Interne notater

STATISTISK SENTRALBYRÅ

89/1 12. januar 1989

NOEN EKSEMPLER PÅ ENKLE MODELLER FOR LATENT STRUKTURANALYSE

av

Anders Rygh Swensen

1. Innledning..... 1
2. Latent model for tverrsnittsdata..... 5
3. Latent model for paneldata..... 9
4. Tilpassning av modellene til data om menn og kvinner..... 12
5. En mer kompleks modell..... 14
6. Forholdet mellom Goodmans og Habermans parametriseringer. 18
7. Estimeringsmetode..... 21
8. Oppsumering og konklusjon..... 21

Referenser..... 22

Appendix: Kjøreoppsett og utskrift for Habermans program. 23

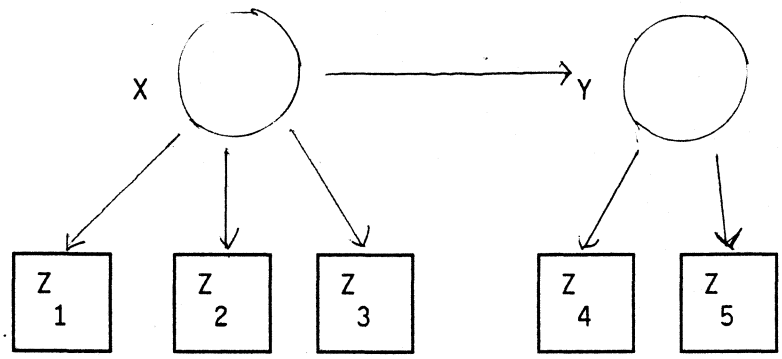
1.Innledning.

Vi skal betrakte et par situasjoner der observasjonsmaterialet består av kategoriske variable som viser høy korrelasjon. Utgangspunktet er en forestilling om at de ulike variablene er uttykk for et felles begrep, og at den høye korrelasjonen kan begrunnes av denne sammenhengen. I samfunnsfag er dette en situasjon som ofte forekommer, og vi skal presentere en formell metode for å analysere slike data. I avsnitt 1-5 skal vi gi noen eksempler på bruk av slike modeller, mens avsnitt 6 og 7 er av mer teknisk natur.

Hvis data er innhentet på et tidspunkt og viser høy korrelasjon, enten mellom alle variablene eller mellom subgrupper av dem, er det nærliggende å tro at det er en felles underliggende struktur som er årsaken. At de observerte variablene har høy samvariasjon, men ikke fullstendig overensstemmelse, er uttrykk for at denne strukturen ikke observeres direkte, men må begrunnes ut fra en teori om de forholdene en studerer.

Dette er bakgrunnen for modeller av typen LISREL, som er en videreutvikling av faktoranalysemodellene og bygger på forutsetningen om at responsvariabelen er kontinuerlig fordelt. Disse modellene ble i sin tid utviklet for data fra intelligenstester. Den underliggende faktor er intelligens, som en kan anta består av en rekke komponenter, og observasjonene er resultatet av forskjellige tester. Dette er en situasjon som har analogier innen en rekke felter i adferdsfag og naturvitenskap. Faktoranalysemodeller har da også blitt brukt i en rekke sammenhenger.

Ideen kan illustreres ved følgende figur der sirkler indikerer uobserverte variable og bokser observerte indikatorer.



Figur 1. Eksempel på enkel latent model.

Her kan vi for eksempel la X være de fysiske sidene ved arbeidsmiljø og Y være fysiske plager. Disse variablene observeres ikke direkte, men må operasjonaliseres ved en rekke indikatorer. De to sidene i diagrammet illustrer ideen bak faktoranalyse. Den observerte korrelasjonen mellom observasjonene skyldes de uobserverte underliggende faktorer.

LISREL modellene tillater i tillegg å modellere og teste årsakssammenhenger mellom de latente variable. Dette gjøres ved å stille opp et simultant likningssystem mellom de ukjente faktorene eller latente variable, som er en annen betegnelse. På figuren indikerer en pil en slik mulig sammenheng: Forhold ved det fysiske arbeidsmiljø medfører fysiske plager.

