Een model voor de analyse van latente tijdbudgetten

P. Verboon[1]

P. G. M. van der Heijden[2]

J. de Leeuw[3]

Samenvatting

Tijdbudgetten vatten samen hoe objecten hun tijd besteden. Een belangrijk kenmerk van tijdbudgetten is het feit dat de rijen van de datamatrix optellen tot dezelfde constante. In dit artikel wordt een model beschreven voor de analyse van tijdbudgetten, waarbij rekening gehouden wordt met deze speciale eigenschap. Verder worden er enkele uitbreidingen van het model besproken en wordt ter illustratie een voorbeeld gepresenteerd. Tenslotte wordt het model vergeleken met een aantal verwante technieken, zoals latente klassenanalyse en correspondentieanalyse.

[1] Vakgroep Datatheorie. Rijksuniversiteit Leiden. Middelstegracht 4. 2312 TW Leiden. tel. 071-273827.

[2]Vakgroep Methoden en Technieken. Rijksuniversiteit Leiden.

[3]Depts. of Mathematics and Psychology. University of California at Los Angeles.

Inleiding

Dit artikel behandelt de analyse van tijdbudgettten. Tijdbudgetten worden verzameld indien men geinteresseerd is in de wijze waarop (groepen) mensen of dieren (of, algemener, objecten) hun tijd besteden. Tijdbudgetten hebben betrekking op de wijze waarop objecten hun tijd verdelen over verschillende, elkaar uitsluitende aktiviteiten. We kunnen dit type gegevens samenvatten in een kruistabel. In deze kruistabel worden de objecten gerepresenteerd in de rijen. De aktiviteiten van deze objecten vinden we in de kolommen van de tabel. In een cel staat de proportie van de totale tijd die een bepaald object aan een aktiviteit besteed heeft. Elke rij van deze kruistabel telt dus op tot 1. In tabel 1 staat een voorbeeld van een kruistabel met tijdbudget data. In dit voorbeeld staat van vijf objecten de proportie tijd vermeld die besteed is aan de activiteiten **A, B, C** en **D**.

Tabel 1. Voorbeeld tijdbudgetdata

	A	B	C	D
1	.38	.14	.27	.21
2	.45	.07	.19	.29
3	.23	.23	.47	.08
4	.40	.12	.23	.25
5	.47	.06	.11	.36

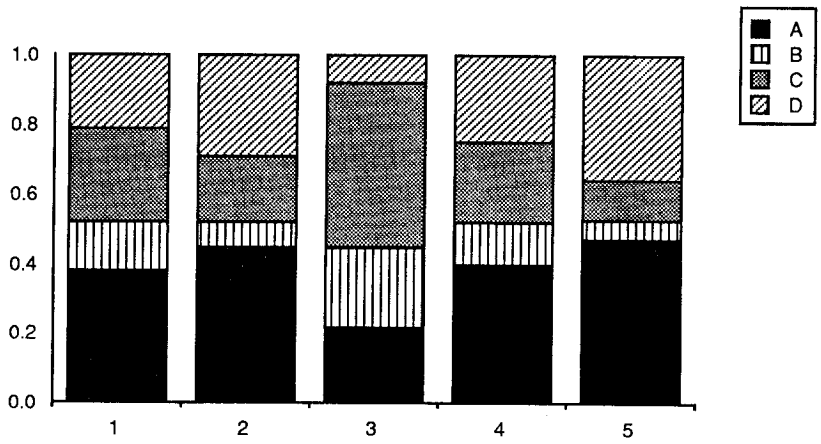
Een matrix met tijdbudgetdata kan op verschillende manieren verkregen worden. In de eerste plaats kunnen tijdbudgetten geschat worden via de zogenaamde 'random spot check methode' (zie Gross, 1984). Bij deze methode wordt voor ieder object op willekeurige momenten de aktiviteit genoteerd waar dit object op dat moment mee bezig is. Het aantal maal dat een bepaalde aktiviteit bij een object geobserveerd wordt, is een indicatie voor de totale tijd die dit object aan deze aktiviteit besteed. Het idee is dat wanneer bijvoorbeeld een persoon van de 100 keer dat zijn gedrag geobserveerd wordt, daarvan 20 keer aan het werk is, hij 20% van z'n tijd doorbrengt met werken.

Bij een dataverzamelingsmethode die hiermee verwant is, wordt gebruik gemaakt van 'piepertjes' (zie bijvoorbeeld Robinson, 1985). Hierbij krijgt een persoon een pieper mee die op willekeurige momenten een signaal geeft. Elke keer dat het signaal gegeven wordt moet deze persoon opschrijven waar hij op dat moment mee bezig is. Deze methode levert hetzelfde type gegevens als de 'random spot check methode'.

Een andere methode om tijdbudgetdata te verkrijgen is door ze af te leiden uit zogenaamde 'event history' data. Bij 'event history' data is voor elk object niet alleen de duur van verschillende activiteiten precies bekend, maar ook de volgorde waarin deze plaatsvinden. Dit type data is bijvoorbeeld te verkrijgen door personen een dagboek bij te laten houden. Ook deze data kunnen in een matrix als in tabel 1 worden samengevat. In deze samenvatting speelt echter de volgorde van de verschillende gedragingen geen rol meer.

