DISCUSSIE OVER ONDERWIJS IN MULTIVARIATE ANALYSE II

J. de Leeuw

— 1 —

Dit korte opstel is bedoeld om het boek van Van de Geer (1971) te plaatsen in het geheel van technieken die in de data analyse aangeduid worden als multivariate analyse (voortaan MVA).

— 2 —

Hiertoe maken we gebruik van een tweetal konceptuele dimensies. In de eerste plaats is er de onderscheiding tussen konfirmatief en eksploratief gebruik van MVA. Als we konfirmerend bezig zijn postuleren we een bepaald model en toetsen dit door gebruik te maken van geobserveerde gegevens. Eksploratie gaat hier als het ware aan vooraf. Een stochastisch model is bij eksploratie niet noodzakelijk aanwezig (of hoogstens impliciet aanwezig). Diverse transformaties, plottechnieken, en redukties van de gegevens worden gebruikt om systematische verbanden te vinden (die dan vervolgens konfirmatief onderzocht kunnen worden, gebruik makend van een andere verzameling gegevens).

— 3 —

De tweede dimensie is klassiek en heeft betrekking op het schaalniveau dat de variabelen tenminste moeten hebben. We onderscheiden (in toenemende algemeenheid) de intervalschaal, de ordinale schaal en de nominale schaal.

— 4 —

Dit geeft het schema:

	Interval	Ordinaal	Nominaal
Exploratief	EXINT →	EXORD →	EXNOM
	↓	↓	↓
Konfirmatief	KONINT →	KONORD →	KONNOM

— 5 —

In deze dimensie ligt een zekere ordening besloten. In de eerste plaats kunnen nominale technieken gebruikt worden in de speciale gevallen waarin we ordinale en/of interval informatie hebben. Ordinale technieken kunnen gebruikt worden op interval data. En tenslotte kunnen konfir-

matieve technieken voor eksploratie gebruikt worden. Dit geeft een partiële orde in termen van algemeenheid, die in het schema aangeduid is met pijlen. De EXINT technieken hebben de meest beperkte toepasbaarheid, de KONNOM technieken zijn het meest algemeen.

— 6 —

Natuurlijk zijn de grenzen in dit schema meer vloeiend dan uit het schema zelf blijkt. Eksploratieve technieken kunnen konfirmatief gemaakt worden door toevoeging van een (ad hoc) stochastisch (deel)model, eksploratieve technieken worden voor 'intuïtieve' konfirmatie gebruikt, konfirmatieve technieken zoals gelijktijdige betrouwbaarheidsintervallen en gelijktijdige testprocedures zijn bij uitstek geschikt voor gekontroleerde eksploratie. Binnen bepaalde technieken kan het konfirmatieve aspect min of meer specifiek zijn: er zijn structuren van zwakkere en sterkere hypothesen die eksploratief gebruikt worden omdat de stochastische interactie van de diverse statistieken niet geheel te overzien is.

— 7 —

Een derde, meer technische, dimensie onderscheidt de konfirmatieve technieken in eksakte en benaderende procedures. Eksakte procedures hebben bepaalde optimale eigenschappen, onafhankelijk van de grootte van de steekproef, benaderende technieken hebben deze optimale eigenschappen slechts asymptotisch en zijn dus over het algemeen slechts bruikbaar als de steekproef (zeer) groot is. Over het algemeen hebben benaderende procedures het voordeel dat ze gebruikt kunnen worden om minder specifieke (bijvoorbeeld niet-parametrische) modellen te analyseren. Het gebrek aan strukturele aannamen wordt als het ware gekompenseerd door de grootte van de steekproef, die de grenswaarde stellingen van de waarschijnlijkheidsleer toepasbaar maakt. De eksploratieve technieken zou men op deze dimensie kunnen karakteriseren door te stellen dat we daar net doen alsof we te maken hebben met de populatie. Er is geen stochastisch model nodig, alle eigenschappen van de gegevens zijn statistisch signifikant.

— 8 —

Het boek van Van de Geer onderscheidt zich van de andere boeken over MVA omdat het vrijwel uitsluitend gewijd is aan de EXINT technieken, of, eksakter, aan de EXINT aspekten van diverse klassieke technieken. Er zijn weliswaar een aantal boeken over faktoranalyse en padanalyse die ook voor bijna honderd procent EXINT zijn, er zijn weliswaar een aantal boeken en artikelen over regressie en komponent analyse die een grote plaats inruimen voor typisch eksploratieve overwegingen,

maar voorzover ik weet is er geen enkel algemeen boek over MVA dat zo kompleet en rigoreus de konfirmatieve aspekten van MVA negeert. De behandeling van de multinormaal verdeling, en faktoranalyse op basis van maximale aannemelijkheid zijn schijnbare uitzonderingen, maar zij vallen dan ook duidelijk buiten de opzet van het boek. In het verloop van dit opstel zullen we de zes verschillende vormen van MVA apart bespreken, tezamen met de verschillende algemene principes ('ingangen') die als hulpmiddel gebruikt (kunnen) worden om het gebied van de MVA geheel of gedeeltelijk onder één noemer te brengen. Vele van deze ingangen zijn eksploratief van aard, en worden in het boek van Van de Geer besproken of tenminste aangeduid.