Også i andre situasjoner er det rimelig å vente høye korrelasjoner mellom de størrelsene en observer. Hvis data innhentes på flere tidspunkter om samme person ,og det dreier seg om noenlunde stabile kjennetegn, er det også grunn til å anta en høy grad av korrelasjon mellom opplysningene på hvert tidspunkt. Den enkleste form for tidsavhengighet er såkalt Markovavhengighet. En konsekvens av denne antagelsen er at bare nåtidstilstanden er av betydning. Hvordan man er kommet dit, spiller ingen rolle. Det er åpenbart at bare i de færreste situasjoner der en tar utgangspunkt i data om sosiale forhold, kan en rettferdiggjøre en slik antagelse. Dette kan skyldes flere forhold. For det første kan det være slike fenomener at hele forhistorien har betydning, selv om en kunne korrigere for alle mulige bakgrunns opplysninger. Det finnes imidlertid en enklere forklaring, som ikke automatisk bør utelukkes. Sett at bare nåtidstilstanden er av betydning for hvert individ, men at tilbøyeligheten til å skifte tilstand avhenger av et kjennetegn en ikke har opplysinger om. Resultatet vil være et en observerer overganger som ikke bare kan forklares av forhistorien, men også av det uobserverte kjennetegnet. Her har en ideen bak en enkel utvidelse av Markovmodellen som det er nærliggende å undersøke.

Blandingsparameteren, altså andelen av populasjonen som har det uobserverte kjennetegn, gjør at det er en formell likhet med modellene som antas for tverrsnittsdata. I begge tilfeller vil den høye korrelasjonen være et resultat av at parameteren er ukjent. Hadde vi kjent den og bare konsentrert oss om data der den er identisk, ville resultatet vært en enklere struktur: Henholdsvis uavhengighet og Markovantagelsen.

Når det gjelder tverrsnittsdata eksisterer det et motstykke til faktoranalysemodellen, som tar utgangspunkt i at responsvariabelen er kategorisk, såkalte latente strukturmodeller. De ble i sin tid foreslått av sosiologen Paul Lazarsfeld. Det virker som om muligheten til å modellere sammenhengen mellom de latente variable er mer begrenset i dette oppsettet, noe som kan forklare at modellene ser ut til å være mindre populære blant samfunnsvitere enn LISREL modellene.

Egenlig brukes latente strukturmodeller bare som betegnelse når den reduserte strukturen, den som gjenstår når en ser bort fra blandingen, er ren uavhengighet. Vi skal imidlertid være litt mer fleksible i språkbruken, og benytte betegnelsen også om modellene for longitudinelle data som ble omtalt ovenfor. Poenget vi ønsker å fremheve er at en relativt komplisert struktur reduseres til en enklere ved å trekke inn en uobserverbar latent variabel.

For å illusterere bruken av denne tankegangen skal vi se på en noe mer konkret problemstilling med utgangspunkt i to spørsmålsgrupper i Levekårsundersøkelsen fra 1980, 1983 og 1987. De to spørsmålsgruppene dreier seg om arbeidsmiljø og psykisk helse. For arbeidesmiljøet blir det spurt om daglige tunge løft, belastende arbeidsstillinger og om mye gjentatte og ensidige bevegelser. Svaralternativene er ja og nei. For psykisk helse dreier spørsmålene seg om forekomst av hyppig hjertebank, nervøsitet og følelse av depresjon. Her er det tre svaralternativer ofte, av og til og aldri.

Tabell 1. Svarfordeling i 1980, 1983 og 1987 på tre spørsmål om arbeidsmiljø, og endring i svar i de tre årene på tre spørsmål om arbeidsmiljø.

Spørsmål			Tidspunkt		
Tunge løft	Belastende arbeidsstilling	Ensidige bevegelser	1980	1983	1987
Ja	Ja	Ja	82	79	101
Ja	Ja	Nei	67	71	57
Ja	Nei	Ja	19	18	13
Ja	Nei	Nei	27	24	21
Nei	Ja	Ja	47	54	48
Nei	Ja	Nei	27	22	34
Nei	Nei	Ja	109	106	91
Nei	Nei	Nei	238	242	251
			616	616	616

Tidspunkt			Spørsmål		
1980	1983	1987	Tunge løft	Belastende arbeidsstilling	Ensidige bevegelser
Ja	Ja	Ja	113	129	118
Ja	Ja	Nei	29	28	55
Ja	Nei	Ja	15	25	31
Ja	Nei	Nei	38	41	53
Nei	Ja	Ja	28	40	53
Nei	Ja	Nei	22	29	31
Nei	Nei	Ja	36	46	51
Nei	Nei	Nei	335	278	224
			616	616	616

For den første spørsmålsgruppen er det 616 personer som har svart alle tre årene og som var over 20 år i 1980, for den andre gruppen er det 917. Ved å dele inn svarene i to kategorier (ja/nei for spørsmålene om arbeidsmiljø og aldri/ ofte og av og til for spørsmålene om psykisk helse) får en ialt 8x8x8 = 512 kombinasjoner ved å ta hensyn til alle svaralternativene i de tre årene.

