Mededelingen

Menno Wolters
Vakgroep Theorie en Methodologie
Sociologisch Instituut
Rijksuniversiteit Utrecht

Bijdrage Voorjaarsvergadering
Sociaal-Wetenschappelijke Sectie
Vereniging Voor Statistiek
30 mei 1979, Jaarbeurs, Utrecht

# EMPIRISCHE TOEPASSINGEN VAN MEERDIMENSIONALE SCHAALANALYSE

(1) In de hier volgende opmerkingen wordt het begrip *meerdimensionale schaalanalyse* gehanteerd in de ruime zin van het woord. Dat wil zeggen, dat eronder ook worden begrepen de zgn. *gewogen meerdimensionale schaalanalyse*, het meerdimensionaal *ontvouwen* enz. Op de eendimensionale schaalanalyse wordt hieronder niet verder ingegaan; bedacht moet evenwel worden dat het onderscheid van analysevormen naar de dimensionaliteit van hun uitkomsten een willekeurig onderscheid is, dat in dit geval slechts wordt gerechtvaardigd door technische verschillen.

(2) Afgezien van zuiver illustratief gebruik kan de meerdimensionale schaalanalyse op 3 manieren in wetenschappelijk onderzoek gehanteerd worden:

(a) *Exploratief*: de onderzoeker poogt zonder uit te gaan van een theorie(tje) in de data een hem onbekende structuur te vinden.
(b) *Metend*: de onderzoeker poogt op basis van een theoretisch model in de data een hem onbekende structuur te vinden.
(c) *Toetsend*: de onderzoeker poogt een theoretisch model en/of een door hem een verwachte structuur aan de data te toetsen.

Onder een theorie of theoretisch model wordt hier verstaan een stelsel van geformaliseerde relaties tussen onafhankelijke en afhankelijke variabelen, waarbij de data de afhankelijke, en de te schalen concepten de onafhankelijke variabelen vormen. Onder een structuur wordt hier verstaan de getalsmatige invulling van de onafhankelijke variabelen in de operationele context van het onderzoek.
*Voorbeeld: een theorie van politiek keuzegedrag; de structuur van de te kiezen Nederlandse politieke partijen.*

(3) Het lijstje van gebruiksmogelijkheden in (2) kan worden aangevuld met een vierde, evenwel bij mijn weten nooit gehanteerde, mogelijkheid:

(d) *Inductief*: de onderzoeker poogt een vooraf gegeven structuur in de data te vinden om aldus de best passende theorie te ontdekken.

(4) De *exploratieve* toepassing (zonder gegeven theorie, zonder gegeven structuur) komt nogal eens voor; er kleven evenwel een aantal bezwaren aan:

(a) De ervaring heeft geleerd, dat de uitkomsten altijd goed interpreteerbaar schijnen (m.a.w. haal 10 verschillende oplossingen uit 1 dataset en u heeft 10 verschillende plausibele interpretaties).
(b) Impliciet moet altijd een model gekozen worden, aangezien de computer slechts 1 bepaalde rekenprocedure kan worden opgedragen. De theoretische assumpties die in dit model besloten liggen, worden in de exploratieve toepassing evenwel niet geëxpliciteerd.
(c) Het gevaar bestaat dat de resultaten worden gebruikt alsof het een metende of inductieve toepassing geldt.

Voor exploratieve dataänalyse kan beter gebruik worden gemaakt van technieken voor directe kwantificatie waaraan een interpretatie als meerdimensionale schaalanalyse kan worden gegeven (b.v. het in Leiden ontwikkelde HOMALS).

