MULTIVARIATE ANALYSE VAN LINEAIRE STRUCTURELE MODELLEN

J. de Leeuw en A. Mooijaart
University of California, L.A. en Rijksuniversiteit Leiden

MULTIVARIATE ANALYSE

Een variabele is een functie f, gedefinieerd op een domein X, en met waarden in een bereik Y. Zo is bijvoorbeeld de functie f die aan alle Nederlandse kinderen tussen 10 en 15 jaar hun lengte in centimeters toekent een variabele. Merk op dat bereik en domein een onderdeel zijn van de definitie van een variabele. Zo is de functie g, die aan alle kerktorens in de Bommelerwaard hun lengte in centimeters toekent, een andere variabele dan f. Het bereik is hetzelfde, maar het domein verschilt. Ook de functie h, die aan alle Nederlandse kinderen tussen 10 en 15 jaar hun lengte in millimeters toekent is verschillend van f. Immers het domein is hetzelfde, maar het bereik verschilt.

Een multivariabele bestaat uit een aantal functies die allemaal hetzelfde domein hebben. Zo kunnen we aan de Nederlandse kinderen tussen 10 en 15 jaar hun lengte in centimeters, hun score op de Raven-test, het beroep van hun vader, hun geslacht, en de kleur van hun ogen toekennen. Dit definieert variabelen f_1 t/m f_5, die tezamen een multivariabele vormen. Het is belangrijk op te merken dat variabelen niet noodzakelijk numerieke waarden hebben. Beroep van de vader is in principe een verbaal label. In dit verband spreken we wel van een nominale variabele. Sociologen zullen de neiging hebben beroepen in te delen in klassen met ongeveer gelijke sociale status. Wanneer we de variabele beroep als bereik een dergelijke statusladder geven, dan is het bereik geordend, en noemen we beroep een ordinale variabele. We kunnen de diverse categorieën zelfs kwantificeren, dat wil zeggen hun labels vervangen door nummers, en dan maken we beroep numeriek. Eenzelfde soort verhaal kan men ophangen voor oogkleur. Variabelen zijn dus niet noodzakelijk numeriek, en in de sociale wetenschappen zijn numerieke variabelen mischien zelfs eerder uitzondering dan regel.

In veel sociaal-wetenschappelijke studies bekijken we multivariabelen. Dit geldt natuurlijk voor survey-onderzoek, maar ook voor de variantie-analytische proefopzetten uit de functieleer, waarin we het verband tussen experimentele condities en uitkomsten bestuderen. Zowel condities als uitkomsten zijn echter variabelen, gedefinieerd op het domein gevormd door de proefpersonen. En het zal duidelijk zijn dat met name de condities zeer vaak helemaal niet numeriek zijn.

Van de Geer heeft zich vanaf het begin van zijn wetenschappelijke carrière met de multivariate analyse, dat wil zeggen met de analyse van multivariabelen, bezig gehouden. Zonder overdrijving kunnen we zeggen dat vanaf ongeveer 1965 de studie van de multivariate analyse zijn belangrijkste wetenschappelijke activiteit was, en het onderwijs in de multivariate analyse zijn belangrijkste onderwijsactiviteit. Hij heeft over de multivariate analyse drie boeken geschreven. Het eerste van die boeken (Van de Geer, 1967) is tamelijk traditioneel. Het valt op door veelvuldig gebruik van geometrische terminologie, en door het vrijwel volledig ontbreken van statistische overwegingen en

resultaten. Juist daardoor sluit het naadloos aan bij de psychometrische traditie in de multivariate analyse, die we ook vinden bij bijvoorbeeld Thurstone en Thomson. Het derde boek (Van de Geer, 1986) is buitengewoon origineel. De geometrische benadering van multivariate analyse, en van matrixalgebra in het algemeen, wordt consequent voortgezet. Als gevolg hiervan lijkt het boek eigenlijk meer op de klassieke werken uit de analytische meetkunde dan op de multivariate analyse boeken van de statistische school. Het tweede boek (Van de Geer, 1971) is in zekere zin een buitenbeentje. Het leidende idee is hier niet de ellips of het snijdende vlak, maar het paddiagram. Het boek werd geschreven tijdens een verblijf aan het Institute of Advanced Studies in Palo Alto, ten tijde van de opkomst van de padanalyse in de econometrie en sociometrie. De verworvenheden van de econometrie (stelsels van simultane vergelijkingen) en van de sociometrie (causale analyse) zijn in het boek verwerkt, en geïntegreerd met klassiek psychometrisch materiaal (factoranalyse). Dat levert een succesvolle combinatie op, die zijn tijd een jaar of tien vooruit was.

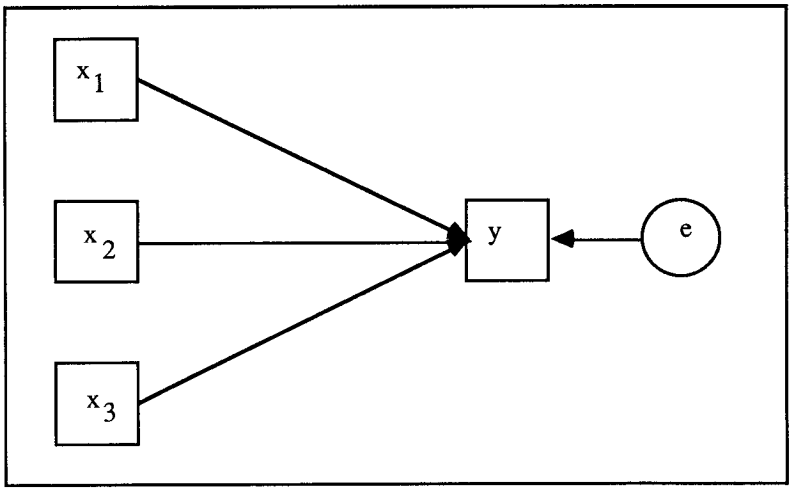
De geometrische benadering van de multivariate analyse is door veel leerlingen van Van de Geer overgenomen. Hierbij zijn ook het boek van Coombs (1964), en de ontwikkeling van meerdimensionale schaalmethoden, van groot belang geweest. In hun hoofdstuk in dit boek gaan Heiser en Meulman nader in op deze specifiek geometrische aanpak. In dit hoofdstuk zullen we ons bezighouden met het fundamentele idee dat de modellen en technieken ontwikkeld in de econometrie en sociometrie niet wezenlijk verschillen van de gebruikelijke psychometrische technieken. Alleen moeten we er rekening mee houden dat, met name in sociaal-wetenschappelijk onderzoek, de gebruikelijke vooronderstellingen van lineaire statistische modellen (numerieke multivariabelen op intervalschalen, die bovendien normaal verdeeld zijn, of ten minste lineaire regressies hebben) niet opgaan. We zullen daarom de causale analyse, analyse van simultane stelsels, en de factor-analytische modellen uit de psychometrie als een geheel behandelen, en we zullen aangeven op welke plaatsen de klassieke methoden, die bijvoorbeeld in Van de Geer (1971) besproken worden, aangepast moeten worden om in de sociale wetenschappen van groter nut te zijn. Het spreekt daarbij, gegeven de opzet van dit boek, min of meer vanzelf dat we vooral aandacht zullen besteden aan werk dat in Leiden uitgevoerd is, en dat te beschouwen is als voortzetting en verdere uitbreiding van het werk van Van de Geer.