De analyse van tijdbudgetdata heeft tot doel de verbanden tussen de objecten enerzijds en de activiteiten anderzijds te onderzoeken. De volgende vraagstellingen zouden bijvoorbeeld van belang kunnen zijn: zijn er personen met een, in vergelijking tot de anderen, duidelijk afwijkend activiteitenpatroon? Zijn er personen met vrijwel identieke activiteitenpatronen? Zijn er activiteiten die alleen bij bepaalde personen veel voorkomen?

Van der Heijden (1987) en De Leeuw en Van der Heijden (1987) geven een beknopte opsomming van analysemethoden uit verschillende disciplines, zoals ethologie, sociologie en antropologie, om dit soort vragen aan te pakken. In de eerste plaats kunnen de data weergegeven worden in een histogram. Zie figuur 1. Dit geeft voornamelijk een overzichtelijke presentatie van de goed 'gevulde' activiteiten, maar geeft verder weinig inzicht in de mogelijk aanwezige verbanden in de kruistabel. Een ander nadeel is dat, wanneer het aantal activiteiten of het aantal objecten groot wordt, de grafiek al snel 'onleesbaar' is.



Figuur 1. Histogram van voorbeeld-data.

In veel onderzoeken komt men niet veel verder dan het maken van tabellen, of dit soort grafische representaties, in allerlei variaties (zie bijvoorbeeld Szalai, 1972; Parkes & Thrift, 1980; Staikov, 1982; Harvey et al., 1984). Hoe beperkt ook, een voordeel van deze aanpak is dat de resultaten op eenvoudige wijze zijn over te brengen op een breder (dan een louter

wetenschappelijk) publiek. Ook worden er wel standaard multivariate technieken gebruikt, waarbij men dan correlaties tussen kolommen met frekwenties berekent (zie Elliott, 1984). Het nadeel hiervan is dat een correlatie geen geschikte maat is om de gelijkenis tussen kolommen met frekwenties te berekenen, en er bovendien geen rekening gehouden wordt met de speciale eigenschap van tijdbudgetdata dat de rijen optellen tot een constante.

Een andere mogelijkheid is om specifieke methoden voor de analyse van kruistabellen te gebruiken, zoals correspondentieanalyse (Gifi, 1981; Greenacre, 1984). Hierbij worden zowel objecten als aktiviteiten als punten in een ruimte afgebeeld, waarbij de afstanden tussen punten de samenhang in de matrix weergeven. In het kader van tijdbudgetdata laat correspondentie analyse een groeiende belangstelling zien (zie bijvoorbeeld Saporta, 1981; De Leeuw et al., 1985, Van der Heijden, 1987). Correspondentieanalyse lijkt een geschikte methode voor de analyse van dit type gegevens met rijen die optellen tot 1, omdat de afstandsmaat die correspondentieanalyse gebruikt, de chi-kwadraatafstand, weergeeft hoeveel rijen (en kolommen) met conditionele proporties van elkaar afwijken. Een nadeel is echter dat bij correspondentieanalyse de mogelijkheid van toetsing ontbreekt (tenzij men recentelijk voorgestelde modellen van Goodman, 1985, 1986, gebruikt) en dat de resultaten ervan wat moeilijker hanteerbaar zijn voor niet wetenschappelijk publiek (cf. De Leeuw en Van der Heijden, 1987). Loglineaire analyse technieken hebben als nadeel dat, indien er een relatie tussen de objecten en de categorieen aanwezig is, het aantal te interpreteren parameters snel erg groot wordt, en bovendien geen gebruik wordt gemaakt van het feit dat de proporties per rij optellen tot 1.

In dit artikel bespreken wij een ander model om tijdbudgetdata te analyseren (zie ook De Leeuw & Van der Heijden, 1987). Dit model biedt de mogelijkheid om op een eenvoudige manier eventuele ingewikkelde verbanden weer te geven die de onderzoeker meer inzicht kunnen geven in de structuur van zijn data. Het model maakt gebruik van de speciale eigenschappen van tijdbudgetdata, en indien de tijdbudgetdata verzameld zijn met de 'random spot check' methode, of met behulp van bovengenoemde piepertjes, is het mogelijk om de houdbaarheid van het model te toetsen.

Het latente tijdsbudgetten model

We gaan uit van een s x m matrix waarbij s het aantal objecten voorstelt en m het aantal aktiviteiten. We nemen aan dat de gegevens verzameld zijn met de 'random spot check' methode of met behulp van piepertjes. Het aantal keer dat een object i bezig was met aktiviteit j wordt weergegeven met n_{ij}, waarbij i=1...s en j=1...m. Het totaal aantal metingen per object wordt bepaald door het design en wordt voorgesteld door n_{i+} (een index wordt vervangen door '+' indien opgeteld is over de corresponderende weg van de matrix). De proportie $p_{ij} = n_{ij}/n_{i+}$ kan nu beschouwd worden als schatter van π_{ij}: de theoretische proportie tijd besteed door object i

aan aktiviteit j in een voor object i typerende periode. Indien object i staat voor een groep, dan kan πij geinterpreteerd worden als de theoretische proportie tijd voor een typisch groepslid.

