

DATA-ANALYSE EN MODELKEUZE: EEN INDELING VAN STANDAARD-ANALYSE-
TECHNIEKEN IN MULTIVARIAAT EN MEERDIMENSIONEEL

A.C.M. SCHAKENRAAD

Samenvatting

Bij data-analyse in experimenteel en niet-experimenteel onderzoek heeft men de beschikking over een omvangrijk arsenaal aan standaard-modellen. In dit overzichtsartikel wordt een verband gelegd tussen standaard-modellen en de mathematisch-statistische theorie die eraan ten grondslag ligt en toepassingsvelden van data-analyse, te weten data-analyse als "criterium" en data-analyse als "techniek".

Aan het gegeneraliseerde lineaire model, dat zowel lineaire als discrete multivariate technieken omvat, ligt een mathematisch-statistische theorie ten grondslag. Dit maakt deze technieken geschikt voor verklarend onderzoek (data-analyse als "criterium"). Bij verdelingsvrije multivariate technieken en meerdimensionele methoden ontbreekt een waarschijnlijkheidsmodel. Deze technieken bewijzen hun nut in exploratief onderzoek (data-analyse als "techniek").

A.C.M. SCHAKENRAAD
MERIT
RIJKSUNIVERSITEIT LIMBURG
POSTBUS 616
6200 MD MAASTRICHT

1. INLEIDING

De grote bekendheid van multivariate en meerdimensionele technieken heeft geleid tot een groot aantal publicaties waarin de resultaten van uiteenlopende technieken vergeleken worden, of waarin aangegeven wordt, in hoeverre sommige technieken gecombineerd kunnen worden. Enkele voorbeelden: Peters (1986) geeft aan hoe de resultaten van meerdimensionele schaling en hiërarchische clusteranalyse gecombineerd kunnen worden. De resultaten van clusteranalyse, principale componentenanalyse en correspondentieanalyse worden vergeleken in Eppink/Van Dinteren (1986). Gillespie (1977) vergelijkt log-lineaire technieken met dummie-regressieanalyse. Ook een titel als "correspondence analysis used complementary to log-linear analysis" (Van der Heijden/De Leeuw, 1985) is relevant in deze context. Blijkbaar is het mogelijk om de resultaten van diverse technieken aan elkaar te koppelen. Men moet echter niet uit het oog verliezen dat data-analysetechnieken op totaal verschillende waarschijnlijkheidsmodellen kunnen berusten en andersoortige informatie kunnen verschaffen over de structuur van het datamateriaal. Het besef hoe een aantal alternatieven kunnen worden vergeleken bij de modelkeuze vormt de rode draad in Eppink (1988).

Doel van dit artikel is enige ordening aan te brengen in het grote arsenaal aan standaard data-analysetechnieken. De mathematisch-statistische basis van de betreffende techniek zal als indelingscriterium fungeren. De gepresenteerde indeling sluit aan bij belangrijke definities van data-analyse.

Laten we van start gaan met een bespreking van een aantal visies op data-analyse. Daarna volgt een mogelijke indeling van multivariate data-analysetechnieken. In dit verband worden lineaire, discrete en verdelingsvrije multivariate technieken besproken. Los van de multivariate technieken staan de meerdimensionele technieken. Deze krijgen in een aparte paragraaf aandacht.

2. DATA-ANALYSE

Data-analyse betekent op de eerste plaats, de term zegt het al, het analyseren van data. Als zodanig is data-analyse een fase in het onderzoeksproces. Het begrip data-analyse kan echter ook in pregnante zin gebruikt worden. In dat geval impliceert data-analyse een visie op de wijze waarop data geanalyseerd worden. De pregnante betekenis van data-analyse schuilt vaak in definities. Zo definieert Segers (1977) data-analyse als de analyse van variabelen en hun onderlinge relaties. Deze definitie impliceert een tamelijk forse inperking van het domein der data-analyse, aangezien data volgens deze

definitie noodzakelijkerwijze in de vorm van variabelen tot ons dienen te komen. Een analyse op basis van gelijkenissen tussen onderzoekseenheden zou niet onder de noemer van data-analyse vallen.