PADMODELLEN

Het fundamentele idee achter padmodellen illustreren we met een aantal kleine diagrammen. In Figuur 1 zien we een model dat weergeeft hoe een variabele y beïnvloed wordt door drie variabelen x_1, x_2, en x_3. Daarnaast is er ook nog een residu dat een zekere rol speelt, maar dit residu wordt niet geobserveerd. Niet-geobserveerde (ook wel: latente) variabelen geven we aan met een cirkel, geobserveerde variabelen met een vierkant.

Figuur 1 is een regressiemodel: de onafhankelijke variabelen (ook wel: predictoren) x_1, x_2, en x_3 bepalen de afhankelijke variabele y. In een vergelijking geformuleerd kunnen we het model schrijven als

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e.$$



Figuur 1. Regressiemodel met 3 onafhankelijke x-variabelen en 1 afhankelijke y-variabele.

We veronderstellen dus dat de afhankelijke variabele y een lineaire combinatie is van de drie onafhankelijke variabelen x_1, x_2, en x_3, plus een residu. In de gebruikelijke statistische versies van dit model nemen we aan dat predictoren en residu normaal verdeeld zijn en onafhankelijk van elkaar.

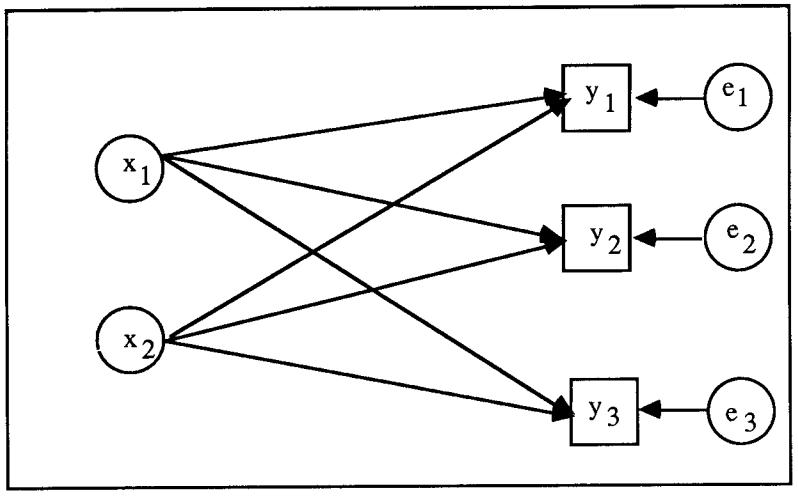
Figuur 2 toont een wat gecompliceerder model. Er zijn twee latente predictoren x_1 en x_2, die drie geobserveerde variabelen y_1, y_2, en y_3 voorspellen. De vergelijkingen zijn

$$y_1 = b_{01} + b_{11}x_1 + b_{21}x_2 + e_1,$$

$$y_2 = b_{02} + b_{12}x_1 + b_{22}x_2 + e_2,$$

$$y_3 = b_{03} + b_{13}x_1 + b_{23}x_2 + e_3.$$

Dit model is een factor analyse model: x_1 en x_2 zijn factoren, die de onderlinge samenhang van y_1, y_2, en y_3 moeten verklaren. In de gebruikelijke statistische versies van het factor analyse model nemen we aan dat de factoren en residuen normaal verdeeld en ongecorreleerd, dus onafhankelijk van elkaar, zijn.



Figuur 2. Factor-analyse.

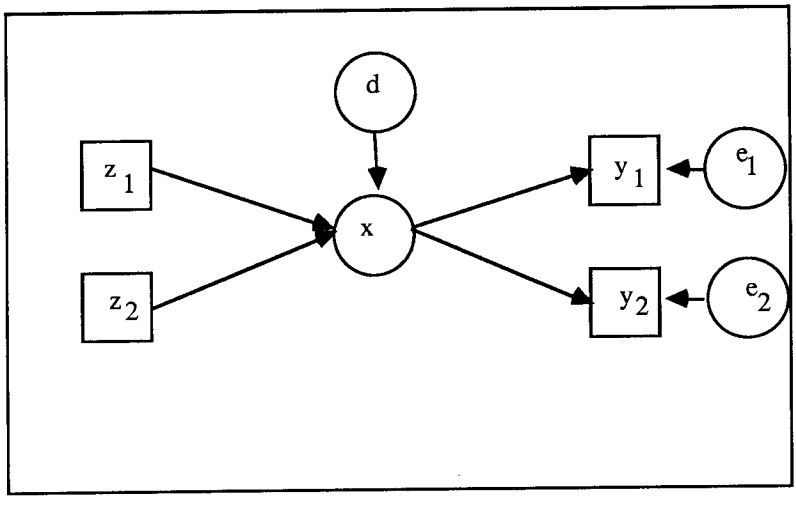
In Figuur 3 staat tenslotte een nog wat ingewikkelder model, een zogenaamd MIMIC model (Jöreskog en Goldberger, 1975). De vergelijkingen zijn

$$x = a_0 + a_1 z_1 + a_2 z_2 + d,$$

$$y_1 = c_1 + b_1 x + e_1,$$

$$y_2 = c_2 + b_2 x + e_2.$$

We zien dat het eerste stuk een regressiemodel is, en het tweede stuk een factor analyse model. Het algemene idee zal nu hopelijk duidelijk zijn. We kunnen allerlei soorten pijldiagrammen opstellen, en die pijldiagrammen vertalen in lineaire structurele modellen. Op systematische wijze gebeurt dit in het programma LISREL (Jöreskog en Sörbom, 1984). LISREL-modellen vormen een veel algemenere klasse dan de simpele voorbeelden die wij besproken hebben. Een LISREL-model combineert overigens ook regressiemodellen, die geheel in termen van latente variabelen gedefinieerd zijn, met factor analyse modellen, die de relatie tussen de latente variabelen en hun geobserveerde indicatoren vastleggen.
De statistische theorie waarop LISREL gebaseerd is gaat uit van normaal verdeelde latente variabelen. Omdat geobserveerde variabelen lineaire combinaties van latente variabelen zijn, betekent dit dat ook de geobserveerde variabelen normaal verdeeld moeten zijn.