Indien we aannemen dat de metingen onafhankelijk van elkaar zijn, dan volgen de n_{ij} in rij i een multinomiale verdeling met $E(n_{ij}) = n_{i+} \pi_{ij}$. Deze aanname is slechts realisties indien de metingen met voldoende tussenruimte zijn gedaan. Indien bovendien de objecten elkaars activiteiten niet beinvloeden, dan mogen we aannemen dat de n_{ij} in de gehele matrix een product-multinomiale verdeling volgen. We zullen in het vervolg aannemen dat deze aannamen niet al te zeer geschonden worden. We hebben nu de mogelijkheid om te toetsen of de verdelingen van de p_{ij} voor verschillende objecten i verschillen. Daarvoor berekenen we verwachte proporties op grond van het onafhankelijkheidsmodel. In tabel 2 staan in de linker matrix de verwachte proporties voor dit model weergegeven. De $\chi 2$ is 70,9 met 12 vrijheidsgraden, indien we aannemen dat de proporties in tabel 1 gebaseerd zijn op 100 waarnemingen per rij. Dit is significant hetgeen betekent dat het model verworpen moet worden. De objecten en aktiviteiten zijn dus niet onafhankelijk.

Tabel 2. Verwachte waarden bij p=1 (onafhankelijkheid) en p=2

	A	B	C	D
1	.39	.12	.25	.24
2	.39	.12	.25	.24
3	.39	.12	.25	.24
4	.39	.12	.25	.24
5	.39	.12	.25	.24

	A	B	C	D
1	.37	.13	.28	.22
2	.44	.09	.18	.30
3	.23	.23	.47	.08
4	.40	.11	.23	.26
5	.49	.05	.12	.35

Onafhankelijkheid van rijen en kolommen zal in interessante toepassingen vrijwel altijd worden verworpen. We stellen voor de aard van de afhankelijkheid nader onderzoeken met een latente tijdbudgetten model. Dit model wordt als volgt geschreven:

$$\pi_{ij} = \Sigma_k \alpha_{ik} \beta_{jk}. \qquad (1)$$

Hierbij indexeert k de latente budgetten, waarbij k=1...p, waarbij p door de onderzoeker wordt bepaald. De volgende restricties gelden bij dit model: $\Sigma_j \pi_{ij} = 1$, d.w.z. de rijen van de datamatrix tellen op tot 1, $0 \geq \alpha_{ik} \geq 1$, $0 \leq \beta_{jk} \leq 1$, d.w.z. de parameterschattingen liggen tussen 0 en 1, en tenslotte: $\Sigma_k \alpha_{ik} = 1$ en $\Sigma_j \beta_{jk} = 1$. Het idee achter het latente tijdbudgetten model is dat er p typische latente tijdbudgetten β_{jk} zijn die het gedrag van alle objecten bepalen.

De activiteiten in elk latente tijdbudget tellen op tot 1: vandaar $\Sigma_j \beta_{jk} = 1$. Ieder latent tijdbudget β_{jk} verklaart het tijdbestedingspatroon van een object in meer of mindere mate, hetgeen wordt aangegeven door de grootte van een gewicht α_{ik} dat een object i op een latent budget k heeft. Het tijdbestedingspatroon van elk object i is voor 100% samengesteld uit de latente tijdbudgetten, vandaar $\Sigma_k \alpha_{ik} = 1$. Merk op dat voor p=1 dit model het eerder gebruikte onafhankelijke model is. Het aantal bij dit model passende vrijheidsgraden is op de gebruikelijke manier te berekenen, namelijk als het aantal onafhankelijke cellen minus het aantal onafhankelijk te schatten parameters. Het aantal onafhankelijke cellen is s(m-1). Bij p latente budgetten zijn er p(m-1) onafhankelijke beta-parameters en (p-1)s onafhankelijke alfa-parameters. Dit betekent dat het aantal vrijheidsgraden gelijk is aan:

$$\# df = [s(m-1)] - [p(m-1) + (p-1)s]. \qquad (2)$$

Wanneer we het model met p=2 latente budgetten uitrekenen voor het voorbeeld, dan vinden we de schattingen voor de modelparameters die staan vermeld in tabel 3 en 4. De verwachte proporties staan in de rechter matrix van tabel 2. De $\chi 2$ is 0.7 met 4 vrijheidsgraden, hetgeen niet significant is. Het model met twee latente budgetten geeft een goede beschrijving van de data .

We kunnen de meest opmerkelijke resultaten als volgt omschrijven. Het eerste latente budget wijkt het meest af van het budget onder het onafhankelijkheidsmodel door een hogere waarde voor **C** (.47 vergeleken met .25) en een kleinere voor **D** (.07 vergeleken met .24). Het tweede budget heeft juist een hoge waarde voor **D** en een lage voor **C**. De waarden van **A** en **B** zijn voor beide budgetten in mindere mate verschillend aan de waarden bij onafhankelijkheid. Aan de gewichten zien we dat de tijdbesteding van het derde object vrijwel volledig bepaald wordt door het eerste budget, deze scoort dus meer dan gemiddeld op **C** en minder op **D**. Voor object vijf geldt precies het tegenovergestelde. De andere objecten liggen meer tussen deze extreme objecten in. Een interpretatie is toelaatbaar dat er twee typische manieren van tijdbesteding zijn, waar elk object op een of andere wijze tussen in zit.

