MULTIVARIATE ANALYSE METHODEN VOOR DISCRETE VARIABELEN. KIJKEN NAAR GEHEIME BESTANDEN?

KM 3(1981)
pag 82-84

R.D. Gill & B.F. Schriever[1]

Samenvatting

In /2/ analyseren Israëls e.a. een vierdimensionale kruistabel met verschillende exploratieve technieken, zonder gebruik te maken van kennis omtrent de betekenis van de variabelen. Zelfs het feit dat sommige variabelen ordinaal zijn wordt niet gebruikt. Ten eerste willen wij een korte opmerking maken over het nut van het werken met geheime bestanden in een exploratieve analyse. Daarna gaan we in op de technische vraag hoe de volgorde van de categorieën van ordinale variabelen in zo'n geheim bestand gereconstrueerd kan worden.

1. Geheime bestanden

Als doel van een exploratief onderzoek noemen Israëls e.a. het zoeken naar patronen in de gegevens, zodat men in een confirmatieve fase meer houvast heeft bij het specificeren van nuttige modellen. Zij beweren dat het, bij zo'n exploratieve analyse, een probleem is voor de onderzoeker om objectief te werk te gaan: hij wordt geleid door zaken die hij zelf interessant vindt, die hij eruit wil krijgen of die hem toevalligerwijs opvallen.

De onderzoeker heeft een aantal verwachtingen, gebaseerd op zijn achtergrond kennis, van de resultaten die de analyses kunnen opleveren. Sommige verwachtingen worden door de analyses bevestigd, maar, nog belangrijker, soms komen er onverwachte zaken naar voren. Liever gezegd: niet van te voren verwachte; want meestal zal de onderzoeker achteraf pas kunnen beargumenteren dat hij die en die verschijnselen toch had kunnen verwachten. Bijvoorbeeld, van te voren verwachten we misschien dat het inkomen met de leeftijd stijgt, maar uit de analyses blijkt dat bij de hoogste leeftijdsklassen een omkering plaats vindt (zie /2/). Achteraf kunnen we dit verklaren, maar het was nodig om ons hieraan te herinneren.

Subjectiviteit is, in een exploratieve fase althans, geen bezwaar. Als we iets interessants vinden, dan zullen we onszelf en anderen in een confirmatieve fase ervan willen overtuigen dat het reëel is. Als het bovendien de bedoeling is om buiten de bestaande gegevens te generaliseren, zal de bewijskracht het sterkst zijn, wanneer we dan met nieuwe gegevens werken.

Subjectiviteit is een groter probleem als men niet zelf bezig is een of ander verschijnsel te onderzoeken, maar alleen een neutrale en overzichtelijke beschrijving van de gegevens wil geven die andere onderzoekers of beleidsvoerders kunnen gebruiken. (Dit is niet alleen bij de analyse methodes, maar ook al bij het verzamelen van de gegevens een probleem, zie /1/.)

[1]Mathematisch Centrum, Amsterdam.

In §3.2 van /2/ is in het geheime bestand het ordinale karakter van enkele variabelen gereconstrueerd. In feite met een kleine en interessante afwijking: de leeftijdsklassen 3 en 4 zijn verwisseld. Bovendien wordt, impliciet, een ordinaal karakter gegeven aan de categorische variabele bedrijfstak. Immers §3.2 wekt de suggestie dat als men in de volgorde B2,B4,B6,B5,B3,B1 door de bedrijfstakken loopt het inkomen steeds hoger wordt.

Er is een aardige relatie aan te geven tussen de methode die hier gevolgd is en de technieken correspondentie analyse en Homals. Deze technieken bezitten een eigenschap waardoor het zoeken naar volgordes "gemechaniseerd" kan worden; iets wat bij grotere problemen nuttig kan zijn.