Figuur 3. MIMIC model.

Bovendien moeten we de beschikking hebben over een willekeurige steekproef uit een precies gedefinieerde populatie waarover we uitspraken willen doen. Maar wat gebeurt er nu wanneer we dit soort vooronderstellingen niet goed kunnen maken, of zelfs wanneer het duidelijk is dat deze vooronderstellingen helemaal niet opgaan? We moeten daarbij twee gevallen onderscheiden.
In de eerste plaats kan het zo zijn, dat we met numerieke variabelen te maken hebben, maar dat het duidelijk is dat deze numerieke variabelen niet normaal verdeeld zijn. En in de tweede plaats is het mogelijk dat de variabelen, of ten minste sommige variabelen, in het onderzoek in het geheel niet numeriek zijn maar bijvoorbeeld nominaal of ordinaal. In beide gevallen zullen verschillende soorten van aanpassingen van de klassieke technieken nodig zijn.

CAUSALE ANALYSE

Voordat we echter op de technische kant ingaan, moeten we eerst wat opmerkingen maken over de algemeen methodologische kant van het gebruik van dit soort methoden. We baseren ons daarbij vooral op De Leeuw (1978, 1984a, 1985), maar er zijn ondertussen veel verwijzingen te geven naar literatuur waar hetzelfde soort uitspraken gedaan worden. De LISREL-methodologie is in veel deelgebieden van de sociale wetenschappen, met name in de sociologie, bijzonder populair. Bij sommige tijdschriften krijgt men de indruk dat het

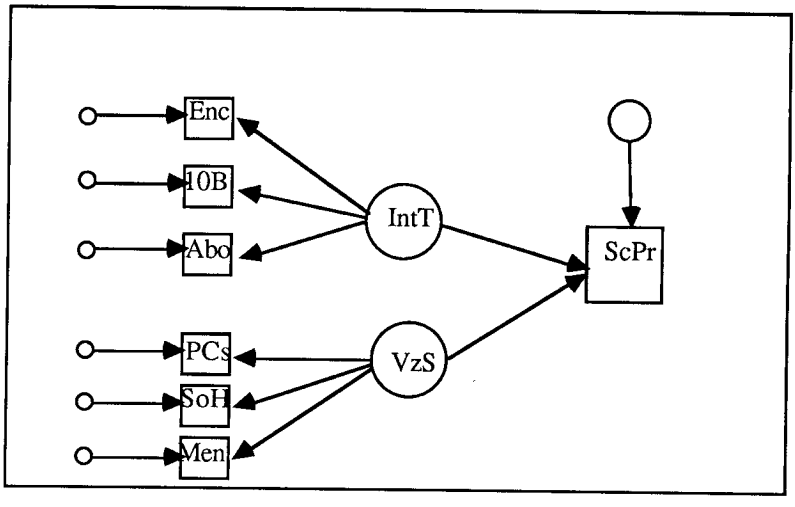
gebruik van LISREL een noodzakelijke en bijna een voldoende voorwaarde is om een artikel gepubliceerd te krijgen. Het is daarbij duidelijk dat LISREL en soortgelijke technieken aansluiten bij de behoefte om redelijk tot zeer ingewikkelde theorieën over mechanismen, die het gedrag bepalen van individuen en groepen individuen, te operationaliseren en te toetsen. In disciplines waarin men geneigd is tot uitgebreide theorievorming zelfs wanneer er geen empirische evidentie beschikbaar is, en waarin causale uitspraken in termen van 'effecten' van variabelen die andere variabelen 'bepalen' de regel zijn, is LISREL een uitkomst. Het lijkt erop alsof tot voor kort nogal mistige praktijken door het gebruik van LISREL en aanverwante technieken nu op verantwoorde wijze uitgevoerd kunnen worden. Immers de theorieën moeten nu in ieder geval geoperationaliseerd worden op zo'n manier dat ze in het algemene LISREL-model passen, en de theorieën kunnen bovendien statistisch getoetst worden. We kunnen nu precies nagaan of een theorie past bij de gegevens die we hebben, en ook hoeveel variantie van de belangrijkste variabelen in het onderzoek door ons model 'verklaard' wordt.

Helaas is de situatie aanzienlijk minder rooskleurig dan een groot aantal sociologen, en sociaal-wetenschappelijke methodologen, schijnen te denken. In de eerste plaats, zoals we boven al aantoonden, is er bij sociaal-wetenschappelijk onderzoek zelden sprake van normaal verdeelde variabelen en willekeurige steekproeven uit een welomschreven populatie. Dit is een technisch bezwaar, en verderop in dit artikel zullen we laten zien dat het met wat inspanning ondervangen kan worden. Niettemin impliceert onze constatering, dat de statistische eigenschappen van LISREL in veel gevallen met een grote korrel zout genomen moeten worden. En dit doet natuurlijk iets af van de exactheid en verantwoordheid waarmee een groot deel van de toepassingen gepresenteerd wordt.

Een essentiëler bezwaar is, dat het hypothetisch-deductieve image van LISREL grotendeels schijn is. Het gaat niet zo dat men eerst een theorie opstelt, die dan precies operationaliseert, en vervolgens toetst. De theorie, als er al sprake is van een theorie, is in het algemeen vaag, en waarom bepaalde pijlen in het diagram er niet zijn en andere wel is vaak verre van duidelijk. Natuurlijk vindt het computerprogramma keurig netjes waarden voor alle regressiecoëfficiënten, en geeft het de belangrijkheid van alle causale paden. En vanaf dat moment kan de fantasie van de socioloog of de onderwijskundige het roer overnemen, en ontaardt het gebruik van LISREL in het oude, vertrouwde 'hineininterpretieren', nu binnen het algemene kader van het causale model. Zo vinden we bijvoorbeeld uitspraken, dat de opleiding van de moeder belangrijker is voor de cognitieve ontwikkeling van het kind dan de opleiding van de vader. Dit soort uitspraken blijkt vaak gebaseerd op padcoëfficiënten die beide klein zijn, en die nauwelijks verschillen.

Het voorbeeld in de vorige paragraaf brengt ons op het meest ernstige bezwaar dat men tegen de LISREL-methodologie kan inbrengen. Het gebruik van causale terminologie in dit soort regressiemodellen is uitermate misleidend, en men kan niet volstaan met wat voorbehouden aan het begin van het artikel, om vervolgens los te barsten in een orgie van 'effecten', 'invloeden', 'bepaaldheden', 'belangrijkheden', enzovoort. Wat regressiemodellen laten zien is dat er bepaalde verbanden tussen variabelen bestaan, en dat die verbanden al of niet verdwijnen wanneer er gecontroleerd wordt voor variatie in andere variabelen. Maar die verbanden zijn altijd heel specifiek (correlaties of covarianties), en de soort controle is dat eveneens (partiële correlatie). Alle uitspraken die gedaan worden, gaan ervan uit dat er voor alle relevante variabelen gecontroleerd is (de 'ceteris paribus' clausule), en dat bovendien de relaties die geobserveerd en gecontroleerd worden inderdaad lineair zijn. Dat wil zeggen: niet alleen de relaties tussen de indicatoren die we in het

model gebruiken moeten lineair zijn, maar ook de relaties tussen de latente constructen waarover we uitspraken doen.