Tabel 3. α_{jk} schattingen bij twee latente budgetten.

k	1	2
1	.57	.43
2	.35	.65
3	.99	.01
4	.47	.53
5	.20	.80

Tabel 4. β_{jk} schattingen bij twee latente budgetten.

k	1	2
A	.23	.55
B	.23	.01
C	.47	.02
D	.07	.42

Omdat het model goed past, hebben we geen reden aan te nemen dat een derde latent tijdbudget nodig is om tot een adekwate beschrijving van de tijdbesteding te komen. Het uiteindelijke doel van het latente tijdsbudgetten model zal duidelijk zijn. Het probeert een spaarzame beschrijving te geven van de data met behulp van typerende tijdsbudgetten die zijn terug te vinden in de β_{jk}. De α_{ik} laten vervolgens zien hoe de objecten op deze typerende budgetten laden, of anders gezegd, hoeveel er verklaard wordt van het tijdbestedingspatroon van een object.

Schattingsmethode

Omdat we aannemen dat de data een steekproef vormen uit een product-multinomiale verdeling, kunnen we de twee-weg matrix met tijdbestedingsobservaties nij als het ware opblazen tot een drie-weg matrix, met in de derde weg voor elk latente budget een plakje, en vervolgens 'maximum likelihood schatters' verkrijgen met behulp van het EM-algoritme (Dempster, Laird, and Rubin, 1977). Om de α_{ik} en β_{jk} te berekenen, moeten we stapsgewijs te werk gaan. In stap 1, waarin de likelihood wordt gemaximaliseert, definieren we dan

stap 1 $$n_{ijk} = [n_{ij}/\pi_{ij}\,]\,\alpha_{ik}\,\beta_{jk}, \tag{3}$$

waarbij n_{ij} de originele datamatrix is en α_{ik} en β_{jk} de op dat moment beste schatters tijdens het itereren van het algoritme, die de waarden π_{ij} opleveren.

Nieuwe waarden van schattingen α_{ik} en β_{jk} worden vervolgens verkregen in stap 2:

stap 2 $$\alpha_{ik} = n_{i+k} / n_{i++} \tag{4}$$

$$\beta_{jk} = n_{+jk} / n_{++k} \tag{5}$$

Hieruit volgen onmiddelijk de restricties dat $\alpha_{i+} = \beta_{+k} = 1$. De in stap 2 geschatte parameters worden vervolgens weer in stap 1 ingevoerd. Dit proces herhaalt zich totdat convergentie is bereikt. Dit algoritme convergeert in ieder geval naar een locaal maximum. Het convergentiebewijs van het algoritme voor het maximaliseren van de likelihood wordt gegeven in De Leeuw and Van der Heijden (1987).

Voorbeeld

De data uit dit voorbeeld zijn ontleend aan Gross et al. (1985) en zijn eerder geanalyseerd in De Leeuw en Van der Heijden (1987) en Van der Heijden (1987). De data zijn verzameld met de random spot check methode. De objecten zijn mannen, vrouwen en kinderen van vier stammen Indianen uit het Amazone gebied. We zullen dit respectievelijk de variabelen sexe en stam noemen. Volledig gekruist leveren deze variabelen twaalf groepen op. Er waren zes elkaar uitsluitende activiteiten gemeten, te weten : niets doen, slapen, verzorgen, "geestelijke bezigheden", huishoudelijke aktiviteiten en jagen. Voor een uitgebreide beschrijving van deze stammen en gedragingen verwijzen we naar Gross et al. (1985) en Werner et al. (1979).

Metingen zijn verricht in zeven perioden van twee uur, beginnend om 6 uur 's ochtends en eindigend 8 uur 's avonds. Voor deze analyse is opgeteld over de verschillende perioden, dit resulteerde in een 12 x 6 datamatrix. Een cel van deze matrix geeft een schatting van de totale hoeveelheid tijd die door een groep aan een bepaalde activiteit is besteed gedurende de totale 14 uur. Uit 12 in Gross et al. (1985) gepresenteerde figuren (voor elke groep een figuur), vergelijkbaar met figuur 1, zijn proporties p_{ij} afgeleid. Deze proporties zijn bij de analyses gebruikt (het bleek niet mogelijk de data zelf te verkrijgen). In eerste instantie zijn er drie analyses gedaan met respectievelijk een, twee en drie latente budgetten in het model. In de tabellen 5 en 7 staan de schattingen van de parameters weergegeven. In tabel 6 staan voor de modellen de likelihood ratio (fit) met het bijbehorende aantal vrijheidsgraden vermeld.

De ladingen van de groepen op de latente budgetten staan vermeld in tabel 5. Hieraan is te zien dat bij een analyse met twee latente budgetten in het model (p=2) de kinderen hoog laden op het tweede budget, terwijl vooral de vrouwen een hoge lading hebben op het eerste budget. Het tweede budget beschrijft dus voornamelijk het tijdbestedingspatroon van de kinderen terwijl het patroon van de vrouwen vooral door het eerste 'verklaard' wordt. De positie van mannen daarentegen is afhankelijk van de stam. Zo besteden Mekranoti mannen hun tijd op dezelfde wijze als Mekranoti vrouwen, terwijl de Xavente man meer een bestedingspatroon te zien geeft dat op dat van kinderen lijkt. Voor de overige stammen laden de mannen ongeveer gelijk op beide budgetten.