Meerling (1981) ziet data-analyse als het zoeken naar een afbeelding van empirische relaties op een formeel model. Een definitie, die een ruimer toepassingsgebied mogelijk maakt dan eerstgenoemde. Meerling wijst er vervolgens op, dat data-analyse twee soorten toepassingen kent: in het ene soort stelt men via toetsing de houdbaarheid van een formele theorie vast - men noemt dit het gebruik van data-analyse als "criterium" - en in het andere soort is het doel uitsluitend het afbeelden van data waarbij houdbaarheid van een formele theorie niet ter discussie staat - data-analyse wordt dan als "techniek" gebruikt.¹

Beide toepassingsmogelijkheden van data-analyse vinden we terug in het werk van De Leeuw. Deze auteur geeft het onderscheid tussen beide toepassingsmogelijkheden als volgt weer: data-analyse als "techniek" "... does not fit an explicit algebraic or stochastic model to the data by estimating free parameters, but simply tries to replace a lot of unstructured input data by a smaller amount of possibly more structured data (without too much loss of information, or with maximum reproducibility)".²

Swanborn geeft beide toepassingsvelden beknopt weer in de volgende definitie: "Data-analyse heeft meestal de vorm van een reductie van de onoverzichtelijke hoeveelheid informatie die in de data-matrix ligt opgesloten, teneinde tot een bondige uitspraak te komen, of om een aan het begin van het onderzoek geformuleerde uitspraak te toetsen".³

Men is niet genoodzaakt tot het maken van een keuze tussen data-analyse als "criterium" en data-analyse als "techniek". Een combinatie van beide benaderingswijzen in één onderzoek behoort ook tot de mogelijkheden. Zo kan men voorafgaand aan het toetsen van hypothesen globaal inzicht in de dataset verkrijgen door de data in een tweedimensionele afbeelding te presenteren. Met name Tukey zal een voorstander zijn van een dergelijke aanpak. Voor Tukey is data-analyse meer dan inferentiële statistiek alleen. Hij definieert data-analyse zodanig dat beide toepassingsvelden onvermijdelijk deel uitmaken van data-analyse, want data-analyse dient volgens hem tenminste de volgende aspecten te bevatten: "... procedures for analyzing data, techniques for interpreting the

¹Meerling (1981), p. 23-24. Cf. de Leeuw (1973), p. 16

²De Leeuw (1973), p. 14

³Swanborn (1981), p. 38

results of such procedures, ... , all the machinery and results of (mathematical) statistics which apply to analyzing data".⁴

We zijn toegekomen aan het schema van multivariate technieken. Uit dat schema kan men in de meeste gevallen eenvoudig afleiden, welke multivariate technieken aansluiten bij data-analyse als "techniek" en welke bij data-analyse als "criterium".

3. MULTIVARIATE ANALYSE

Een speurtocht door recente werken, waarin multivariate technieken centraal staan, levert ons een groot aantal indelingen en schema's op.⁵ Kenmerkend van alle gevonden indelingen is dat slechts deelverzamelingen van de multivariate technieken ingedeeld worden. Gelet op de doelstelling van die werken - in het gros wordt slechts een specifieke groep multivariate technieken onderzocht - is dat niet vreemd. In de door ons gepresenteerde indeling krijgt daarentegen een ruimer aantal gangbare technieken een plaats. Minder bekende multivariate technieken zoals bayesiaanse factoranalyse zullen echter buiten de indeling vallen (Euverman/Vermulst, 1983).

Startpunt van het schema vormt de inhoud van de datamatrix. Men kan een scheiding aanbrengen in de verzameling methoden en technieken naar gelang de informatie in de datamatrix. Indien het informatie over één of meer variabelen betreft, of indien het samengevatte informatie over variabelen betreft - hierbij valt te denken aan correlatiecoëfficiënten - dan ligt een tweede indelingscriterium voor de hand: het aantal variabelen dat de datamatrix bevat. Men kan nu univariate, bivariate en multivariate methoden en technieken onderscheiden.

In univariate analyse staat telkens één variabele in het middelpunt van de belangstelling. De betrokken variabele kan beschreven worden volgens een aantal verdelingskarakteristieken zoals maten van centrale tendentie en spreidingsmaten. Naast het beschrijven van een variabele kan men ook hypothesen met betrekking tot de variabele onderzoeken, zoals: "Is het gemiddelde in de steekproef een betrouwbare benadering van het gemiddelde in de populatie?"; of: "Verschillen

⁴Tukey, geciteerd door De Leeuw (1973), p. 10 en door Gifi (1981), p. 19

⁵Bijvoorbeeld: Brouwer et al. (1983), p. 1; Fienberg (1980), p. 3; Gifi (1981), p. 37-51; Hagenaars/Segers (1980), p. XI-XIV; Lammers (1983), p. 2-10; Lammers (1985), p. 1; Meerling (1981), p. 7; Pedhazur (1982), p. IX-X; Saris/Stronkhorst (1984), p. 286-302; Van de Wollenberg (1985), p. 7; Wrigley (1985), p. 7

twee steekproeven significant in gemiddelde?". Een wel zeer elementaire vorm van univariate analyse is de berekening van percentages.