Figuur 4. Schoolloopbaanmodel.

Laten we het model in Figuur 4 eens bekijken. De geobserveerde variabele 'schoolprestaties' wordt 'verklaard' met twee constructen 'intellectueel thuisklimaat' en 'voorzieningen in de school'. Intellectueel thuisklimaat wordt gemeten met de drie indicatoren 'is er een encyclopedie in het gezin', 'lezen de ouders meer dan 10 boeken per jaar', en 'aantal abonnementen'. Schoolvoorzieningen wordt gemeten met 'aantal personal computers op school', ' is er een cursus Spaans en/of Hebreeuws', en 'aantal mentoren per leerling'. Stel we vinden dat het pad van thuisklimaat naar schoolprestaties een waarde van .24 krijgt, en het pad van schoolvoorzieningen naar schoolprestaties een waarde van .16. Sociologen zeggen dan, dat het intellectuele klimaat in het gezin belangrijker is voor de schoolprestaties van het kind dan de voorzieningen op school. Maar bij het doen van deze uitspraak hebben we twee stappen gemaakt, die volstrekt uit de lucht gegrepen zijn. In de eerste plaats hebben we, alleen maar door de latente variabelen een in principe willekeurige naam te geven, de stap van het empirische niveau van de indicatoren naar het theoretische niveau van de begrippen gemaakt. En in de tweede plaats hebben we de gevonden regressiecoëfficiënten vertaald in het causale begrip 'belangrijkheid', waar ze, alweer in principe, niets mee te maken hebben. De waarde van een regressiecoëfficiënt wordt bepaald door de overige variabelen in het model, door de schattingsmethode, door het al dan niet lineair zijn van de regressies, door de betrouwbaarheid en validiteit van de indicatoren, door de meetfouten in de schoolloopbaanvariabele, enzovoort. De conclusie die

we op basis van onze LISREL-analyse vinden, en vervolgens tegen veel geld aan het Ministerie van Onderwijs verkopen, is gebaseerd op goedkope verbale goocheltrucs. Het voorbeeld is met opzet zo gekozen, dat de tegenwoordige regering de conclusie graag in deze vorm wil vernemen. Bij eventuele kabinetswisseling kunnen we zonder veel moeite onze LISREL-analyses in de gewenste richting aanpassen.

AL DAN NIET NORMAAL VERDEELDE NUMERIEKE VARIABELEN

Het basisidee van LISREL is de analyse van covarianties. En covariantie is een getal dat de mate van samenhang tussen twee variabelen weergeeft. In feite is het niets anders dan een gemiddeld kruisprodukt van de scores op twee variabelen. Zo'n kruisprodukt zal positief zijn als er een positieve samenhang tussen de variabelen is, d.w.z. een toename van de scores op de ene variabele gaat samen met een toename van de scores op de andere variabelen. Indien zo'n kruisprodukt negatief is, geldt het tegenovergestelde. We noemen covarianties wel tweede orde-kruisprodukten, omdat de samenhang van twee variabelen wordt weergegeven. Dit betekent dat we met behulp van covarianties niet de gelijktijdige samenhang tussen drie variabelen kunnen weergeven, daarvoor hebben we derde orde-kruisprodukten nodig. We zullen hier later nog op terugkomen als we analyses voor niet-normale verdelingen nader zullen bekijken. We kunnen overigens via een truc wel de gemiddelden van de variabelen in de analyses van covarianties betrekken. Het gemiddelde van een variabele is immers niets anders dan een gemiddeld kruisprodukt van een variabele met een variabele die enkel scores 1 heeft. Het analyseren van covarianties en gemiddelden is interessant in situaties waarbij er verschillende groepen van bijv. personen zijn. Met behulp van deze tweede orde- kruisprodukten kunnen we dan zowel het verschil in gemiddelden van de groepen analyseren, als het verschil in covarianties. Dit soort analyses komt men vooral voor bij zg. cohort-studies. Zoals we in de figuren 1 t/m 4 hebben gezien zijn er onbekende parameters. Deze parameters noemen we de model-parameters. In de praktijk moeten we deze parameters schatten. We zullen daarom bekijken hoe deze parameters geschat kunnen worden.

In het algemeen geldt dat een model nooit perfect opgaat. Een model is slechts iets waarvan we denken dat het de samenhang van de variabelen goed weergeeft. Het is verder heel goed mogelijk dat als we een model hebben gevonden dat de samenhang goed weergeeft, er nog vele andere modellen zijn die die samenhang evengoed weergeven. Via pure data-analyse (d.w.z. analyse van gegevens met minimale a priori kennis) zijn deze structurele modellen dan ook niet te gebruiken. Als data-analyticus gaan we er dan ook van uit dat het model (d.w.z. het pijlendiagram) geleverd is door iemand die een expert is in het inhoudelijke gebied dat we onderzoeken. De data-analyticus kan dan alleen zeggen of een dergelijk model bevestigd wordt (liever of het verworpen moet worden) op grond van de analyse van de geobserveerde variabelen.

Zoals we gezegd hebben zijn in LISREL covarianties van belang. We kunnen echter twee soorten covarianties onderscheiden: de covarianties zoals ze gevonden worden in de steekproef (als het gemiddelde van een kruisprodukt) en covarianties zoals ze volgens het model zouden moeten zijn als we de onbekende parameters zouden kennen. De steekproef covarianties verzamelen we in een rijtje (ook wel vector genoemd) en dat rijtje noemen we s. Hetzelfde doen we door de covarianties op grond van de modelparameters, deze verzamelen we in een rijtje genaamd σ. Deze σ kennen we echter niet omdat de

elementen van dit rijtje functies zijn van de onbekende modelparameters. Wat we dus willen is het schatten van deze parameters zodanig dat s en σ zoveel mogelijk op elkaar lijken. Er zijn vele manieren om dit op elkaar lijken te definiëren. We beginnen hier echter met het zg. kleinste kwadraten criterium. Dit criterium zegt dat we die parameters zoeken die als resultaat hebben dat het gekwadrateerde verschil van de elementen van s en σ zo klein mogelijk is. In formule schrijven we dat als: zoek die parameters zodanig dat:

$$\Phi = \Sigma_{i=1}^{m}\Sigma_{j=1}^{m}(s_{ij} - \sigma_{ij})^2$$

minimaal is. In deze formule zijn s_{ij} en σ_{ij} de twee genoemde covarianties tussen de variabelen i en j. Deze covarianties staan in de rijtjes s en σ. Het schatten van de onbekende parameters is nu niets anders dan het zoeken van die parameters zodanig dat Φ minimaal is. Dat is een wiskundig probleem waar we hier niet verder op in zullen gaan.