(5) De *metende* toepassing van de meerdimensionale schaaltechnieken is het meest gebruikelijk; in het algemeen worden deze technieken behandeld in de literatuur over het meten (meettheorie e.d.). Auteurs als Shepard en Young hebben dan ook als een van de doeleinden van de meerdimensionale schaalanalyse omschreven *metrische informatie te ontlenen aan nonmetrische data*; hierbij gaat het om het ophogen van het meetniveau. In vele toepassingen wordt beoogd een onbekende structuur te meten in het licht van een apriori gestelde theorie, die overeenkomt met het gehanteerde meerdimensionale schaalmodel.
Het belangrijkste bezwaar van metend gebruik van meerdimensionale schaaltechnieken is gelegen in de onvolmaaktheid van deze technieken. Hun vermogen de meest eenvoudige structuur in gegeven data te vinden is beperkt. Veelal worden overbodige dimensies gepresenteerd, of stopt het zoekproces bij een *lokaal minimum* (suboptimale stress) of bij een *gedegenereerde oplossing* (waarin een groot aantal relaties getrivialiseerd zijn). Een bekend probleem in dit verband is het verschijnen van een *hoefijzer* als gevolg van een of meer van deze ovolkomenheden.
Aangezien ook hier geldt dat alle oplossingen (gedegenereerd, voorzien van overbodige dimensies, verkregen bij suboptimale stress, of niet) wel van een plausibele interpretatie kunnen worden voorzien, zal het duidelijk zijn dat veel meten in dit verband schijn-meten is.
Een tweede bezwaar, dat evenwel meer de toepassingen dan de mogelijkheden van het metend gebruik betreft, is dat het apriori gestelde theoretisch model veelal onvoldoende is getoetst tegen alternatieve modellen.

(6) De *toetsende* hantering van de meerdimensionale schaalanalyse heeft totnogtoe veel te weinig aandacht gekregen, niet in de laatste plaats wellicht, doordat veel van de technieken zijn ontwikkeld door meettheoretici. Hierdoor kleven er thans een tweetal bezwaren aan deze toepassingsmogelijkheid:

(a) Vele der ontwikkelde programma's bieden niet de mogelijkheid een vooraf ontwikkelde structuur in te lezen. Weliswaar is het voor de geoefende programmeur niet moeilijk op dit punt wijzigingen in het programma aan te brengen, maar vooralsnog maakt dit gebrek toetsend gebruik door de doorsnee-gebruiker niet mogelijk.

(b) Een zeer groot bezwaar is dat statistische maten en toetsen geheel ontbreken, waardoor slechts deterministisch (en niet: probabilistisch) toetsend gebruik mogelijk is, en dan nog binnen enge grenzen waar het gaat om de parsimonische keuze uit twee of meer modellen.

In feite vormen metend en toetsend gebruik de polen van een continuum, waarop dataänalyse binnen een theoretisch kader en op grond van een ten dele gegeven structuur kan worden bedreven. *Voorbeeld: een gegeven theorie van politiek keuzegedrag en een structuur van de Nederlandse politieke partijen die lokaal onbekend is.* Een bijzonder geval op dit continuum vormt de zogenaamde *externe analyse* waarin de structuur van 1 verzameling punten is gegeven en getracht wordt de structuur van een tweede verzameling punten te vinden aan de hand van data die worden voorgesteld als afstanden tussen telkens een punt uit de eerste en een punt uit de tweede verzameling.

(7) Aan de *inductieve* benadering kleven uiteraard dezelfde bezwaren als aan de toetsende benadering.

(8) Het meest algemene model dat aan de meerdimensionale schaalanalyse ten grondslag ligt kan worden geformuleerd als:

$$\delta_{ijk} \overset{opt}{=} d_{ijk} = \sqrt[r]{\sum_{a=1}^{n} w_{ia} w_{ja} w_{ka} (x_{ia} - x_{ja})^r} ,$$

waarin:

$\delta_{ijk}$ = het element in cel (i,j) van de k-de datamatrix

$\overset{opt}{=}$ = optimale overeenstemming van data en afstanden

$d_{ijk}$ = de voor i, j en k gewogen afstand tussen i en j

$r$ = de Minkowski parameter

$a$ = de a-de dimensie

$n$ = het aantal dimensies

$w_{ia}$ = het gewicht van dimensie a voor rij-element i

$w_{ja}$ = het gewicht van dimensie a voor kolom-element j

$w_{ka}$ = het gewicht van dimensie a voor matrix k

$x_{ha}$ = de coördinaatwaarde van rij- of kolomelement h op dimensie a

De data vormen een drieweg (i×j×k) matrix met i, j > 1 en k > 0.