Tabell 1 og tabell 2 viser noen marginale fordelinger for de to spørsmålsgruppene.

Som vi kan se er det hvert år en sterk korrelasjon mellom svarene i den forstand at svært mange svarer nei/aldri på alle tre spørsmål, og at det også er et betydelig antall som svarer ja/ofte og av og til på alle. Det siste er mest typisk for spørsmålene om arbeidsmiljø. Ved å betrakte svarfordelingen over tid ser en at for hvert spørsmål i de to gruppene er det en klar tendens til at alle svarer det samme i alle tre undersøkelsene. Men det er også en relativt stor andel som svarer forskjellig. Det er faktisk overrraskende at fordelingen på de to svartypene er såvidt konstante med så store overganger på individnivå.

Vi skal illustrere tankegangen bak de såkalte latente variabel modellene ved å tilpasse noen enkle modeller av denne typen til de 2x2x2 tabellene som er gjengitt ovenfor. For ikke å nedlesse framstillingen med detaljer vil hovedideene i modellene bli kommentert.

Tabell 2. Svarfordeling i 1980, 1983 og 1987 på tre spørsmål om psykisk helse, og endring i de tre årene i svarene på de tre spørsmålene.

Spørsmål			Tidspunkt		
Kraftig hjertebank	Nervøsitet	Depresjon	1980	1983	1987
Av og til/ ofte	Av og til/ ofte	Av og til/ ofte	33	31	32
Av og til/ ofte	Av og til/ ofte	Aldri	14	22	16
Av og til/ ofte	Aldri	Av og til/ ofte	4	11	9
Av og til/ ofte	Aldri	Aldri	20	23	31
Aldri	Av og til/ ofte	Av og til/ ofte	76	68	56
Aldri	Av og til/ ofte	Aldri	66	65	68
Aldri	Aldri	Av og til/ ofte	54	50	55
Aldri	Aldri	Aldri	652	647	648
			917	917	917

Tidspunkt			Spørsmål		
1980	1983	1987	Kraftig hjertebank	Nervø-sitet	Depre-sjon
Av og til/ ofte	Av og til/ ofte	Av og til/ ofte	31	61	15
Av og til/ ofte	Av og til/ ofte	Aldri	33	28	15
Av og til/ ofte	Aldri	Av og til/ ofte	25	33	15
Av og til/ ofte	Aldri	Aldri	78	68	25
Aldri	Av og til/ ofte	Av og til/ ofte	28	28	21
Aldri	Av og til/ ofte	Aldri	68	69	36
Aldri	Aldri	Av og til/ ofte	68	49	38
Aldri	Aldri	Aldri	568	581	752
			917	917	917

2. Latent modell for tversnittsdata.

Hovedideen bak modellen er følgende. Vi tenker oss at arbeidsmiljø kan deles inn i to typer, fysisk anstrengende og ikke anstrengende. Personer med fysisk anstrengende arbeid har størst tilbøyelighet til å svare ja på alle spørsmål. Men vi forlanger ikke at det skal være noen nøyaktig overensstemmelse i den forstand at en ut fra viten om type arbeidsmiljø nøyaktig skal kunne forutsi hvilke svar en person vil gi. Det er mange grunner til å tillate en slik løsere binding. For det første er det forhold som berører innsamling av data, som feil fra intervjuernes side og feil koding. Men det er også mer fundamentale grunner. I ethvert opplegg for spørreskjemaundersøkelser vil det være usikkerhet knyttet til valg av spørsmål og utforming av spørsmålsteksten. Intervjuobjektene vil oppfatte dette forskjellig avhengig av personlige og sosiale forhold. Selv om en to personer har en arbeidsituasjon som er så lik at en ved planleggingen av undersøkelsen venter at de skal gi samme svar, er det for mye forlangt at de absolutt skal gjøre det. Hva som oppleves som ensidige bevegelser, er ikke entydig gitt. Det samme gjelder tunge løft og belastende arbeidssituasjon. Men å innse at absolutt overensstemmelse er for mye å forlange behøver ikke nødvendigvis medføre at man må gi avkall på alle krav til sammenheng. Ved å betrakte bestemte arbeidssituasjoner er det grunn til å tro at andelen personer som svarer ja på spørsmålene er høyere enn tilsvarende andeler for personer med andre arbeidsforhold.