Op zich lijkt bovengenoemde procedure eenvoudig. Er kunnen zich echter problemen voordoen. Stel dat we een oplossing van de onbekende parameters hebben gevonden, dan is het goed mogelijk dat er nog andere oplossingen van de parameters zijn die hetzelfde minimum van Φ geven. De twee oplossingen kunnen we dan op grond van het gekozen criterium niet van elkaar onderscheiden. In een dergelijk geval spreken we van een niet-geïdentificeerd model.

Een heel bekend voorbeeld van een dergelijk niet-geïdentificeerd model is het factormodel met ongecorreleerde factoren. Om toch tot een unieke oplossing van de parameters (factorladingen en communaliteiten) in factoranalyse te komen, wordt een oplossing geroteerd naar een oplossing die een aantal aantrekkelijke eigenschappen heeft. In de praktijk betekent dit dat we een oplossing zoeken die "optimaal interpreteerbaar" is. Dit kan weer op vele manieren gedefinieerd worden. De meest bekende rotatietechniek voor een model met ongecorreleerde factoren is VARIMAX. Deze techniek roteert een oplossing zodanig dat de factorladingen ofwel absoluut groot zijn, ofwel ongeveer 0 zijn. Een andere manier om tot een unieke oplossing te komen in factoranalyse (maar ook bij andere modellen) is het fixeren van bepaalde parameters. Veelal betekent dit dat geëist wordt dat sommige factorladingen (meer algemeen parameters) 0 zijn. Zoals duidelijk zal zijn bij beide methoden gebruiken we extra informatie die niet in de data aanwezig is. In een volgende paragraaf zullen we dan ook een methode bespreken waarbij het wel mogelijk is om tot een unieke oplossing te komen op grond van informatie die in de data aanwezig is.

Het criterium van de kleinste kwadratenmethode is slechts één van de criteria die we kunnen kiezen. Een ander veel gebruikt criterium is de methode van de meest aannemelijke schatters. Deze methode gaat er van uit dat de variabelen normaal verdeeld zijn, d.w.z. dat aangenomen wordt dat de scores op de variabelen een steekproef uit een normaalverdeling is. Het is dan mogelijk schatters van de parameters te vinden die een aantal statistisch aantrekkelijke eigenschappen heeft. Ook is het dan mogelijk te toetsen of een bepaalde dataset past bij een bepaald model; dit is de zg. toets van het model. Een dergelijke toets van het model is van belang omdat, zoals we eerder hebben gezegd, het toetsen van een bepaald model vaak het meest belangrijke is bij het analyseren van een structureel model. In de praktijk echter blijkt vaak dat de variabelen helemaal niet normaal verdeeld zijn, zodat we geen eenvoudige methode hebben om bijvoorbeeld de toets

van een model uit te voeren. Het toepassen van de methode van de meest aannemelijke schatters is dan ook in de praktijk niet zonder problemen.
Een andere methode die wel aantrekkelijke statistische eigenschappen heeft in het geval de variabelen niet normaal verdeeld zijn is de zg. asymptotisch distrubutievrije (ADF) methode. Zie De Leeuw (1983), Mooijaart en Bentler (1985a). Deze methode gaat uit van de zg. centrale limietstelling in de statistiek. Dat betekent in ons voorbeeld dat aangenomen wordt dat als de steekproef maar erg groot is (voor sommige gevallen is aangetoond dat een grootte van minimaal 400 genoeg is) de covarianties s_{ij} normaal verdeeld zijn, met gemiddelde s_{ij} en de covariantie tussen s_{ij} en s_{kl} gelijk aan $\gamma_{ij,kl}$. In een dergelijk geval kiezen we als te minimaliseren functie:

$$\Phi = \Sigma_{i=1}^{m} \Sigma_{j=1}^{m} \Sigma_{k=1}^{m} \Sigma_{l=1}^{m} \gamma_{ij,kl} (s_{ij} - \sigma_{ij})(s_{kl} - \sigma_{kl}),$$

waarbij s_{ij} en σ_{ij} eerder zijn gedefinieerd en $\gamma_{ij,kl}$ een bepaalde schatter van de elementen van de inverse matrix van Γ (de matrix met elementen $\gamma_{ij,kl}$) is. Een voordeel van deze methode is dat we zonder een statistische verdeling aan te nemen voor de geobserveerde variabelen aantrekkelijke statistische eigenschappen hebben, zoals een toets voor het model. Een nadeel echter is dat deze methode uitgaat van grote steekproeven. Gelukkig hebben we in de sociale wetenschappen vaak grote steekproeven, zodat dit nadeel het voordeel van deze methode in de praktijk meestal niet overtreft.

Bovenstaande ADF-methode is erg algemeen, in de zin dat bijv. de methode van de meest aannemelijke schatters een speciaal geval is als we voor $\gamma_{ij,kl}$ bepaalde waarden kiezen. Ook de methode van de kleinste kwadraten is een speciaal geval van bovenstaande formule. Een andere interessante eigenschap kan worden afgeleid met deze methode: robuustheid van de meest aannemelijke schatters en de bijbehorende toets van het model. Robuustheid betekent dat als assumpties die gemaakt worden, geschonden worden de betreffende methode toch een juist resultaat geeft. Dit kan praktisch van belang zijn. Immers de methode van de meest aannemelijke schatters is veel eenvoudiger uit te rekenen dan de ADF-methode. Bijvoorbeeld als we 20 variabelen hebben dan is een ADF methode zoals boven beschreven nagenoeg niet te hanteren, omdat het veel te veel rekentijd vergt. Ook kan een nadeel van de ADF-methode zijn dat als de steekproef klein is we nogal instabiele schatters vinden. (Dit zou wellicht opgelost kunnen worden door meer robuuste schatters van de steekproefmomenten te bepalen. Onderzoek hierover is gaande.) Over de robuustheid van de methode van de meest aannemelijke schatters kan het volgende opgemerkt worden: indien de fouten onafhankelijk zijn van de latente variabelen en de fouten zelf normaal verdeeld zijn, dan geeft de methode van de meest aannemelijke schatters toch de juiste schatters van de parameters en de juiste toets van het model. Dit betekent dat als bijvoorbeeld in LISREL de latente variabelen niet normaal verdeeld zijn maar bovengenoemde assumpties wel opgaan, LISREL toch de juiste schatters en toets van het model geeft. Zie Mooijaart en Bentler (1987).