In tabel 7 zien we in de eerste kolom het latente budget voor p=1. Uit (1) is te zien dat voor p=1 het onafhankelijkheidsmodel gefit wordt. Met behulp van dit onafhankelijke model, dat slechts de marginale kolomproporties fit, kunnen we voor het model met p=2 latente budgetten concluderen dat kinderen meer "niets doen" dan gemiddeld (.781) en dat vrouwen meer dan gemiddeld hun tijd besteden aan "geestelijke bezigheden" (.338). Met 'gemiddeld' wordt hier dus de verwachte waarde bij onafhankelijkheid bedoeld.

Een analyse met p=3 levert, ruwweg, voor iedere categorie van sexe een eigen budget op. Het blijkt dat mannen meer "geestelijk bezig zijn" en meer jagen dan gemiddeld. Verder is het duidelijk dat de variabele sexe veel belangrijker is dan stam als we de tijdbestedingsverschillen

willen verklaren. De gewichtsverschillen binnen stammen zijn namelijk veel groter dan die tussen stammen.

Tenslotte is te zien uit tabel 6 dat de fit aanzienlijk beter wordt wanneer we meer latente budgetten gebruiken, terwijl natuurlijk het aantal vrijheidsgraden afneemt. Aangezien we bij dit voorbeeld niet beschikken over de ruwe data kunnen we de chi-kwadraat statistiek niet gebruiken om de significantie van het model te toetsen.

Table 5. α_{ik} parameter schattingen voor verschillende waarden van p.

		p=2		p=3		
stam	**sexe**	k=1	k=2	k=1	k=2	k=3
Mekranoti	man	.763	.237	.107	.832	.061
	vro	.725	.275	.253	.109	.638
	kin	.054	.946	.876	.084	.040
Kanela	man	.558	.442	.346	.448	.206
	vro	.782	.218	.220	.019	.761
	kin	.038	.962	.929	.009	.063
Bororo	man	.458	.542	.363	.625	.012
	vro	.828	.172	.146	.207	.647
	kin	.108	.892	.783	.200	.017
Xavente	man	.331	.669	.520	.457	.024
	vro	.982	.018	.002	.216	.783
	kin	.134	.866	.799	.110	.091

Tabel 6. Fit en vrijheidgraden voor verschillende p.

p	fit	df
1	2.519	55
2	.963	38
3	.370	21

Table 7. β_{jk} parameter schattingen voor verschillende waarden van p.

	p=1	p=2		p=3		
gedrag	k=1	k=1	k=2	k=1	k=2	k=3
niets	.594	.391	.781	.817	.437	.391
slapen	.060	.031	.087	.095	.032	.034
zorgen	.032	.068	.000	.000	.000	.116
geestelijk	.174	.338	.023	.005	.271	.348
huis	.093	.105	.081	.080	.096	.110
jagen	.047	.067	.028	.003	.163	.000

Modelleren van structuur in de objecten

Tot nu hebben we de objecten in de matrix als gelijkwaardig beschouwd. Echter, wanneer de objecten te karakteriseren zijn door gekruiste categorieen van twee (of meer) variabelen, zoals in ons voorbeeld de variabelen stam en sexe, kan het van belang zijn om het effect van de afzonderlijke variabelen te onderzoeken. Zo kan men zich afvragen of verschillen in tijdsbesteding alleen veroorzaakt worden door het geslacht of door de stam waartoe een persoon behoort. Om dit soort vragen te beantwoorden, moeten we de tijdbudgetten opvatten als een drie-weg matrix, en kan het oude model (1) geschreven worden als :

$$\pi_{isj} = \Sigma_k \, \alpha_{isk} \, \beta_{jk}, \tag{6}$$

waarbij s en i de rijvariabelen indexeren. Om de α_{isk} en β_{jk} te berekenen definieren we nu stap 1:

stap 1 $$n_{isjk} = [n_{isj}/\pi_{isj}] \, \alpha_{isk} \, \beta_{jk}, \tag{7}$$

waarbij n_{isj} de geobserveerde datamatrix is en α_{isk} en β_{jk} de op dat moment beste schatters tijdens het itereren van het algoritme. Nieuwe waarden van α_{isk} en β_{jk} worden in stap 2 berekend volgens (8) en (9).

stap 2 $$\alpha_{isk} = n_{is+k} / n_{is++}, \tag{8}$$

$$\beta_{jk} = n_{++jk} / n_{+++k}. \tag{9}$$

Hieruit volgt wederom dat $\alpha_{is+} = \beta_{+k} = 1$. De gevonden waarden worden weer in stap 1 ingevoerd en het proces herhaald zich totdat convergentie is bereikt.