Dette er nokså banalt og uproblematisk. Det springende punkt i de latente modellene er at de gjør krav på å forklare den sammenhengen man kan observere mellom svar på ulike spørsmål ved å trekke inn de latente typene. Hvis man hadde kjent typen arbeidsmiljø, ville kjennskap til hva en person har svart på et spørsmål ikke gi noen ytterligere indikasjon på hva en vedkommende vil svare på et annet spørsmål. For å være helt konkret: Fra tabell 1 er det klart at hvis en person svarer nei på spørsmål om tunge løft er det god grunn til å vente at vedkommende også svarer nei på spørsmålet om ensidige bevegelser. Godtar man argumentasjonen om to typer arbeidsmiljø og argumentet om at dette forklarer samvariasjone i data, vil det være kjennskap til typen arbeidsmiljø som vil være avgjørende. Vet vi hilken type arbeidsmiljø en person har, vil vi ikke kunne si noe mer om hva vedkommende vil svare på spørsmålet om ensidige bevegelser ved å vite svaret på spørsmålet om tunge løft.

Dette er en argumentasjon som tar utgangspunkt i sentrale sider ved analyse av toveistabeller i samfunnsvitenskap. Mye av poenget ved analyse av slike data er å lete etter variable som forklarer sammenhenger en observerer. Kriteriet på at dette er tilfelle er nettopp at når man tar hensyn til verdien på den tredje variabelen forsinner samvariasjonen mellom de to opprinnelige variable. Det nye ved de latente modellene er at man ganske enkelt postulerer at en slik tredje variabel finns, men ikke gjør noe forsøk på å observere den direkte i første omgang.

En latent modell for spørsmålene om psykisk helse kan gis en tilsvarende begrunnelse som ovenfor.

Et fellestrekk ved begge modellene er at interessen er sentert rundt et begrep det egentlig ikke finnes noen objektiv definisjon av. Med utgangspunkt i svar på en del spørsmål er vi interessert i å si noe mer om variabelen eller begrepet. Validitetsproblematikken, spørsmålet om hva som egentlig måles og estimeres, er derfor av sentral

betydning. Det er nødvendig å sammenholde de empiriske resultatene med de forestillinger en har om fenomenet som undersøkes. Det er således legitimt å stille spørsmål om arbeidsmiljø slik det er benyttet ovenfor egentlig er noe fruktbart begrep. Er det grunnlag for å tro at arbeidsbetingelsene skiller seg så klart at det er rimelig å operere med to grupper, og å anta at svarfordelingen i de to gruppene er så forskjellig at det kan påvises ved hjelp av de spørsmålene som er stilt? God tilpasning av en statistisk modell kan ikke alene gi svar på slike spørsmål. Men modeller kan være til hjelp. Dårlig tilpasning indikerer at de empiriske resultatene svarer dårlig til de teoretiske forestillingene. Likeledes vil parameterestimater som avviker sterkt fra det en kan forvente være en indikasjon på at ikke alt stemmer.

Siden mye av hensikten med disse eksemplene er å illustrere et par enkle modeller med latente variable, skal vi ikke gå ytterligere inn på disse spørsmålene. Vi antar altså at det er godtgjort at det er av interesse å kartlegge størrelsen på de latente klassene og hvordan svarfordelingen er innen hver av klassene.

La oss nå se litt på hvordan ideene ovenfor kan formaliseres. La (i,j,k) betegne svaret på en bestemt spørsmålskombinasjon, der i,j,k antar verdien 1 hvis svaret er ja/ ofte eller av og til og verdien 2 hvis svaret er nei/aldri. La sannsynligheten for å tilhøre gruppa med et lite anstengende fysisk arbeid være p, og sannsynligheten for å tlhøre gruppa med fysisk anstrengende arbeid være 1 - p. For en person i den første gruppa antar vi nå at sannsynligheten for å svare (i,j,k) kan skrives

$$p_{i|2}\ p_{j|2}\ p_{k|2}$$

der i betgner tunge løft, j belastende arbeidssituasjon og k ensidige bevegelser.

Dette uttrykker ideen vi beskrev ovenfor om at hvis vi vet at en person hører til en bestemt latent klasse, vil vi ikke få noe ytterligere viten om hva en person vil svare på et bestemt spørsmål ved å kjenne svarene på ett eller flere av de andre.

Nå vet vi ikke hvilken gruppe en bestemt person hører til, slik at sannsynligheten for svarkombinasjon (i,j,k) vil være

$$(1-p)\ p_{i|1}\ p_{j|1}\ p_{k|1} \quad + \quad p\ p_{i|2}\ p_{j|2}\ p_{k|2}$$

som altså uttrykker sannsynligheten for at en person hører til den første gruppa og svarer (i,j,k) pluss sannsynligheten for at vedkommende hører til den andre gruppa og svarer (i,j,k).