— 9 —

Zo is een van de meest voor de hand liggende overeenkomsten tussen de diverse EXINT technieken dat ze geformuleerd kunnen worden als eigenwaarde-eigenvektor problemen. De overeenkomstige KONINT interpretatie is dat ze geformuleerd kunnen worden als speciale gevallen van kanonische analyse in de zin van Hotelling.

Het belangrijkste verschil in deze kontekst tussen de psychometrikus (of sociaal wetenschappelijke data theoreticus) en zijn EXINT techniek enerzijds en de statistikus en zijn KONINT techniek anderzijds is het volgende: de eerste bekijkt de orthogonale komponenten apart en tracht ze te interpreteren (en op basis van een niet te formaliseren foutentheorie bekend als 'root staring' de onbelangrijke komponenten negeert), terwijl de tweede statistieken berekent die het gehele systeem van komponenten samenvatten ten einde een van te voren geformuleerde hypothese van globale onafhankelijkheid te toetsen. (Een zelfde onderscheid kan ook gemaakt worden tussen EXORD en KONORD en tussen EXNOM en KONNOM technieken).

Tussenvormen gebaseerd op logische analyse van hypothesen en overeenkomstige partitionering van globale statistieken in (additieve, onafhankelijke) komponenten zijn evenwel mogelijk.

Door de gekompliceerde stochastische relaties die hier een rol gaan spelen en door de veelheid van mogelijke beslissingen liggen deze tussenvormen in het eerder genoemde grensgebied tussen konfirmatie en eksploratie. Een behandeling van MVA gebaseerd op deze tussenvormen wordt verdedigd door Koornstra in zijn bespreking van Van de Geer's boek.

— 10 —

In deze sectie geven we een bespreking van de zes onderscheiden vormen van MVA, met de belangrijkste ingangen. De ingangen worden

genoemd in arbitraire volgorde en zonder logische of methodologische prioriteiten aan te duiden.

EXINT:
- Het pseudo-statistische uitsplitsen van kwadratensommen in tussen en binnen komponenten.
- Het vinden van causale relaties.
- Het principe van de kleinste kwadraten (toegepast op skorematrix of op kovariantiematrix).
- De n-dimensionale lineaire meetkunde.
- Lineaire operaties (c.q. kombinaties).

EXORD:
- n-dimensionale lineaire meetkunde.
- Kleinste kwadraten.
- Monotone regressie.

EXNOM:
- Tussen - binnen.
- Kontiguiteit.
- Additieve kombinaties.
- Afstandsmaten.

KONINT:
- Multinormale verdeling.
- Maximale aannemelijkheid.
- Sporen en determinanten.
- Asymptotisch optimale schatters en tests.

KONORD:
- Permutatie tests.
- Rangorde tests.

KONNOM:
- De multinominale verdeling.
- De Neyman-Wald methode.
- De Pearson-Fisher methode.
- Orthonormale funkties.

— 11 —

Gebruik makend van deze trefwoorden is het niet moeilijk het meeste werk op het gebied van MVA eenvoudig te klassificeren. De boeken van Anderson (1958) en Morrison (1961) behandelen KONINT met als voornaamste ingang de multinormale verdeling en zijn eigenschappen. De belangrijkste recente ontwikkelingen op dit gebied zijn de exacte verdelingen die door James, Gleser, Constantine, Consul, en anderen gevonden werden voor de diverse statistieken die logisch uit het beginsel van de aannemelijkheidsverhouding voortvloeien (de Wilks-Bartlett statistieken).

Het boek van Roy (1957) is wat minder konventioneel omdat het zich konsentreert op tests en meer algemene inferente procedures gebaseerd op de (grootste, kleinste) eigenwaarde(n). Recente ontwikkelingen hier zijn voornamelijk gebaseerd op het omvangrijke analytische en rekenwerk van Pillai en zijn medewerkers. Een andere belangrijke ontwikkeling in

KONINT, die niet zozeer iets nieuws brengt maar diverse zaken vanuit een nieuw algemeen gezichtspunt benadert, is de 'structurele' analyse van kovariantie matrices zoals uiteengezet door Anderson, Bock, Bargmann, Mukherjee, Jöreskog.

Speciaal geval is de faktoranalyse, waarvan de nieuwste EXINT versies behandeld worden door Harman (1967) en de nieuwste KONINT versies door Lawley & Maxwell (1971). De ontwikkelingen op gebied zijn één van de meest duidelijke illustraties van de invloed van komputertechnologie op de ontwikkeling van MVA. Ander systematiserend EXINT werk dat gebruik maakt van de kleinste kwadraten ingang is de NILES-NIPALS benadering van Wold en zijn school (dit omvat Harman's MINRES en Carroll-Chang-Harshman's CANDECOMP-PARAFAC).