Door het oorspronkelijke model (1) als volgt te herformuleren, is het mogelijk geworden om het effect van elke variabele afzonderlijk te onderzoeken door restricties aan de rijparameters van het model op te leggen. Een mogelijk interessante restrictie is te stellen dat een variabele geen invloed heeft op de rijparameters. Dit kan door te sommeren over deze variabele. Het effect van die variabele wordt zo uit het model gehaald. Passen we dit toe op het voorbeeld van de Amazone indianen, dan kunnen we het effect van de variabelen sexe en stam afzonderlijk gaan bekijken door het effect van de andere variabele te verwijderen. De variabelen sexe en stam worden respectievelijk door i en s geindexeerd. We analyzeren de data nogmaals, nu met het effect van stam verwijderd. De α_{isk} wordt dan berekend met :

$$\textit{stap 2} \qquad \alpha_{isk} = n_{i++k} / n_{i+++}. \tag{10}$$

Wanneer het effect van sexe verwijderd wordt dan ziet de berekening van α_{isk} eruit als :

$$\textit{stap 2} \qquad \alpha_{isk} = n_{+s+k} / n_{+s++}. \tag{11}$$

De resultaten van beide analyses met twee latente budgetten zijn in de onderstaande tabellen samengevat. De notatie [sexe] geeft aan dat alleen de variabele sexe van belang is. In dit geval hebben bijvoorbeeld alle rijen voor de mannen dezelfde gewichten op de latente budgettten. Het model waaruit het effect van de variabele stam is verwijderd, blijkt bijna even goed te passen als het ongerestricteerde model (zie tabel 8). Het model met de variabele sexe verwijderd past een stuk slechter. De parameterschattingen staan vermeld in tabel 9 en 10.

Tabel 8. Fit en vrijheidgraden voor verschillende modellen.

	p=2		p=3	
model	fit	df	fit	df
[sexe,stam]	0.96	38	0.37	21
[sexe]	1.10	47	0.59	39
[stam]	2.43	46	2.33	37

Table 9. α_{isk} parameter schattingen voor [sexe] en [stam].

		p=2		p=3		
model		k=1	k=2	k=1	k=2	k=3
[stam]	**Mekranoti**	.23	.76	.31	.51	.18
	Kanela	.63	.37	.12	.30	.57
	Bororo	.58	.42	.46	.11	.42
	Xavente	.65	.34	.24	.22	.54
[sexe]	**man**	.46	.54	.06	.71	.23
	vrouw	.15	.85	.73	.12	.15
	kind	.94	.06	.04	.11	.85

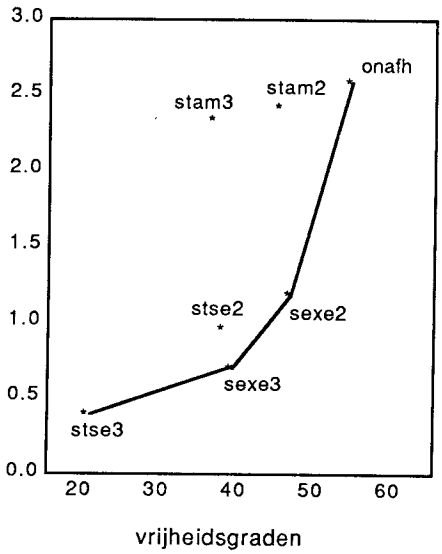
Table 10. β_{jk} parameter schattingen voor [sexe] en [stam].

	p=1	[stam]		[sexe]	
gedrag	k=1	k=1	k=2	k=1	k=2
niets	.59	.68	.50	.78	.40
slapen	.06	.04	.08	.09	.03
zorgen	.03	.04	.02	.00	.07
geestelijk	.17	.18	.17	.03	.32
huis	.09	.04	.15	.07	.11
jagen	.05	.02	.07	.03	.07

In tabel 10 staan in de eerste kolom nogmaals de verwachte waarden vermeld onder het onafhankelijkheidsmodel. Wanneer we kijken naar het model [stam], d.w.z. het model waarin de rol van sexe is onderdrukt, dan zien we bij de twee latente budget dat alleen de categorieen 'niets doen' en 'huiselijke aktiviteiten' veel afwijken van de waarden bij onafhankelijkheid. Tabel 9 laat vervolgens zien dat de Mekranoti hoog laden op het tweede budget, terwijl de andere stammen hoog laden op het eerste. De Mekranoti stam wordt in deze analyse dus onderscheiden van de andere stammen. Dit verschil wordt veroorzaakt doordat er bij de

Mekranoti meer huiselijke aktiviteiten worden verricht, terwijl er minder 'niets' wordt gedaan. Bij [sexe] zien we weer het onderscheid vrouwen tegenover kinderen, waarbij in het budget van de kinderen 'niets doen' een hoge waarde heeft. Het budget van de vrouwen wordt gekenmerkt door een hoge waarde voor 'geestelijke activiteiten'. Het gedragspatroon van mannen wordt door beide budgetten in ongeveer gelijke mate verklaard

Tenslotte is er nog een analyse gedaan met drie latente budgetten. Hiervan zijn de fit en de gewichten vermeld. Om nu te zien welke analyses de beste resultaten opleveren, kunnen we de fit waarden uitzetten tegen de vrijheidsgraden die erbij horen (zie ook Verbeek, 1984). Zie figuur 2.



Het cijfer achter de naam geeft het aantal onderliggende latente budgetten aan. De naam 'stse' staat voor het model [sexe,stam]. Rechtsboven in de figuur vinden we het onafhankelijksmodel. Dit model heeft de hoogste likelihood-ratio en het hoogste aantal vrijheidsgraden.