In bivariate analyse gaat het om de relatie tussen twee variabelen. Dikwijls wordt deze relatie uitgedrukt in associatiematen zoals de produkt-moment correlatiecoëfficiënt, de rangcorrelatiecoëfficiënt en de chi-kwadraat grootheid. Bivariate analyse wordt in vele onderzoeken toegepast, hoewel men de beschikking heeft over meer dan twee met elkaar in verband staande variabelen. In feite schieten afzonderlijke bivariate analyses in dergelijke situaties tekort: men is niet in staat op eenvoudige en overzichtelijke manier inzicht te verwerven in de onderlinge relaties tussen de variabelen. De uit afzonderlijke bivariate analyses verkregen informatie is fragmentarisch. De simultane relaties in de groep variabelen blijft verborgen. Multivariate analyse biedt dan soulaas.

In multivariate analyse worden de verbanden in een groep van minstens drie variabelen simultaan bestudeerd op basis van univariate, bivariate en in sommige gevallen ook multivariate verdelingskarakteristieken. Het resultaat is altijd een formeel model waarin de relaties kwantitatief uitgedrukt worden. Het formele model bestaat uit een algebraïsche vergelijking of uit een grafische weergave. Het geaccepteerde resultaat heeft per definitie geleid tot datareductie.

Gifi (1981) stelt voor om schaaltechnieken te plaatsen onder de noemer van meerdimensionele methoden. In dit verband is de beschouwing van Gifi over definities van multivariate analyse interessant. Een voorbeeld van een te beperkte definitie van multivariate analyse is: "de analyse van steekproeven met onafhankelijke waarnemingen uit een multivariate normale verdeling". Een te ruime definitie is: "de reductie van datamatrices tot formele modellen". Eerstgenoemde definitie beperkt multivariate analyse tot lineaire modellen; de tweede impliceert dat bijvoorbeeld schaaltechnieken eveneens deel uitmaken van de verzameling multivariate technieken. In een volgende paragraaf wordt kort ingegaan op gelijkenisgegevens en meerdimensionele schaaltechnieken.

We nemen de draad van de indeling weer op. Een vaak gehanteerde indeling van multivariate methoden berust op het meetniveau van de variabelen. Zo'n indeling staat tamelijk ver van de realiteit van met name niet-experimenteel onderzoek waarin:

1. het gekozen meetniveau bij de operationalisering dikwijls arbitrair is in die situaties waarin op meerdere meetniveaus gemeten kan worden;
2. de resultaten van het gekozen meetniveau niet zelden onderhevig zijn aan transformaties zoals logarithmisering; en
3. veel multivariate methoden niet vereisen dat alle variabelen hetzelfde meetniveau bezitten, maar enige flexibiliteit bezitten in dit opzicht.

Het meetniveau wordt als indelingscriterium voor multivariate methoden overvleugeld door een essentiële criterium: de mathematisch-statistische theorie die ten grondslag ligt aan multivariate methoden. Men kan de mathematisch-statistische theorie indelen in een gedeelte, dat betrekking heeft op de steekproef - het steekproefmodel - en een gedeelte, waarin op basis van de resultaten uit het steekproefmodel het formele model gevormd wordt. Dit gedeelte wordt het verklaringsmodel genoemd (zie bijvoorbeeld Lammers, 1983).

De gemeenschappelijke verdeling van de variabelen in de populatie determineert de kans op een bepaald scorepatroon van een waarnemingseenheid in de steekproef. De populatieverdeling is echter onbekend. Een steekproefmodel geeft aan hoe waarschijnlijk, of hoe aannemelijk, het scorepatroon in de populatie is, gelet op het aantal malen dat dit scorepatroon in de steekproef waargenomen wordt. Een verklaringsmodel leidt hieruit een formeel model af met parameters en hun significantie. Bij het schema van multivariate methoden hebben we de aanwezigheid van mathematisch-statistische theorie als indelingscriterium genomen.

Het schema verdeelt multivariate methoden simpelweg in drie groepen. Op de eerste plaats worden lineaire multivariate technieken onderscheiden. De mathematisch-statistische theorie die aan dit soort technieken ten grondslag ligt, is gebaseerd op eigenschappen van de multivariate normale verdeling. Op de tweede plaats vinden we kwalitatieve of discrete multivariate methoden. Hier vormt multinomialiteit de basis van de mathematisch-statistische theorie. De laatste groep multivariate technieken bevat technieken die gekenmerkt worden door het ontbreken van een mathematisch-statistische theorie. We zullen deze daarom voortaan aanduiden als verdelingsvrije multivariate technieken.⁶

3.1 LINEAIRE MULTIVARIATE ANALYSE

Eigenschappen van de multivariate normale verdeling - we noemen deze in het vervolg de multinormale verdeling - liggen ten grondslag aan de mathematisch-statistische theorie van lineaire multivariate methoden. De multinormale verdeling heeft als verdelingsparameters μ , Σ en ρ . Deze verdeling impliceert dat alle conditionele en marginale verdelingen eveneens normaal zijn. Het omgekeerde is niet zonder meer waar. Een normale verdeling voor elk van de