— 12 —

Het in de vorige sectie genoemde boek van Roy bevat ook een inleiding in de (benaderende) Pearson-Fisher methode voor KONNOM. Ander pionierswerk op dit gebied werd gedaan door Neyman en Rao. Medewerkers van Roy zoals Mitra, Diamond, Bhapkar, Grizzle, en Ogawa hebben veel systematisch werk in KONNOM gedaan, en mensen als Good, Goodman, Lindley, Kullback, en Birch hebben dit nog geperfectioneerd.

Er is nog geen boek over KONNOM hoewel Rao (1965) er aardig wat bladzijden aan spandeert, en hoewel Goodman's recente artikelen zonder meer tot een goed en representatief boek zouden kunnen worden gebundeld. Zoals gezegd bevatten de boeken van Van de Geer de meest systematische inleiding in EXINT. EXORD omvat de 'niet-metrische' technieken gepropageerd door Shepard, Guttman, Kruskal, Carroll, Roskam, Green, Wish, Lingoes (diverse boeken van deze mensen zijn onlangs verschenen of in voorbereiding). EXNOM in de meest voor de hand liggende versie wordt behandeld door De Leeuw (1972). Laatstgenoemde systematiseert wat verspreid en niet al te bekend werk van Fisher, Guttman, Carroll, Lancaster, en anderen om (via de tussen-binnen ingang) een kompleet beeld te geven van EXNOM (met ekspliciete specialiseringen naar EXORD en EXINT). KONORD tenslotte is vrijwel uitputtend behandeld in het recente boek van Puri & Sen (1971). De meeste ontwikkelingen in de eksploratieve technieken zijn er (vanzelfsprekend) op gericht om meer konfirmatieve elementen in te passen. We hebben Koornstra's aanvulling op Van de Geer's EXINT benadering genoemd, EXNOM zou op dezelfde manier aangevuld kunnen worden door de multinominale in plaats van de multinormale verdeling te gebruiken, en waar nodig over te stappen op benaderende argumenten. De duidelijke trend in EXORD is het invoeren van konfirmatieve

aspekten door middel van Monte Carlo methoden en het ontwerpen van differentiële modellen voor drieweg-gegevens en individuele verschillen. Het lijkt mogelijk (hoewel zeker niet eenvoudig) bij deze laatste modellen de inferentiële statistiek (bijvoorbeeld de theorie van incidentele en strukturele parameters) op een natuurlijke manier te laten aansluiten.

— 13 —

Welke boeken zijn nog nodig om een zo kompleet mogelijk beeld te geven van MVA zoals het tegenwoordig beoefed wordt? De boeken van Anderson en Morrison kunnen gemakkelijk uitgebreid worden met het recente werk over exacte verdelingen en strukturele analyse. Ze blijven dan volledige en bruikbare inleidingen in multinormale KONINT.

Voor de sociale wetenschappen lijkt een kombinatie van de benadering van Van de Geer met de modifikaties in konfirmatieve richting voorgesteld door Koornstra voorlopig ideaal. Een ander zinvol boek zou te schrijven zijn door gelijktijdige strukturele analyse van multinormale dispersie matrices en vektoren van gemiddelden te kombineren met de zware machinerie van gekomputeriseerde maximale aannemelijkheid en aannemelijkheidsverhoudingen. Dit leidt tot een verregaande generalisatie van het lineaire model zoals besproken door Searle (1971) en een konseptueel eenvoudige (om niet te zeggen mechanische) methode tot het benaderend konfirmatief onderzoeken van modellen. Het accent komt te liggen op de modelbouw en op logische analyse van aannamen, de statistische aspecten blijven in het duister. Tenslotte zou een zelfde benadering van multinomiale gegevens een zeer nuttig komplement hiervan zijn (terwijl een vergelijkbare, hoewel wat meer specifieke, toepassing op Markov processen al geschreven is door Billingsley in 1961). Mogelijkerwijs is het raadzaam voor deze twee laatste projekten een aparte inleiding te schrijven in de benaderende (grote steekproeven) statistiek, hetzij in de zin van Wald-LeCam-Hajek-Roussas, hetzij in de zin van Pearson-Fisher-Neyman-Rao. Aan deze laatste projekten wordt gewerkt.

Literatuur:

T. W. Anderson: An introduction to multivariate statistical analysis. New York, Wiley, 1958.

P. Billingsley: Statistical inference for Markov Processess. Chicago, Univ. Chicago Press, 1961.

J. De Leeuw: Canonical analysis of categorical data. Leiden, Psychol. Institute, 1973.

H. H. Harman: Modern factor analysis. Chicago, Univ. Chicago Press, 1967 (2e edition).

D. N. Lawley & A. E. Maxwell: Factor analysis as a statistical method. London, Butterworths, 1971 (2e edition).

D. F. Morrison: Multivariate statistical methods. New York, McGraw-Hill, 1967.

M. L. Puri & P. K. Sen: Nonparametric methods in multivariate analysis. New York, Wiley, 1971.

C. R. Rao: Linear statistical inference and its applications. New York, Wiley, 1965.

S. N. Roy: Some aspects of multivariate analysis. New York, Wiley, 1957.

S. R. Searle: Linear models. New York, Wiley, 1971.