Figuur 2. Likelihood-ratio uitgezet tegen aantal vrijheidsgraden voor verschillende modellen.

Er dient bij het kiezen voor een uiteindelijk model een afweging gemaakt te worden tussen goede fit en spaarzaamheid. Figuur 2 kan hierbij behulpzaam zijn. Een algemene regel is dat, bij gelijke fit, er gekozen dient te worden voor het model dat het minste aantal parameters heeft, d.w.z. het model met de meeste vrijheidsgraden. Voor modellen die op een horizontale lijn liggen, kiezen we dus voor het meest rechtse model in figuur 2. Een tweede regel is dat voor modellen die even spaarzaam zijn, we kiezen voor het model met de beste fit. Voor modellen die op een verticale lijn liggen, kiezen we dus het onderste model. In figuur 2 kunnen we zien dat de modellen die boven en links van de getrokken lijn liggen relatief slecht fittende modellen zijn.

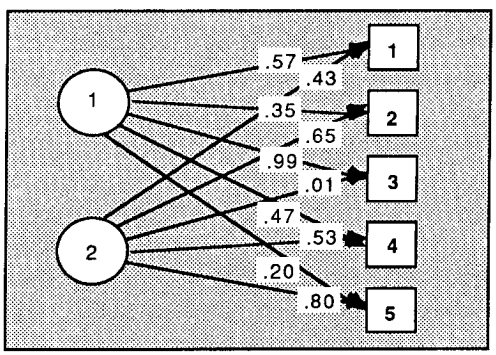
Er is voor deze modellen namelijk altijd een beter model te vinden. We zien nu dat [sexe,stam] en [sexe] met drie latente budgetten een goed resultaat opleveren. Indien we de data willen beschrijven met twee latente budgetten dat houden we alleen [sexe] over als geschikt model.
We hebben hier alleen modellen besproken waarbij effecten van bepaalde variabelen geheel onderdrukt werden. Het is ook mogelijk een model te postuleren waarin alleen 'hoofdeffecten' voor sexe en stam zitten, maar interactie niet is toegestaan. Momenteel wordt door ons gewerkt aan het aanpassen van dit soort modellen. Daarnaast is het ook mogelijk de latente budgetten op soortgelijke wijze te restricteren, bijvoorbeeld als de activiteiten multivariaat zijn, of als we tijdbudgetten hebben voor verschillende perioden. Ook hieraan wordt momenteel gewerkt.

Relaties met andere technieken

1. *Latente klassen analyse*

Het latente klassen model (voor een eenvoudige introductie, zie McCutcheon, 1987; zie ook Hagenaars, 1985) kan ook geschreven worden als (1). Er gelden echter andere restricties, namelijk: $\Sigma_i \alpha_{ik} = 1$ en $\Sigma_j \beta_{jk} = 1$. Verder geldt niet de restrictie dat $\Sigma_j \pi_{ij} = 1$. Door de verschillende restricties op de parameters van beide modellen kan er met beide technieken een ander type vragen beantwoord kan worden. De benadering van de geobserveerde waarden door de verwachte waarden zal bij een gelijk aantal latente factoren in de meeste gevallen niet veel verschillen.

Bij latente klassen analyse wordt er vanuit gegaan dat de geobserveerde variabelen veroorzaakt worden door een of meer niet geobserveerde of latente variabelen. Dit is hetzelfde uitgangspunt als bij factoranalyse en in die zin is het mogelijk factoranalyse te beschouwen als een vorm van latente klassen analyse. Het verschil is natuurlijk dat factoranalyse wordt uitgevoerd met kwantitatieve variabelen en latente klassen analyse met discrete of categorische variabelen.



Figuur 3. Model met twee latente budgetten.

De vergelijking van het latente tijdsbudgetten model en factor analyse kunnen we illustreren aan de hand van het in de inleiding besproken voorbeeld. Figuur 3 illustreert het eenvoudigste model dat ten grondslag ligt aan beide technieken.

De gewichten of, in factor analyse terminologie, de factorladingen, zijn in de figuur aangegeven. De cirkels stellen de twee latente budgetten voor, die gegeven zijn in tabel 4. De vierkantjes zijn de geobserveerde tijdsbestedingen van de vijf objecten.

Een belangrijk verschil met factor analyse is ook - naast het feit dat de variabelen discreet zijn - het karakter van de onderliggende factoren. Bij factor analyse worden de scores op de variabelen verklaard door de latente factoren, terwijl bij het tijdbudgetten model het gedragspatroon van de personen wordt verklaard door onderliggende factoren of latente budgetten. We gaan met het latente tijdbudgettten model dus op zoek naar typerende objecten (bijvoorbeeld, de vrolijke fliere fluiter, of het zorgzame type, waar wij allen iets van hebben) en niet naar typerende variabelen.

2. *Correspondentie analyse*

Wanneer we correspondentie analyse vergelijken met het latente tijdbudgetten model dan blijken beide technieken grote overeenkomst te vertonen. Deze overeenkomst doet zich voor als we de benadering van het model met k latente budgetten vergelijken met de benadering van een (k-1) dimensionele correspondentieanalyse oplossing (zie ook Gilula, 1984, die deze gelijkenis constateerde voor correspondentieanalyse en latente klassen analyse).