Hovedantagelsen i modellen er at hadde en visst hvilken latent klasse en person tilhører, ville svarene på de tre spørsmålene vært uavhengige av hverandre. Eventuell observert avhengighet mellom svar på ulike spørsmål skyldes bare at den observerte tabellen er agregert over en ukjent dimensjon. Det er derfor to springende punkter i resonnementet: For det første at en inndeling i to grupper er rimelig. Dette er et fundamentalt begrepsmessig problem, og er i siste instans avhengig av hvor fruktbart det teoretiske utgangspunktet er. Dernest trengs en begrunnelse av antagelsen om betinget uavhengighet. Den representerer de sammenhengene modellen ikke gir noen mulighet for å forklare. Selv om vi vet at en person hører med til gruppa med lite

fysisk krevende arbeidssituasjon, er det mye forlangt å kreve at vedkommende svarer nei på alle tre spørsmål. Som nevnt tidligere er det mer rimelig å nøye seg med å forlange at andelen som svarer nei på alle tre spørsmål skal være høy i denne gruppa.

Tabell 3 viser resultatet av estimering for årene 1980, 1983 og 1987 i de to spørsmålsgruppene.

Tabell 3. Resultat av estimering av latent modell med to klasser for spørsmål om arbeidsmiljø og psykisk helse i 1980, 83 og 87.

		p	$p_{\cdot\|1}$			$p_{\cdot\|2}$			G^2
		klasse	tunge løft	bel. arb. sit	ens. bev.	tunge løft	bel. arb. sit	ens. bev.	
Arbeids-	1980	0.59	0.33	0.11	0.42	0.92	1.00	0.69	1.47
		0.64	0.33	0.00	0.42	0.88	1.00	0.67	3.22
miljø.	1983	0.59	0.34	0.11	0.41	0.94	1.00	0.70	7.21
		0.63	0.37	0.00	0.41	0.89	1.00	0.68	9.76
	1987	0.61	0.31	0.06	0.36	0.94	0.96	0.74	0.00
		0.61	0.34	0.00	0.38	0.91	1.00	0.72	2.63
		klasse	hj.b.	nerv.	depr.	hj.b.	nerv.	depr.	
Psykisk	1980	0.83	0.69	0.07	0.29	0.97	0.94	0.93	0.00
		0.80	0.76	0.00	0.43	0.98	1.00	0.92	6.04
helse.	1983	0.76	0.68	0.25	0.40	0.98	0.97	0.96	0.00
		0.80	0.71	0.00	0.47	0.95	1.00	0.92	17.98
	1987	0.84	0.62	0.18	0.30	0.96	0.94	0.94	0.00
		0.81	0.71	0.00	0.48	0.94	1.00	0.92	14.46

Den øverste linjen angir reultatet av estimering i en latent modell uten noen restriksjoner. Slik vi har parametrisert modellen, er det sannsynligheten for å svare nei som angis i tabellen. Det er en minus sannsynligheten for å svare ja.

For arbeidsmiljøspørsmålene ser vi at materialet deles i to grupper med omlag 60% i den gruppa som kan antas å representere personer med lite fysisk krevende arbeidsmiljø. Den andre gruppa utgjør omlag 40% i alle tre årene. Legg merke til at spørsmålene om ensidige bevegelser stort sett besvares med ja i den ene latente gruppa og med nei i den andre. Siden det bare er to latente grupper betyr det at den latente inndelingen, svarer relativt godt til svaret på spørsmålet om belastende arbeidsstilling. Dette inntrykk bekreftes for eksempel ved å betrakte kryssproduktforholdene i de 2x2 tabellene som framkommer ved å betinge fordelingen i 1987 med hensyn på:

(i) svaret på spørsmålet om belastende arbeidsstilling (1.3 og 1.7)
(ii) svaret på spørsmålet om tunge løft (2.9 og 3.9).
(iii) svaret på spørsmålet om ensidige bevegelser (14.7 og 20.0)

Den latente inndelingen og inndelingen representert ved svaret på spørsmålet om belastende arbeidsstillling er derfor relativt sammenfallende.

For spørsmålene om psyko-somatiske forhold ser vi at det er nervøsitet som ligger nærmest den latente inndelingen.