⁶Een vierde groep kan gevormd worden door bayesiaanse multivariate technieken, waarvan er enige op grondige wijze besproken worden in Euverman/Vermulst (1983). De mathematisch-statistische theorie is opgebouwd rond het theorema van Bayes en de bayesiaanse statistiek.

betrokken variabelen is een noodzakelijke, doch geen voldoende voorwaarde voor een multinormale verdeling. Er moet in zo'n geval namelijk nog aan een tweede conditie voldaan zijn: een onderlinge onafhankelijkheid der variabelen.

Naast genoemde kenmerken heeft een multinormale verdeling zeer belangrijke eigenschappen met betrekking tot lineaire transformaties en statistische onafhankelijkheid. Stel dat een verdeling multinormaal is, dan gelden de volgende eigenschappen.

Allereerst is er een eigenschap die te maken heeft met lineaire transformaties. Een lineaire transformatie is een bewerking van de waarden van een variabele door bij alle waarden een constante, ongelijk aan nul, op te tellen, en/of alle waarden met zo'n constante te vermenigvuldigen. Door lineaire transformaties uit te voeren op variabelen die gezamenlijk multinormaal verdeeld zijn, kan een nieuwe multinormale verdeling geconstrueerd worden, waarin de nieuwe variabelen, die nu factoren of componenten heten, onderling statistisch onafhankelijk zijn. Op dit principe berusten factoranalytische modellen.

Een tweede eigenschap heeft te maken met eventuele statistische onafhankelijkheid in de set variabelen. Indien de variabelen multinormaal verdeeld zijn én tevens onderling statistisch onafhankelijk, dan geldt:

1. dat de gemeenschappelijke kansdichtheidsfunctie uitgedrukt kan worden als het produkt van de marginale kansdichtheidsfuncties;
2. dat de som van de variabelen een normaal verdeelde variabele is;
3. dat de som van lineair getransformeerde variabelen een normaal verdeelde variabele oplevert.

Met deze punten moeten de begrippen lineariteit en additiviteit in verband gebracht worden. Lineariteit verwijst naar een lineaire functie tussen bijvoorbeeld de variabelen X en Y . Een functie, die weer te geven is met een rechte lijn. De grootte van het effect van X op Y hangt in zo'n geval niet af van de waarde van X . Additiviteit houdt in dat de waarde van een variabele uit te drukken is als een optelsom van andere, eventueel gewogen, variabelen die onderling onafhankelijk zijn. De grootte van het effect van X_1 op Y hangt dan niet af van de waarde van X_2 . Lineariteit en additiviteit vinden we onder andere terug bij regressiemodellen.⁷

Lineaire multivariate methoden vereisen dat de variabelen tenminste op intervalniveau gemeten zijn. De lengte van de intervallen moet bekend zijn; zo

⁷De besproken theorie heeft nu duidelijk gemaakt, waarom het verstandig is een principale componentenanalyse of een factoranalyse vooraf te laten gaan aan een regressieanalyse.

mag de frequent gehanteerde ordinale vijfpuntsschaal lopend van "zeer mee eens" (waarde 1) tot "zeer mee oneens" (waarde 5) strikt genomen niet in een lineaire methode gebruiken, tenzij de lengte van de intervallen gecontroleerd is en eventueel aangepast aan een normale verdeling.

Tot slot: de veronderstelling is dat uit de populatie een steekproef getrokken wordt waarin de steekproefelementen onafhankelijk zijn. Dat wil zeggen dat de kans om geselecteerd te worden voor ieder element even groot is. Dit is het principe van de bekende "random" steekproef.

De assumpties van lineaire technieken duidt men ook wel aan als lineariteit, additiviteit, normaliteit, interval of ratio meetniveau en "random" steekproef. Deze assumpties staan nauw met elkaar in verband.

In de praktijk wordt de toepassing van lineaire multivariate technieken niet zelden ondermijnd door schending van de assumpties en door bezwaren op onderzoeksmethodologisch terrein.