Dit meest algemene model is nimmer in een algorithme omgezet en waarschijnlijk bevat het ook teveel parameters (is het te weinig bepaald) om dat mogelijk te maken. Het is evenwel een nuttig uitgangspunt om alle bestaande (opties van) programma's aan te relateren. In alle gevallen gaat het om vereenvoudigingen van bovenstaand model.

(a) *metrisch/nonmetrisch*: In een metrische analyse wordt identiteit van δ en d nagestreefd; in een nonmetrische analyse wordt slechts een monotoon verband tussen beide verlangd.

(b) *replicaties (k)*: In een analyse zonder replicaties is de datamatrix tweeweg (i×j); de parameter k duidt de replicaties aan. Replicaties staan b.v. voor herhaalde waarnemingen, of waarnemingen aan meer individuen, van elk van wie een tweeweg matrix wordt verkregen.

(c) *gewichten (w)*: De oudste modellen en programma's kennen geen parameters w, waarin individuele variatie van de rij- of kolomelementen of van de replicaties tot uitdrukking wordt gebracht. Later zijn modellen ontwikkeld met 1, zelden 2, van de bovengenoemde gewichtsparameters. Het waardegebied der gewichten wordt al dan niet beperkt tot > of $\geq$ 0.

(d) *conditionaliteit*: Van rij-conditionaliteit is sprake, indien de $w_i$ voor alle a (maar niet voor de i) gelijk zijn; van kolomconditionaliteit wordt gesproken indien alle $w_j$ gelijk zijn. Het begrip doelt erop, dat de data op een per rij (kolom) verschillende schaal gemeten kunnen zijn, zoals i.g.v. rating scales. Matrixconditionaliteit doelt op een soortgelijke situatie voor de gewichten der replicaties $w_k$.

(e) *partitionering*: Hetzelfde idee (dat data-elementen op diverse schalen gemeten kunnen zijn), maar niet gebonden aan rij, kolom of replicatie, is in een enkel programma uitgewerkt als partitionering: de data worden ingedeeld in onderscheiden partities.

(f) *vierkante of rechthoekige datamatrix*: Indien de elementen i en j uit dezelfde verzameling stammen (zodat tegenover elke cel (i,j) een cel (j,i) staat) is de datamatrix vierkant, in het andere geval rechthoekig. In geval van rechthoekige data wordt de analyse wel met *ontvouwing* aangeduid; het 'klassieke' model van meerdimensionale schaalanalyse had betrekking op vierkante data. Een rechthoekige (i×j) datamatrix wordt vaak opgevat als deel van een grotere vierkante((i+j)×(i+j)) matrix.

(g) *asymmetrische datamatrix*: Een door Young c.s. bestudeerd bijzonder geval van ontvouwing, waarin de i en j ogenschijnlijk tot dezelfde verzameling (*bij voorbeeld: handeldrijvende staten*) doch in wezen tot verschillende verzamelingen (*importerende, resp. exporterende staten*) behoren. De datamatrix is dan vierkant van vorm, maar rechthoekig van karakter (cel (i,j) verschilt van cel (j.i)) en wordt rechthoekig geanalyseerd onder gebruik van parameters $w_{i\psi}$ of $w_{j\psi}$.
N.B.: het 'klassieke' model van ontvouwen kent deze parameters w niet.
(h) *externe analyse*: Een bijzonder geval van ontvouwing, waarbij de $x_i$ apriori bepaald zijn en de $x_j$ in de analyse worden bepaald.