Een recent computerprogramma ontwikkeld door Muthen, genaamd LISCOMP, kan gezien worden als een generalisatie van het LISREL-programma. In het LISCOMP-programma is het mogelijk ook niet normaal verdeelde variabelen te analyseren met behulp van de ADF-methode. Ook is het mogelijk om categorische variabelen in de analyse te betrekken. Hierbij worden deze categorische variabelen gezien als partities van niet geobserveerde continue variabelen, de zogenaamde latente response variabelen.

De ADF methode is een methode om structurele modellen voor niet normaal verdeelde

variabelen te analyseren. Deze methode gaat er echter van uit dat we covarianties van variabelen onderzoeken. We zouden echter ook in plaats van tweede orde-coëfficiënten derde orde- coëfficiënten kunnen onderzoeken. Dit is iets wat zelden gedaan wordt, opvallend omdat er toch interessante informatie in hogere orde- coëfficiënten aanwezig kan zijn. Een argument waarom dit niet gedaan wordt ligt in de populariteit van de normaalverdeling. Indien variabelen normaal verdeeld zijn, dan is deze verdeling volledig bepaald door de gemiddelden van de variabelen en alle covarianties. In een dergelijke situatie is er geen additionele informatie meer in de hogere orde-coëfficiënten, en dus heeft het analyseren van deze coëfficiënten geen nut. Zijn de variabelen echter niet normaal verdeeld dan kunnen hogere orde-coëfficiënten wel degelijk informatie geven die van belang is. Bijvoorbeeld de scheefheid van een variabele wordt bepaald door het derde moment van de variabele. Zouden we dus scheve variabelen alleen onderzoeken met eerste en tweede orde- coëfficiënten, dan wordt er informatie weggegooid. Een manier om toch deze hogere orde-coëfficiënten te analyseren kan gedaan worden door deze coëfficiënten te schrijven in termen van model-parameters. Bijvoorbeeld in factoranalyse met onafhankelijke factoren kunnen we de tweede en derde orde-coëfficiënten schrijven als:

$$\sigma_{ij} = \Sigma_{r=1}^{p} f_{ir}f_{jr}$$

$$\sigma_{ijk} = \Sigma_{r=1}^{p} \mu_r f_{ir}f_{jr}f_{kr},$$

waarbij μ_r het derde moment van de r-de faktor is. Dus naast de factorladingen hebben we ook een parameter voor de scheefheid van de factoren (Mooijaart, 1978, Mooijaart, 1985b). Het is niet moeilijk aan te tonen dat we dergelijke modellen met hogere orde-coëfficiënten ook kunnen analyseren met de ADF- methode. Het is dus hier ook mogelijk om bijvoorbeeld een toets van het model uit te voeren.

Tot nu toe hebben we het uitsluitend gehad over lineaire relaties tussen de variabelen. In het bijzonder over lineaire relaties tussen de indicatoren en de latente variabelen. Het is echter ook mogelijk om niet-lineaire modellen te formuleren en te analyseren. We kunnen bijvoorbeeld aannemen dat er een polynome relatie is tussen de variabelen (zie Mooijaart en Bentler, 1985b). Ook dit is een model voor niet-normaal verdeelde geobserveerde variabelen, hoewel de latente variabelen wel normaal verdeeld kunnen zijn. Dit soort modellen kunnen erg zinvol zijn in de praktijk. Bijvoorbeeld in een onderzoek naar verschillende meningen over politieke vraagstukken kunnen variabelen als "hoeveel geld moet er besteed worden aan ontwikkelingshulp" of "hoeveel geld moet er besteed worden aan defensie" lineair gerelateerd zijn aan een politieke links-midden-rechts dimensie. Terwijl variabelen als "moet abortus gelegaliseerd worden" of "mogen verhuurders weigeren een kamer te verhuren aan een homosexueel" kwadratisch met deze politieke dimensie gerelateerd zijn. Dit soort niet-lineaire modellen kunnen eveneens met de boven besproken ADF-methode geanalyseerd worden.

HET GEBRUIK VAN OPTIMAAL SCHALEN

In de vorige paragraaf hebben we gezien wat we kunnen doen wanneer de gebruikelijke veronderstelling dat we met normaal verdeelde numerieke grootheden te doen hebben niet

opgaat. We hebben bovendien een bepaald soort niet-lineaire relaties tussen numerieke variabelen toegelaten. Maar in veel sociaal-wetenschappelijk onderzoek zijn, zoals we eerder al vastgesteld hebben, de variabelen helemaal niet numeriek. Denk maar aan het beroep van de vader, bijvoorbeeld. We kunnen natuurlijk met elkaar afspreken dat alle hoogleraren een score 5 krijgen, alle universitaire hoofddocenten een score 4, alle directeuren van grote bedrijven een score 3, en alle overige inwoners van Nederland een score 1. Maar een dergelijke afspraak is in hoge mate arbitrair. In de sociologie zijn methoden ontwikkeld om tot aanzienlijk minder willekeurig toekennen van getallen te komen. Hierbij groepeert men beroepen in hoofdarbeid en handarbeid, rekening houdend met salaris en hoeveelheid ondergeschikten, en men komt dan vaak tot een indeling in zo'n 7 tot 25 'sociale status' klassen. Daaraan worden dan vaak getalwaarden toegekend, die uit de natuurlijke maatschappelijke ordening van de categorieën van de variabele volgen. Het probleem is natuurlijk dat zelfs bij een dergelijke procedure toch nog een groot aantal willekeurige beslissingen genomen moeten worden, en dat men daardoor alleen met een slecht geweten dit soort variabelen in bijvoorbeeld LISREL kan gebruiken. En voor variabelen als religie geldt dit in nog veel grotere mate.

Daar komt nog bij dat de willekeurigheid van de toekenning van scores niet het enige probleem is. Lineaire structurele modellen zijn lineair, omdat ze er van uitgaan dat de regressies tussen de variabelen lineair zijn. Wanneer dit niet het geval is, dan is de covariantie of correlatie geen goede maat om de samenhang tussen twee variabelen weer te geven. Bij nominale en ordinale variabelen is het lineair zijn van de regressies een probleem. Laten we eens het verband tussen sekse en inkomen bekijken in de beroepsbevolking tussen 20 en 30 jaar. De regressiefunctie vinden we door voor iedere vaste waarde van de eerste variabele het gemiddelde van de tweede variabele te berekenen. De regressie van inkomen op sekse is per definitie lineair. Immers door de twee punten (man, gemiddeld inkomen van mannen) en (vrouw, gemiddeld inkomen van vrouwen) gaat altijd een rechte lijn. Maar door de punten (inkomen < 10000, proportie vrouwen met inkomen < 10000), (10000 < inkomen < 20000, proportie vrouwen met 10000 < inkomen < 20000), ... , die de regressie van sekse op inkomen bepalen gaat waarschijnlijk helemaal geen rechte lijn. Die regressie is misschien wel monotoon, maar het lijkt onwaarschijnlijk dat we zelfs maar bij benadering een rechte lijn zullen vinden. En dan nog ... Wanneer we een rechte lijn vinden voor de regressie van sekse op inkomen, dan vinden we die niet voor sekse op de logaritme van inkomen. Of op de derde macht van inkomen. Met andere woorden: regressies die lineair zijn voor bepaalde transformaties van variabelen zijn dat niet voor andere transformaties. En in de sociale wetenschappen is de precieze vorm waarin we de variabelen uitdrukken, de precieze schaal dus, grotendeels arbitrair.