Dit is eenvoudig in te zien indien we het model dat bij correspondentie analyse gefit wordt opschrijven en vergelijken met (1).

$$\pi_{ij} = \pi_{i+}\,\pi_{+j}\,(1+\Sigma_s\, r_{is}\, c_{js}\, \lambda_s). \qquad (12)$$

Hierbij indexeert s (s=1...t) het aantal dimensies. De overeenkomst met het latente budgetten model is direct te zien wanneer we (1) en (12) vergelijken waarbij p=1 en t=0 gekozen wordt. Beide technieken fitten dan het onafhankelijkheidsmodel. In het algemeen geldt, wanneer we k=q en p=q+1 kiezen, dat π_{ij} benaderd wordt door de som van q+1 producten van rij- en kolomtermen.

Hoewel beide technieken veelal vergelijkbare resultaten op zullen leveren, is dit niet noodzakelijk het geval. In de eerste plaats is dit omdat voor de α_{ik} en β_{jk} andere restricties gelden dan voor de r_{is} en c_{js}, in de tweede plaats vanwege het feit dat beide technieken verschillende functies minimaliseren (zie ook Van der Heijden, 1987, De Leeuw en Van der Heijden, 1987).

Conclusies

We hebben hier een model gepresenteerd voor de analyse van tijdbudgetdata, dat rekening houdt met de speciale eigenschappen van deze data, en het mogelijk maakt uitspraken over de populatie te doen op basis van steekproefgegevens. Naar onze opvatting zijn de resultaten die dit model oplevert gemakkelijk uit te leggen aan leken. Dit lijkt ons een groot voordeel.

Een belangrijke aanname van het model is dat de gegevens getrokken zijn uit een product-multinomiale verdeling. Indien we te maken hebben met event history data, dan is deze aanname duidelijk geschonden. Ook in geval van de random spot check methode of wanneer piepertjes worden gebruikt, zal deze aanname geregeld geschonden zijn. De verkregen chi-kwadraat statistieken kunnen dan niet vergeleken worden met de waarden van de theoretische chi-kwadraat verdeling. Het model is in dit geval slechts bruikbaar voor de exploratie van de gegevens, en niet voor confirmatie.

Referenties

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society**, Series B 39,1-38.

Elliot, D.H. (1984). Statistical Analysis of time budget data. In: **Time budget research**. (Harvey et al., eds.).Frankfurt: Campus Verlag.

Gifi, A. (1981). **Non-linear multivariate analysis.** Leiden: Vakgroep Datatheorie.

Gilula, Z. (1984). On some similarities between canonical models and latent class models for two-way contingency tables. **Biometrika**, 71, 523-529.

Goodman, L.A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. **Annals of Statistics**, 13, 10-69.

Goodman, L.A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. **International Statistical Review**, 54, 243-309.

Greenacre, M.J. (1984). **Theory and applications of correspondence analysis**. New York: Academic Press.

Gross, D.R. (1984). Time allocation: A tool for the study of cultural behavior. **Annual Review of Antropology**, 13, 519-558.

Gross, D.R., Rubin, J.& Flowers, N.M. (1985). **All in a day's time: random spot check data can describe the texture of a day's activities**. Presented at the Annual Meeting of the American Association of Advancement of Science, Los Angeles, May 29,1985.

Hagenaars, J.A.P. (1985). **Loglineaire analyse van herhaalde surveys; panel-, trend- en cohortonderzoek.** Dissertatie, Katholieke Universiteit van Brabant, Tilburg.

Harvey, A.S., Szalai, A., Elliot, D.H., Stone, P.J. & Clarke, S.M. (1984). **Time budget research**. Frankfurt: Campus Verlag.

Heijden, P.G.M. van der (1987). **Correspondence analysis of longitudinal categorical data.** Leiden: DSWO-press.

Leeuw, J. de & Heijden, P.G.M van der (1987). The analysis of time budgets with a latent time-budget model. To appear in: E. Diday et al. (Eds). **Data analysis and informatics 5**. Amsterdam: North Holland.

Leeuw, J. de, Heijden, P.G.M van der & Kreft, I. (1985). Homogeneity analysis of event history data. **Methods of Operations research**, 50, 299-316.

McCutcheon, A.L. (1987). **Latent Class Analysis**. Sage University Paper series on Quantitative Applications in the Social Sciences,064-001. Beverly Hills: Sage Pubns.

Parkes, D.N. & Thrift, N.J. (1980). **Times, spaces and places**. New York: Wiley.

Robinson, J.P. (1985). The validity and reliability of diaries versus alternative time use measures. In: F.T.Juster & F.P.Stafford (Eds.). **Time, goods and well-being**. Michigan: University of Michigan.

Saporta, G. (1981). Méthodes exploratoires d'analyse de données temporelles. **Cahiers du Buro** no 37-38, Paris.

Staikov, Z. (ed.) (1982). **It's about time**. Sofia: Bulgarian Academy of Sciences.

Szalai, A. et al. (1972). **The use of time**. The Hague: Mouton.

Verbeek, A. (1984). The geometry of model selection in regression. In: T.K. Dijkstra, (ed.). **Misspecification analysis**.. Berlin: Springer-Verlag.

Werner, D. et al. (1979). Subsistence productivity and hunting effort in native South America. **Human Ecology**, 7, 303-315.