Wat de normaliteitseis betreft: aan deze eis is meestal alleen voldaan na transformaties van de waarden om normale en bivariaat normale verdelingen te creëren. De betekenis van multinormaliteit ligt echter vooral op het theoretische vlak. De aard van de waargenomen multivariate verdeling in de steekproef hoeft echter nog geen struikelblok voor lineaire multivariate analyse te zijn. Een forse afwijking van multinormaliteit kan onder meer gevolgen hebben voor de significantie van de parameters en de "fit" van het model, maar hierdoor hoeven de resultaten nog niet per se onbruikbaar te zijn. Eén en ander hangt af van welke lineaire methode men gebruikt en van de schattingsprocedure. In het algemeen kan men stellen dat een minder efficiënt gebruik van de data gemaakt wordt bij toepassing van lineaire methoden op niet-normaal verdeelde variabelen. Dit komt omdat gemiddelde, standaardafwijking en correlatie, waar de berekeningen van lineaire methoden op gebaseerd zijn, in zo'n geval minder geschikte verdelingsparameters zijn.

Wat de lineariteit betreft: (die overigens direct samenhangt met multinormaliteit) scheef verdeelde variabelen kunnen wijzen op een kwadratisch in plaats van een lineair verband. In regressiemodellen is dit nog enigszins op te vangen door invoering van een kwadratische term, andere lineaire modellen bezitten deze mogelijkheid niet.

Ter afsluiting van deze paragraaf noemen we de belangrijkste multivariate lineaire methoden: de regressiemodellen (multiële regressie, padanalyse, logistische regressie, multivariate multiële regressie), factoranalytische modellen (factoranalyse, principale componentenanalyse), lineaire structurele modellen, discriminantanalyse, canonische analyse, variantie- en covariantie-analyse.

3.2 DISCRETE MULTIVARIATE ANALYSE

De mathematisch-statistische theorie van de groep discrete multivariate methoden is aan minder voorwaarden gebonden dan die van de lineaire methoden. De assumpties met betrekking tot steekproef- en verklaringsmodel vallen af. Wel komen er enkele andere assumpties voor in de plaats.

Laten we loglineaire modellen als voorbeeld nemen. Als men inferentieel bezig wil zijn, dan moet aan deze modellen een multinomiaal, een produkt-multinomiaal of een poisson steekproefmodel ten grondslag liggen.⁸ Het maakt voor de interpretatie van de geaccepteerde formele modellen nauwelijks verschil volgens welke van de drie genoemde steekproefmodellen de steekproeftrekking feitelijk uitgevoerd is.

Het asymptotische karakter vormt een gevaarlijke eigenschap van genoemde steekproefmodellen. Asymptotisch wil zeggen, dat de verdeling in de populatie pas benaderd wordt bij een steekproef van voldoende omvang. Dit heeft gevolgen voor de inferentiële aspecten van loglineaire modellen: bij een toenemend aantal steekproefelementen wijzigt de significantie van het formele model zich in positieve zin, omdat de toetsingsgrootheid sterk beïnvloed wordt door de steekproefomvang. Met andere woorden: mits de steekproef voldoende groot is dient zich altijd wel een significant model aan.

Een tweede nadelig aspect, ditmaal betrekking hebbend op het verklaringsmodel, is het gebruik van elke celfrequentie - ofwel van elke conditionele verdeling - in het volledige ("verzadigde") model. Of, indien het niet-volledige ("niet-verzadigde") modellen betreft, het gebruik van een groot aantal celfrequenties - ofwel van een groot aantal conditionele verdelingen. In niet-volledige modellen speelt niet elke celfrequentie een rol omdat over één of meer dimensies (variabelen) van de kruistabel gesommeerd is. Nu spelen dus ook marginale verdelingen een rol in het verklaringsmodel. Welke conditionele en marginale verdelingen gebruikt worden in het verklaringsmodel hangt af van het gepostuleerde formele model. Wat we in deze alinea willen zeggen is het volgende. Loglineaire technieken zijn niet noodzakelijkerwijze aangewezen op univariate en bivariate verdelingskarakteristieken; ook hogere orde

⁸Hieruit blijkt onmiddellijk dat het om de analyse van kwalitatieve variabelen gaat. Extensies met kwantitatieve variabelen (covariaten) zijn echter uitvoerbaar.

(conditionele) verdelingen kunnen een rol spelen.⁹ In de volgende alinea gaan we in op de gevaren hiervan.

Loglineaire technieken maken in vergelijking met lineaire technieken een efficiënter gebruik van de data omdat eerstgenoemde technieken naast univariate en bivariate meestal ook gebruik maken van multivariate verdelingskarakteristieken. Het nadeel hiervan is evenwel dat men rekening moet gaan houden met lage celvullingen. Deze hebben een negatieve invloed op de betrouwbaarheid van het gepostuleerde formele model.¹⁰ Conclusie: het aantal variabelen en hiermee het aantal categorieën dat effectief in de analyse opgenomen kan worden is bij loglineaire analyse aan de lage kant.

Tegenover de genoemde minpunten van loglineaire analyse, te weten het asymptotische karakter en het celvullingenprobleem, staan de volgende pluspunten: een groot aantal toepassingsmogelijkheden in vrijwel ieder onderzoek en de mogelijkheid tot het trekken van genuanceerde, statistisch onderbouwde conclusies.

