, 111-147.

G.P. Styan: Hadamard products and multivariate statistical analysis. Linear Algebra Appl., 6, 1973, 217-240.

M. Sugiyama: Religious behaviour of the Japanese. Execution of a partial order scalogram analysis based on quantification theory. U.S.-Japan seminar on theory and applications of MDS and related techniques, La Jolla, 1975.

J.J. Sylvester: Sur la réduction biorthogonal d'une forme linéo-linéaire à sa forme canonique. CRAS, 108, 1889, 651-653.

M.M. Tatsuoka: Multivariate analysis: techniques for educational and psychological research. New York, Wiley, 1971.

M. Tenenhaus: Analyse en composantes principales d'un ensemble de variables nominales et numériques. Revue Statist. Appliqué, 25, 1977, 39-56.

R.M. Thorndike: Correlational procedures for research. New York, Gardner, 1978.

R.M. Thorndike: Canonical analysis and predictor selection. MBR, 1977, 12, 75-87.

R.M. Thorndike & D.J. Weiss: A study of the stability of canonical correlations and canonical components. EPM, 33, 1973, 123-134.

L.L. Thurstone: Multiple factor analysis. Chicago, University of Chicago Press, 1947.

W.S. Torgerson: Theory and methods of scaling. New York, Wiley, 1958.

F.G. Tricomi: Vorlesungen über Orthogonalreihen. Berlin, Springer, 1955.

A.C. Tucker: A structure theorem for the consecutive 1's property. J. Combinatorial theory, B, 12, 1972, 153-162.

J.W. Tukey: The future of data analysis. AMS, 33, 1962, 1-67.

S. Tyan & J.B. Thomas: Characterization of a class of bivariate distribution functions. JMV, 5, 1975, 227-235.

R. Veenhoven & F. Hentenaar: Nederlanders over abortus. Stimezo onderzoek 3, 1975.

J.H. Venter: Probability measures on product spaces. South African Statist. J. 1, 1966, 3-20.

H. Wainer: Estimating coefficients in linear models: It don't make no nevermind. Pychol. Bull., 83, 1976, 213-217.

P. Whittle: Probability. London, Penguin, 1970.

S. Wiegersma: Belangstellingsonderzoek bij de differentiatie na de lagere school. Dissertatie Leiden, 1959.

J.H. Wilkinson: The algebraic eigenvalue problem. Oxford, Clarendon, 1965.

S.S. Wilks: Weighting systems for linear functions of correlated variables when there is no independent variable. PM, 3, 1938, 23-40.

E.J. Williams: Use of scores for the analysis of association in contingency tables BK, 39, 1952, 274-289.

A.L. van de Wollenberg: Smallest space analysis of Guttman's radex. Dept. Psychology, Universiteit van Nijmegen, Rapport 74MA01, 1974.

E. Wong & J.B. Thomas: On polynomial expansions of second order distributions. J. SIAM, 10, 1962, 507-516.

R.C. Wood: On optimum quantization. IEEE/IT, 15, 1969, 248-252.

F. Yates: The analysis of contingency tables with grouping based on quantitative characters. BK, 35, 1948, 176-181.

F.W. Young: A model for polynomial conjoint analysis algorithms. In: R.N. Shepard, A.K. Romney, S.B. Nerlove (eds): Multidimensional scaling: theory and applications in the social sciences. New York, Seminar Press, 1972.

F.W. Young, J. de Leeuw, Y. Takane: Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. PM, 41, 1976, 505-529.

F.W. Young, Y. Takane, J. de Leeuw: The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling feautes. PM, 43, 1978, 279-281.

E. Zvulin: Multidimensional scalogram analysis: the method and its application In: S. Shye (ed): Theory construction and data analysis in the behavioural sciences. San Francisco, Jossey-Bass, 1978.