Bekijk allereerst een tweedimensionale tabel van relatieve frequenties of kansen p_{ij}, met marginalen p_{i+} en p_{+j} ($i=1,\ldots,N$; $j=1,\ldots,M$). De informele analyse van tabel 3.1 in §3.2 heeft, na een herordening van de categorieën, aangetoond dat de voorwaardelijke verdelingen van de kolomcategorieën, voorwaardelijk op de rijcategorieën, "stochastisch geordend zijn", d.w.z.:

$$(*) \qquad \sum_{j=1}^{j_0} \frac{p_{ij}}{p_{i+}} \text{ is dalend in } i \text{ voor iedere } j_0.$$

Intuïtief gezien, schuiven de verdelingen over de kolommen, voorwaardelijk op iedere rij, op naarmate we bij hogere rijen komen. Oftewel: naarmate de index van de rijcategorie hoger wordt gaat meer gewicht naar de kolomcategorieën met hogere indices. Nog intuïtiever, de kolomvariabele stijgt met de rijvariabele: er is een (eenzijdige) ordinale relatie tussen de twee variabelen. Niet noodzakelijkerwijs, maar wel vaak, geldt voor zo'n tabel ook het omgekeerde: naarmate de index van de kolomcategorie hoger wordt, krijgen de rijcategorieën met hogere indices meer gewicht; m.a.w.:

$$(**) \qquad \sum_{i=1}^{i_0} \frac{p_{ij}}{p_{+j}} \text{ is dalend in } j \text{ voor iedere } i_0.$$

Nu kunnen we het volgende bewijzen: als correspondentie analyse op een tabel, die aan (*) en (**) voldoet, wordt toegepast, zal de ordening van de schaalwaarden in de eerste dimensie overeenkomen met de ordening van de categorieën. Anders gezegd: als twee variabelen, bij bepaalde ordeningen van de categorieën, deze vorm van positieve samenhang vertonen, zal correspondentie analyse deze ordening terug vinden.

Analoge resultaten gelden ook voor correspondentie analyse toegepast op samengevoegde kruistabellen (bijv. (S+L+B)×I) en voor Homals, mits voor iedere kruistabel die in de analyse betrokken wordt de twee variabelen op de boven genoemde manier positief samenhangen. De volgordes van de categorieën van één variabele geïnduceerd door de verschillende kruistabellen moeten wel dezelfde zijn. In het algemeen zal de volgorde ook gevonden worden als er

kleine afwijkingen van dit patroon bestaan, of als dit patroon verstoord wordt door tweetallen variabelen die zeer zwak samenhangen.

Zo kunnen we de volgende aanvullende analyse op kruistabellen uitvoeren:

- pas Homals of correspondentie analyse toe;
- herschrijf de kruistabellen met de categorieën in de volgorde die door de schaalwaarden uit Homals resp. correspondentie analyse worden gegeven;
- ga na of de zo verkregen tabellen (bijna) aan (*) en (**) voldoen.

Correspondentie analyse op de tabel (S+L+B)×I en Homals op S,B,L,I (zie resp. fig. 8.1 en fig. 8.3 uit /2/) leveren beide dezelfde volgordes voor de categorieën van de vier variabelen. Bij contrôle blijken de kruistabellen I×S,I×B,I×L,S×B, na ordening van de categorieën, aan (*) en (**) te voldoen. De tabel B×L wijkt een klein beetje van dit patroon af, terwijl S×L er totaal niet aan voldoet.

Onze conclusie is dat Homals en correspondentie analyse handige hulpmiddelen zijn om na te gaan of er onverwachte ordinale relaties tussen de variabelen bestaan.

Literatuur

/1/ J. Irvine, I. Mills & J. Evans "Demystifying Social Statistics". Pluto Press, London (1979).

/2/ A.Z. Israëls, J.G. Bethlehem, J. van Driel, M.E. Jansen, J. Pannekoek, S.J.M. de Ree & D. Sikkel "Multivariate Analyse Methoden voor Discrete Variabelen". KM 2 (1981) 87-149