Voorbeelden van discrete multivariate technieken, die mathematisch-statistisch onderbouwd zijn, zijn naast loglineaire technieken zoals het frequentiemodel, het logfrequentiemodel, het logitmodel, hybride loglineaire modellen (T-model, D-model) en het zogenaamde gemodificeerde padmodel, ook logistieke regressie met louter dummie variabelen.

3.3 VERDELINGSVRIJE MULTIVARIATE ANALYSE

Hoofdkenmerk van de groep verdelingsvrije multivariate methoden is dat de mathematisch-statistische achtergrond ontbreekt. Onderzoekselementen hoeven niet volgens een of ander steekproefmodel geselecteerd te worden. Inferentiële statistiek is meestal onmogelijk. Men kan dan ook in de regel geen uitspraken

⁹Op dit punt gekomen, refereren we naar de bespreking van de lineaire multivariate technieken. Het verklaringsmodel van lineaire technieken maakt louter gebruik van univariate en bivariate verdelingskarakteristieken (μ , σ en ρ) omdat deze karakteristieken volgens de achterliggende theorie de multivariate verdeling perfect kunnen beschrijven. Vandaar dat een computerprogramma als LISREL reeds uitvoerbaar is wanneer gemiddelde, variantie en covariantie beschikbaar zijn, of zelfs wanneer louter produkt-moment correlatiecoëfficiënten beschikbaar zijn. (Gestandaardiseerde covariantie is hetzelfde als correlatie).

¹⁰De zogenaamde structurele nulcellen nopen zelfs tot het treffen van speciale modelspecificaties. Hiermee verdwijnt globaal gesproken hun schadelijke invloed. De negatieve invloed van "random" nulcellen kan echter nauwelijks uitgeschakeld worden.

doen over de significantie van het model of van de parameters.¹¹

Verder zijn veel verdelingsvrije multivariate technieken beperkt omdat geen optimaal gebruik van de data gemaakt wordt. Een techniek als homogeniteitsanalyse is puur gebaseerd op uni- en bivariate verdelingskenmerken waarbij elke theorie over multivariate kenmerken ontbreekt. Bij de bespreking van lineaire technieken hebben we gezien, dat de multinormale verdeling zo'n theorie bij multivariate lineaire technieken mogelijk maakt. Bij de toepassing van een techniek als homogeniteitsanalyse kunnen we slechts hopen dat de bivariate verdelingen voldoende zeggen over de multivariate verdeling.

De beperktheid van verdelingsvrije multivariate methoden vormt geen bezwaar als men op de doelstelling van deze methoden let. Meestal is het doel van verdelingsvrije multivariate methoden het reduceren van de data tot een overzichtelijk tweedimensioneel plaatje, of tot een aansprekende netwerkstructuur. Bovendien brengt de beperktheid ook een voordeel met zich mee: "random" nulcellen vormen in de meeste gevallen geen probleem, zodat naar verhouding zeer veel variabelen en categorieën tegelijkertijd in de analyse kunnen worden opgenomen.

Voorbeelden van verdelingsvrije multivariate technieken zijn door Gifi ontwikkelde niet-lineaire technieken zoals homogeniteitsanalyse (ook meervoudige correspondentieanalyse genaamd), principale componentenanalyse voor ordinale en intervalvariabelen en analyse van profielen.¹² Andere voorbeelden zijn meerdimensionele ontvouwingstechnieken, contrastgroepenanalyse, clusteranalyse en regressie met dummies volgens het kleinste kwadratenprincipe. Verdelingsvrije multivariate toetsen zijn bijvoorbeeld de toetsen van Kruskal-Wallis, Friedman en Cochran.

Het gros van de verdelingsvrije technieken voldoet uitstekend bij toepassing van data-analyse als "techniek". Ze vormen een welkome aanvulling op de lineaire en discrete methoden, die in paragraaf 3.1 en 3.2 besproken werden. Deze passen op hun beurt tezamen met de verdelingsvrije multivariate toetsen bij de benadering van data-analyse als "criterium". Hiermee is de "link" met paragraaf 2 gelegd.

¹¹Technieken waarbij dit wel mogelijk is zijn bijvoorbeeld de toetsen van Kruskal-Wallis, Friedman en Cochran (zie bijvoorbeeld Grootenboer/Luijten 1978).

¹²De niet-lineaire extensies zorgen ervoor dat de te analyseren variabelen niet numeriek beschouwd worden, maar ordinaal of zelfs nominaal.