Det er relativt enkelt å teste hvorvidt den latente og manifeste dimensjonen faller sammen. Den andre linjen i tabell 3 angir resultatene av en estimering under denne forutsetningen. Som en ser G^2 er økningen i G^2, dvs. sannsynlighetskvoten som er det vanlige mål for tilpassning, ikke i noe tilfelle tilstrekkelig ($X^2_{0.95,2}$ fraktilen er 5.99) til å avvise en slik reduksjon i tilfellet med arbeidsmiljøspørsmålene. For den andre spørsmålsgruppen er økningen så stor at det ikke er noen grunn til å tro at den latente inndelingen faller sammen med svaret på spørsmålet om nervøse plager.

For de modellene vi opererer med har vi nå estimert alle ukjente størrelser. Dette gjør det mulig å regne ut sannsynligheten for å høre til en bestemt klasse gitt et visst obsevasjonsmønster. På denne måten får vi en rangering av de 8 ulike svarkategoriene, eller det man ofte kaller en skåre eller en indeks. Problemet her er den innbyrdes rangering av kategorier som (1,1,2) , (1,2,1) og (2,1,1), dvs. der en svarer ja på et av spørsmålene. Siden det ikke er noen naturlig rangering av spørsmålene, kan det være et problem hvis en ønsker å ordne de essensielt tredimensjonale svaraltertnativene langs en endimensjonal skala. Ved å ta utgangspunkt i en latent strukturmodel regner en ganske enkelt ut

$$P(\text{ klasse } 2 \mid \text{observasjon } (i,j,k)) .$$

Tabell 4 viser resultatet av slike beregninger for arbeidsmiljøspørsmålene i 1987. Husk at den første variabelen referer til spørsmålet om tunge løft, den andre til belastende arbeidssituasjon og den tredje til spørsmålet om ensidige bevegelser.

Tabell 4.

Svarfordeling	Skåre
(1,1,1)	0.0021
(1,1,2)	0.0108
(1,2,1)	0.4643
(1,2,2)	0.8146
(2,1,1)	0.0658
(2,1,2)	0.2631
(2,2,1)	0.9660
(2,2,2)	0.9931

Vi ser igjen betydningen av spørsmålet om belastende arbeidsstillig. Det er langt på vei svaret på det som avjør om en hører til en bestemt klasse.

3. Latent model for paneldata.

For paneldataene ser vi at det er en tendens til overrepresentasjon av personer langs digonalen. Det betyr at det en svarer er det samme som forrige gang. Dette er et fenomen som går igjen i mange sammenhenger, og ofte gjør at rene Markovmodeller gir dårlig tilpasning for samfunnsvitenskaplige data, jfr. Bartholomew (1982). Markovantagelsen uttrykker at hva en person svarer bare avhenger av forrige gangs svar, ikke av hva en har svart ved tidligere anledninger. Dette er den enkleste avhengighet en kan ha av fortida. Ved enhver matematisk modellering av paneldata vil det derfor være naturlig å referere seg til denne situasjonen.

Et alternativ til den rene Markovmodell , som vi skal forsøke, er den såkalte "mover-stayer" modell. Den ble første gang foreslått i 1955 av Blumen, Kogan og McCarthy for å beskrive mobilitet mellom næringer i USA. En tenker seg at observasjonene er et resultat av to ulike bevegelser. For det første bevegelsene representert av "movers": En antar at de kan beskrives ved en Markovkjede. Dernest bevegelsene til "stayers": Disse har en sterk tilbøyelighet til å bli i jobben. Deres bevegelser beskrives derfor ved at de er i samme tilstand hele tida. Nå observerer en ikke om en person er "mover" eller "stayer". Denne tilhørigheten representer derfor en latent dimensjon slik vi har benyttet begrepet i forrige avsnitt.

En kan tenke seg at slike latente mekanismer gjør seg gjeldende for arbeidesmiljøspørsmålene. Her representerer "stayers" personer som opplever situasjonen på samme måte ved alle tre høve. "Movers" er en gruppe som enten faktisk har forandret arbeidsvilkår eller oppfatter spørsmålene annerledes. Andelen "stayers" representerer derfor stabiliteten i materialet. Hvis den er nær en, er det høy grad av samsvar mellom svarene ved de tre aneldningene.

En tilsvarende begrunnelse kan en tenke seg for spørsmålene om psykisk helse.

Analogien til modellen for tversnittsdata er derfor klar. I begge tilfeller antar vi at observasjonene stammer fra to grupper, men ut fra observasjonene kan vi ikke fastslå hvilken. Innen hver av gruppene antar vi et bestemt mønster, definert ved forutsetningen om betinget uavhengighet for tverrsnittsadataene og Markovantagelsen for paneldataene.