Dit leidt tot het trachten te vinden van optimale schalingen, dat wil zeggen schalingen van de variabelen waarvoor alle regressies zo lineair mogelijk zijn. We transformeren dus inkomen, en we kwantificeren beroepsniveau en religie, op zo'n manier dat de relaties tussen alle paren variabelen in de analyse zo lineair mogelijk zijn. Dit is één manier om het probleem van de optimale schaling, of van de niet-lineaire multivariate analyse, te benaderen. Uitvoerige discussies van deze benadering, en van een aantal andere benaderingen die dikwijls tot hetzelfde of tot een vergelijkbaar resultaat leiden staan in De Leeuw (1973), in Gifi (1981), en in veel ander werk van de 'Leidse school'. De gesuggereerde techniek is hier dus een tweestaps analyse: eerst wordt nagegaan in hoeverre de regressies gelineariseerd kunnen worden, door middel van een optimaal

schalingsprogramma zoals HOMALS, PRIMALS, of PRINCALS, en vervolgens worden de optimaal geschaalde variabelen in LISREL of in een ADF-programma gestopt. Het is van belang om hierbij in het oog te houden, dat de verdelingsassumpties waarop LISREL gebaseerd zijn na optimale schaling niet meer opgaan. De schalingen van de variabelen zijn nu zelf steekproefafhankelijk, en daardoor wordt de analyse van de steekproefverdelingen gecompliceerder. Een recent resultaat (De Leeuw, 1984b) is wat dit betreft bemoedigend. Wanneer de regressies in de populatie lineair zin, dan levert de ADF-methode op de correlaties tussen de optimaal getransformeerde variabelen resultaten op die statistisch dezelfde eigenschappen hebben als de LISREL schatters. Wanneer de regressies in de populatie niet lineair zijn, maar door schaling gelineariseerd kunnen worden, dan geeft ADF na optimale schatting statistisch gezien betere resultaten.

Een padmodel moet, zoals alle modellen, twee dingen kunnen doen. In de eerste plaats moet het de gegeven toestand goed beschrijven, en in de tweede plaats moet het model gebruikt kunnen worden om goed te voorspellen. Voor dit laatste is het niet alleen nodig dat de modeltoets een goede passing aangeeft, maar moeten bovendien de belangrijkste afhankelijke variabelen zoals bijvoorbeeld inkomen of schoolsucces goed voorspeld worden (met weinig residuele variantie). Deze overweging leidt tot een andere vorm van optimaal schalen, die nog wat verder afstaat van gebruikelijke statistische procedures zoals LISREL. Hierboven hebben we gezien dat we de variabelen kunnen schalen op zo'n manier dat de regressies zo lineair mogelijk worden, en dat dit bepaalde voordelen biedt. We kunnen echter ook een ander criterium hanteren, en schalen op zo'n manier dat de 'verklaarde' variantie van een aantal geselecteerde variabelen, die we de belangrijkste vinden, zo groot mogelijk wordt. Deze vorm van schaling , geïmplementeerd in het programma PATHALS, is niet alleen steekproefafhankelijk, maar bovendien afhankelijk van het padmodel dat we gekozen hebben. Zie De Leeuw (1987). In Figuur 4 bijvoorbeeld schalen we de variabelen zodat zoveel mogelijk variantie van schoolprestaties verklaard wordt. Wanneer we het model veranderen, door wat indicatoren weg te laten, dan zullen in het algemeen de schalingen veranderen. Het is mogelijk te laten zien, dat dit niet gebeurt wanneer de regressies door schaling gelineariseerd kunnen worden. We vinden dan altijd de lineariserende schalingen, onafhankelijk van het model dat we hanteren. Maar in het algemeen zal er modelafhankelijkheid zijn. En eigenlijk is dat ook zoals het hoort. Wanneer we niet voldoende a priori informatie hebben om een schaling vast te leggen, dan is het niet verwonderlijk dat verschillende doeleinden verschillende schalingen opleveren. Wanneer we inkomen willen voorspellen uit religie, dan zullen we een andere schaling van religie vinden dan wanneer we bijbelvastheid willen voorspellen. Een model dat schoolsucces relateert aan sociaal milieu gebruikt een ander aspect van sociaal milieu, en dus een andere schaling, als een model dat sociaal milieu relateert aan gebruik van luxe consumptiegoederen.

CONCLUSIES

We hebben in dit artikel laten zien dat de aanpak van multivariate analyse, die het uitgangspunt vormt van Van de Geer (1971), op diverse manieren uitgebreid kan worden. De voordelen van padmodellen, of van de LISREL-aanpak van sociaal-wetenschappelijke theorievorming, zijn door veel auteurs nogal breed uitgemeten (bijvoorbeeld Saris en Stronkhorst, 1983). Het is duidelijk dat padmodellen naadloos aansluiten bij het

causaliteitsdenken in de sociologie en econometrie, en bij het gebruik van latente variabelen dat uit de psychometrie afkomstig is. We hebben laten zien dat er toch behoorlijk veel gevaren aan het gebruik van een programma als LISREL kleven. Weliswaar moet men de causale vooronderstellingen die men heeft duidelijk operationaliseren in de vorm van een pijldiagram, en dat is een voordeel, maar daarnaast is de feitelijk keuze van dat diagram binnen de klasse van mogelijke pijldiagrammen dikwijls in hoge mate arbitrair. Het onderscheidend vermogen van de statistische toets is in het algemeen niet van dien aard, dat we aangrenzende modellen goed kunnen onderscheiden, en daardoor ontstaat een suggestie van precisie en exactheid die niet waargemaakt kan worden.

We hebben ook gezien dat in veel sociaal-wetenschappelijk onderzoek de veronderstelling van normaal verdeelde manifeste variabelen niet vol te houden is, maar dat dit technische bezwaar ondervangen kan worden door ADF-methoden te gebruiken. Daarvoor hebben we echter nog grotere steekproeven nodig, en in veel gevallen zijn die niet beschikbaar. Of, anders gezegd, ook de statistische theorie waarop LISREL en de ADF-methoden gebaseerd zijn werkt mee om een veelal spurieuze precisie te suggereren. De vereiste vooronderstellingen gaan niet op, dus de toetsen en betrouwbaarheidsintervallen die we berekenen zijn niet valide.