4. HET ALGEMENE OF GEGENERALISEERDE LINEAIRE MODEL

In paragraaf 3 werd gesproken van lineaire, discrete en verdelingsvrije multivariate technieken. In deze paragraaf willen we wijzen op een model dat direct of indirect elke lineaire of discrete multivariate techniek als speciaal geval omvat, mits aan de volgende condities is voldaan:¹³

1. Er kan een zogenaamde "link"-functie opgesteld worden. Dit is een functie die de afhankelijke variabele transformeert in een variabele die geschreven kan worden als een optelsom van eventueel gewogen effecten. De logittransformatie is zo'n "link"-functie.
2. De foutenterm ("errorterm") kan beschreven worden met een exponentiële verdelingsfunctie zoals de normale of een poisson verdeling.

Het model wordt het algemene of gegeneraliseerde model genoemd (GLM). Formeel heeft het de volgende gedaante bij m observaties en n parameters.

$$\begin{array}{ccccc} \underline{Y} & = & \underline{X} & \cdot & \underline{b} \\ (mx1) & & (mxn) & & (nx1) \end{array}$$

Het logitmodel en het klassieke regressiemodel zijn direct uit bovenstaande matrixvergelijking af te leiden. Het GLM biedt echter een theoretisch kader waaruit in feite elke lineaire en discrete multivariate techniek herleidbaar is.

Voor onze indeling van multivariate technieken betekent het dat het onderscheid in drie groepen teruggebracht kan worden tot *grosso modo* twee groepen: GLM technieken en verdelingsvrije multivariate technieken.

5. EENDIMENSIONELE EN MEERDIMENSIONELE ANALYSE

Eén van de typen onderzoeksgegevens waarover men kan beschikken, betreft gelijkennisgegevens of "similarities". Gelijkennisgegevens geven de mate aan, waarin objecten aan elkaar gelijk zijn of aan elkaar gelijk worden geacht. Ze geven als het ware de afstanden aan, die tussen objecten bestaan.

Indien men over dergelijke waarnemingen beschikt, en het doel is te komen tot enige structuur, is men in eerste instantie aangewezen op schaaltechnieken. Deze kunnen ééndimensioneel van aard zijn (bijvoorbeeld Thurstone's techniek voor het

¹³Zie bijvoorbeeld Wonnacott/Wonnacott (1970, 1979); Willekens (1984); Wrigley (1985).

schatten van schaalwaarden van stimuli via paarsgewijze vergelijking) of meerdimensioneel ("two-way" en "three-way" meerdimensionele schaling).

In metrische meerdimensionele schaling wordt ervan uitgegaan dat er een lineaire relatie bestaat tussen de gelijkenisoordelen (=psychologische afstanden) - of andersoortige gelijkenissen - en de euclidische afstanden in het formele model. In de niet-metrische variant wordt alleen de rangorde van de gelijkenissen in beschouwing genomen.

Het analyseren van gelijkenisgegevens is echter geen exclusief voorrecht van schaaltechnieken. Eén specifiek type similariteitsmatrix - namelijk een correlatiematrix - kan invoer zijn voor lineaire multivariate technieken. Op similariteitsmatrices in het algemeen kan men naast schaaltechnieken ook clusteranalyse toepassen. Volgens onze definitie moet clusteranalyse in zo'n geval niet als een multivariate techniek opgevat worden.

6. BESLUIT

Aan het begin van dit artikel constateerden we dat data-analyse twee belangrijke toepassingsvelden kent, te weten data-analyse als "techniek" en als "criterium", die overigens beide in één onderzoek verenigbaar zijn.

In het eerste toepassingsveld is het doel uitsluitend het afbeelden, het reduceren van data zonder dat de houdbaarheid van een formeel model ter discussie staat. Inferentiële aspecten zijn daar dan ook niet aan de orde.

In het tweede toepassingsveld ligt dat geheel anders. Toetsing van hypothesen vormt hier een hoofdbestanddeel.

De gepresenteerde indeling van multivariate data-analysetechnieken bleek aan te sluiten bij beide toepassingsvelden van data-analyse. De GLM-technieken, op te splitsen in lineaire en discrete multivariate technieken, kunnen een rol spelen bij data-analyse als "criterium". Verdelingsvrije methoden bewijzen hun nut bij data-analyse als "techniek".

Twee kanttekeningen moeten bij de indeling geplaatst worden. Op de eerste plaats bestaan er verdelingsvrije multivariate technieken waarmee hypothesen getoetst kunnen worden (Kruskal-Wallis). Deze technieken liggen dus in de lijn van data-analyse als "criterium". Op de tweede plaats is gebleken, dat enkele data-analyse procedures niet onder de noemer "multivariaat" geplaatst konden worden omdat ze gebaseerd zijn op gelijkenisgegevens. Het betrof een aantal schaaltechnieken die aangeduid werden als meerdimensionele methoden. Ook op gelijkenisgegevens gebaseerde ééndimensionele schaaltechnieken vallen in deze groep.