La (i_1, i_2, i_3) betegne svarene i 1980,83 og 87 for en person på spørsmålet om tunge løft. Innen en bestemt latent klasse uttrykker vi sannsynligheten for svarkombinasjon (i_1, i_2, i_3) som

$$p^{123}_{i_1 i_2 i_3} = p^{3|12}_{i_3 | i_1 i_2} \; p^{12}_{i_1 i_2} = p^{3|12}_{i_3 | i_1 i_2} \; p^{2|1}_{i_2 | i_1} \; p^{}_{i_1} .$$

Foreløbig er dette bare en omskrivning av sannsynligheten for en bestemt svarkombinasjon og representerer ingen begrensning. Markov-antagelsen innebærer at

$$p^{3|12}_{i\,|i\,i\,\,i\,\,i} = p^{3|2}_{i\,\,|\,\,i\,\,i\,\,i}\,,$$

som nettopp sier at overgangssannsynligheten fra tidstunkt 2 (1983) til tidspunkt 3 (1987) bare avhenger av tilstanden på tidspunkt 2. I tillegg vil vi anta at utviklingen fra 1980 til 1983 er den samme som utviklingen fra 1983 til 1987, d.v.s.

$$p^{3|2}_{\cdot|\cdot} = p^{2|1}_{\cdot|\cdot} = p^{\,}_{\cdot|\cdot}\,.$$

Dette er en modelforutsetning og burde strengt tatt vært testet. For eksempel er det tre år mellom de to første tidspunktene og fire år mellom de neste. På den andre siden vil det ikke være mulig å operere med to latente klasser for så få tidspunkter som 3 hvis den ikke godtas. Siden hensikten med dette notatet i høy grad er å illustrere og utprøve modeller av denne typen, har vi valgt å godta denne forutsetningen.

For å skille mellom parametrene i de to latente klassene angir vi klassen i en parantes. Det betyr at sannsynligheten for å observere svarkombinasjon (i_1, i_2, i_3) er gitt ved

$$p^{123}_{i\,i\,i} = (1-p)\,p^{123}_{i\,i\,i}(1) + p\,p^{123}_{i\,i\,i}(2)$$

$$= (1-p)\,p^{(1)}_{i}\,p^{(1)}_{i\,|i}\,p^{(1)}_{i\,|i} + p\,p^{(2)}_{i}\,p^{(2)}_{i\,|i}\,p^{(2)}_{i\,|i}$$

Vi lar nå "movers" være klasse 1, slik at denne Markovkjeden er uten noen restriksjoner. For å uttrykke at svaret blir det samme for "stayers" lar vi $p^{(2)}_{i_2|\,i_1}$ være en hvis i_1 er lik i_2 og null hvis de er forskjellige.

Resultatene for de seks svartypene og tre modellene er sammenfattet i tabell 5.

Tabell 5. Resultatet av estimering av latent model for paneldata for seks typer spørsmål og tre modeller.

	p	$p^{(1)}_2$	$p^{(1)}_{11}$	$p^{(1)}_{22}$	$p^{(2)}_2$	$p^{(2)}_{11}$	$p^{(2)}_{22}$	G^2
Tunge løft	0,52	0.60	0.56	0.72	0.76	1.00	1.00	0.08
Belastende arbeids stilling	0.47	0.59	0.55	0.63	0.69	1.00	1.00	1.71
Ensidige bevegelser	0.26	0.49	0.63	0.63	0.85	1.00	1.00	0.01
Hjertebank	0.76	0.71	0.37	0.64	0.99	1.00	1.00	0.08
Nervøsitet	0.57	0.64	0.28	0.64	0.91	1.00	1.00	2.4
Depresjon	0.43	0.72	0.30	0.76	0.96	1.00	1.00	0.29

Parameterestimatene for arbeidsmiljøspørsmålene virker rimelige, og tilpassningen er god: Husk at det er fem fri parametre og syv uavhengige observasjoner. Vi merker oss at stabiliteten definert ved sannsynligheten for å høre til stayers er markert mindre for spørsmålet om ensidige bevegelser enn de to andre. Dette er ikke særlig overraskende. Av tabell 1 ser en uten videre at det er færre som svarer nei alle tre ganger på dette spørsmålet enn på de to andre.