In een groot aantal gevallen gaat zelfs de vooronderstelling van lineaire regressies niet op. In dat geval kunnen we twee dingen doen. We kunnen optimale schalingsmethoden gebruiken, die uitgaan van de algemenere assumptie van lineariseerbare regressies. Theoretisch levert dit weinig problemen op, en ook wat berekeningen betreft is het nauwelijks meer werk dan ADF op niet gekwantificeerde variabelen. Maar alweer: waarschijnlijk moeten de steekproeven die nodig zijn om betrouwbare statistische uitspraken te kunnen doen weer groter zijn. Een tweede mogelijkheid is om de pretentie op te geven dat we naar statistisch generaliseerbare uitspraken op zoek zijn, en om ons te beperken tot exploratief onderzoek van de steekproef die we toevallig gevonden hebben. We kunnen dan zonder gewetensbeperkingen op zoek gaan naar alle mogelijke verbanden in onze multivariabele, maar het is duidelijk dat het gevaar van kanskapitalisatie, dat wil zeggen het ontdekken van verbanden die alleen maar toevallig zijn, groot is.

En in feite schetst de keuze tussen exploratie en confirmatie het dilemma van alle data-analyse, en van de multivariate data-analyse met behulp van padmodellen in het bijzonder. Wanneer we zeer veel scherpe vooronderstellingen maken kunnen we, volgens de gangbare statistische theorie, in ieder geval een groot onderscheidend vermogen verwachten. En wanneer we weinig vooronderstellingen maken, dan kapitaliseren we op kans, en hebben we zeer grote steekproeven nodig om nog een beetje power te bereiken. Ongelukkigerwijs lijkt het er op, dat de gecompliceerde padmodellen met latente variabelen een groot aantal vooronderstellingen nodig hebben om statistisch getoetst te kunnen worden, terwijl niettemin het onderscheidend vermogen gering is. Het blijkt mogelijk om modellen op te bouwen die gebaseerd zijn op assumpties die allemaal aantoonbaar onjuist, of zelfs absurd zijn, en die toch goed bij de gegevens passen. De psychometrische genetica (De Leeuw, 1981) levert wat dat betreft vele goede voorbeelden.

Het is daarom misschien het beste, en in ieder geval het eerlijkste, om toe te geven dat padanalytische modellen exploratief gebruikt dienen te worden, en dat de statistische theorie in veel gevallen alleen maar een precisie suggereert die er in feite helemaal niet is. Vanzelfsprekend is het in deze context het beste om naar globale modellen te kijken, die de variabelen indelen in groepen, en die niet proberen om van iedere individuele padcoëfficiënt uit te maken of hij nu wel of niet nul is. En deze groepsmodellen brengen

ons dicht in de buurt van klassieke regressie en component analyse, en van de pijldiagrammen in Van de Geer (1971). Nog steeds bestaat de illusie dat sociaalwetenschappelijke theorie vanzelf te voorschijn komt, mits men de correlatiematrix maar lang genoeg analyseert, en mits het computerprogramma dat men gebruikt maar ingewikkeld en onbegrijpelijk genoeg is. Na ongeveer honderd jaar ervaring met deze aanpak lijkt het op zijn plaats om hier van hardleersheid te spreken.

LITERATUUR

Coombs, C.H. (1964). *A Theory of Data*. New York, Wiley.

De Leeuw, J. (1973). *Canonical Analysis of Categorical Data*. Dissertatie. Universiteit Leiden. Gepubliceerd (1985), DSWO Press, Leiden.

De Leeuw, J. (1978). De politieke relevantie van correlaties. *Sociologische Gids, 25*, 31-39.

De Leeuw, J. (1981). Psychometrische Genetika. In: H.C.J. Duijker & P.A. Vroon (red.), *Codex Psychologicus*. Amsterdam, Elsevier.

De Leeuw, J. (1983). Models and methods for the analysis of correlation coefficients. *Journal of Econometrics, 22*, 113-137.

De Leeuw, J. (1984a). Bespreking van Saris en Stronkhorst, Causal Modelling in Nonexperimental Research. *Mens en Maatschappij, 60*, 421-423.

De Leeuw, J. (1984b). *Statistical properties of multiple correspondence analysis*. Rapport RR-84-06. Vakgroep Datatheorie, Universiteit Leiden.

De Leeuw, J. (1985). Review of four books on causal analysis. *Psychometrika, 50*, 371-374.

De Leeuw, J. (1987). Nonlinear Path Analysis with optimal scaling. In: P. Legendre en L. Legrende (red.), *Numerical Ecology*, Berlin, Springer.

Gifi, A. (1981). *Nonlinear Multivariate Analysis*. Vakgroep Datatheorie, Universiteit Leiden.

Heiser, W. en Meulman, J. (1987). *Afstandsmodellen voor multivariate analyse* (deze bundel).

Jöreskog, K.G. en Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*, 631-639.

Jöreskog, K.G. en Sörbom, D. (1984). *LISREL VI. Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods*. Mooresville, Scientific Software.

Mooijaart, A. (1978). *Latent Structure Analysis*. Dissertatie, Universiteit Leiden.

Mooijaart, A. (1985a). A note on computational efficiency in asymptotically distribution-free correlation models. *British Journal of Mathematical ans Statistical Psychology, 38*, 112-115.

Mooijaart, A. (1985b). Factor analysis for non-normal variables. *Psychometrika, 50*, 323-342.

Mooijaart, A. en Bentler, P.M. (1985a). The weight matrix in asymptotic distribution-free methods. *British Journal of Mathematical and Statistical Psychology, 38*, 190-196.

Mooijaart, A. en Bentler, P.M. (1985b). Random polynomial factor analysis. *Proceedings of the Fourth International Symposium on Data Analysis and Informatics*.

Mooijaart, A. en Bentler, P.M. (1987). Robustness of normal theory statistics in structural equation models. Research Report PRM 87- Department of Psychology. Leiden

University.
Saris, W.E. en Stronkhorst, H. (1984). *Causal Modelling in Nonexperimental Research*. Amsterdam, Sociometric Research Foundation.
Van de Geer, J.P. (1967). *Inleiding in de Multivariate Analyse*. Arnhem, Van Loghum Slaterus.
Van de Geer, J.P. (1971). *Introduction to Multivariate Analysis for the Social Sciences*. San Francisco, Freeman.
Van de Geer, J.P. (1986). *Introduction to Linear Multivariate Data Analysis*. Leiden, DSWO Press.