LITERATUUR

- BROUWER, F.M., NIJKAMP, P. en RIETVELD, P. (1983) Een Overzicht van Methoden voor de Behandeling van Kwalitatieve Geografische Data; in: OP 'T VELD, D., SMAGT, T. van der, TIMMERMANS, H. (red.) Waar Harde Data Ontbreken; Bohn, Scheltema en Holkema; Utrecht/Antwerpen
- DAVISON, M.L. (1983) Multidimensional Scaling; Wiley; New York/Chichester/etc.
- EPPINK, TH.W.A. and DINTEREN, J.H.G. van (1986) An application of Correspondence Analysis in Urban Geography Compared with Cluster Analysis and Principal Components Analysis; in: TIJDSCHRIFT VOOR ECONOMISCHE EN SOCIALE GEOGRAFIE; vol. 77, nr. 3, p. 213-220
- EPPINK, TH.W.A. (1988) Choice of Mathematical Models in Geographic Research Considering Alternatives; Thesis; Geografisch en Planologisch Instituut KUN; Nijmegen
- EUVERMAN, T.J. and VERMULST, A.A. (1983) Bayesian Factor Analysis; Thesis; RUG; Groningen
- FIENBERG, S.E. (1980) The Analysis of Cross-Classified Categorical Data; The MIT Press; Cambridge, Mass.
- GIFI, A. (1981) Non-Linear Multivariate Analysis; Department of Data Theory; Leiden
- GILLESPIE, M.W. (1977) Log-Linear Techniques and the Regression Analysis of Dummy Dependent Variables, Further Bases for Comparison; in: SOCIOLOGICAL METHODS AND RESEARCH, vol. 6, no. 1
- GROOTENBOER, A. en LUIJTEN, A. (1978) Verdelingsvrije Toetsen; Van Loghum Slaterus; Deventer
- HAYDUK, L.A. (1987) Structural Equation Modelling with LISREL, Essentials and advances; The Johns Hopkins University Press; Baltimore/London
- HEIJDEN, P.G.M. van der and LEEUW, J. de (1985) Correspondence Analysis Used Complementary to Log-Linear Analysis; in: PSYCHOMETRIKA, vol. 50, no. 4, p. 429-447
- LAMMERS, J.G.M. (1983) Tabelanalyse volgens Multiplicatieve Modellen Gezien vanuit een Causaal Perspectief; Dissertatie; Sociologisch Instituut KUN; Nijmegen
- LAMMERS, J.M.G. (1985) Cross-sectionele Data-analyse; Vakgroep Sociologie KUN; Nijmegen
- LEEUEW, J. de (1973) Canonical Analysis of Categorical Data; RUL; Leiden
- MEERLING (1981) Methoden en Technieken van Psychologisch Onderzoek, DEEL 2, Data-analyse en Psychometrie; Boom; Meppel/Amsterdam

- PEDHAZUR, E.J. (1982) Multiple Regression in Behavioral Research: Explanation and Prediction; Holt, Rinehart and Winston; New York/Chicago/etc.
- PETERS, V.A.M. (1986) Multidimensionele Schaling en Hiërarchische Clusteranalyse; Vakgroep Sociologie KUN; Nijmegen
- SARIS, W.E. and STRONKHORST, L.H. (1984) Causal Modelling in Nonexperimental Research: an Introduction to the LISREL approach; Sociometric Foundation; Amsterdam
- SEGRS, J.H.G. (1977) Sociologische Onderzoeksmethoden; Van Gorcum; Assen
- SEGRS, J.H.G. en HAGENAARS, J.A.P., red. (1980) Sociologische Onderzoeksmethoden, Technieken van Causale Analyse II; Van Gorcum; Assen
- SWANBORN, P.G. (1981) Methoden van Sociaal-wetenschappelijk onderzoek: Inleiding in Ontwerpstrategieën; Boom; Meppel/Amsterdam
- WILLEKENS, F.J. (1984) Loglineaire en Gegeneraliseerde Lineaire Modellen in Sociologisch Onderzoek; Intern Rapport nr. 31; NIDI; Voorburg
- WOLLENBERG, A.L. van de (1984) Multivariate Analyse Technieken; Vakgroep Mathematische Psychologie, Psychologisch Laboratorium KUN; Nijmegen
- WONNACOTT, R.J. and WONNACOTT, T.H. (1970, 1979) Econometrics; Wiley; New York/Chichester/etc.
- WRIGLEY, N. (1985) Categorical Data Analysis for Geographers and Environmental Scientists; Longman; London/New York

Ontvangen: 05-06-1988

Geaccepteerd: 20-04-1989