En rimelig hypotese å undersøke kan være om Markovkjeden er stasjonær, som blant annet medfører at fordelingen på de tre tidspunktene er den samme. Det er ikke vanskelig å utføre en formell test basert på sannsynlighetskvoten. Vi skal i stedet beregne stasjonærfordelingen fra den estimerte overgangsmatrisen og sammenligne den med hva som observeres på første tidspunkt. Stasjonærfordeligen finner en ved å løse likningssystemet

$$(1-x,\ x) = (1-x,\ x)\begin{bmatrix} p^{(1)}_{11} & 1-p^{(1)}_{11} \\ 1-p^{(1)}_{22} & p^{(1)}_{22} \end{bmatrix}.$$

Ved å sette inn de estimerte verdiene får en 0.61, 0.54 og 0.50, mens verdiene i tabell 5 er 0.60, 0.59 og 0.49 for de tre spørsmålene om arbeidsmiljø. Det er god overensstemmelse mellom de to tallsettene, slik at mye tyder på at Markovkjeden faktisk er stasjonær.

Den substansielle tolkningen er at det ikke er noen tilbøyelighet til å svare annerledes over tid. Nå er det de samme personene som inngår i undersøkelsen hele tiden slik at de observerte overgangene kan skyldes i alle fall to bevegelser: At personene blir eldre og derfor opplever arbeidssituasjonen som mer belastende og derfor har en større tilbøyelighet til å svare ja på spørsmålene, og på den andre siden at intervjuobjektene faktisk skifter arbeid og finner lettere arbeidssituasjoner. Det siste vil medføre en sterkere tendens til å svare nei. Vi ser altså at i den grad en kan redusere de faktisk observerte overgangene til et resultat av disse to bevegelsene, har de en tendens til å motvirke hverandre.

Det kan være grunn til å stille spørsmålstegn ved svarene for psykosomatiske lidelser. Ser vi for eksempel på svarene om hjertebank er overgangsmatrisen for de to latente klassene

0.71	0.29	1.0	0.0
0.63	0.37	0.0	1.0

Hvis verdiene utenfor hoveddiagonalen er større enn elementene på hoveddiagonalen, medfører det en osillerende adferd. Det er altså en sterk tilbøyelighet til å skifte tilstand fra et tidspunkt til et annet. Det er ingen grunn til å tro at dette er en egenskap man finner uten i særdeles spesielle tilfeller. Spørsmålet er nå om 0.64 - 0.36 i eksemplet ovenfor faller i denne kategorien.

4. Tilpassing av modellene til data om menn og kvinner.

Det er ofte av interesse å tilpasse modeller for menn og kvinner separat. Forskjeller i modeller og parameterestimater kan si noe vesentlig om fenomenet vi er interessert i. Vi har tilpasset modellene fra avsnitt 2 og 3 til arbeidsmiljøspørsmålene, og resultatene er gjengitt i tabell 6 og 7.

Tabell 6. Resultat av estimering av latent modell med to klasser for spørsmål om arbeidsmiljø for menn og kvinner i 1980, 83 og 87.

		p	$p_{\cdot\|1}$			$p_{\cdot\|2}$			G^2
		klasse	tunge løft	bel. arb. sit.	ens. bev.	tunge løft	bel. arb. sit.	ens. bev.	
Menn:	1980	0.59	0.22	0.07	0.46	0.91	0.95	0.73	0.00
		0.59	0.27	0.00	0.48	0.87	1.00	0.72	1.22
	1983	0.55	0.29	0.16	0.45	0.94	1.00	0.77	3.78
		0.63	0.29	0.00	0.45	0.87	1.00	0.74	8.46
	1987	0.59	0.28	0.06	0.45	0.95	0.96	0.78	0.00
		0.59	0.32	0.00	0.47	0.92	1.00	0.76	1.89
Kvinner:	1980	0.65	0.47	0.16	0.30	0.93	1.00	0.65	4.86
		0.71	0.47	0.00	0.30	0.89	1.00	0.63	6.12
	1983	0.63	0.40	0.02	0.35	0.93	1.00	0.61	2.63
		0.64	0.40	0.00	0.35	0.92	1.00	0.61	2.65
	1987	0.65	0.33	0.05	0.19	0.91	0.96	0.61	0.00
		0.64	0.37	0.00	0.22	0.89	1.00	0.66	1.11

Som man ser er det ikke de store endringene for tverrsnittsdata. Det kan være noe dårlig tilpasning: Særlig for kvinner. På den andre siden er modellen og parameterestimatene svært stabile over tid. Vi legger igjen merke til betydningen av spørsmålet om belastende arbeidssituasjon. Når det betinges med hensyn på dette spørsmålet, er det liten samvariasjon mellom ensidige bevegelser og tunge løft. Det er med andre ord et sterkt sammenefall mellom den latente dimensjonen og svaret på spørsmål om belastende arbeidsstilling. Av interesse er også at parameterestimatene svarende til spørsmålene om tunge løft og ensidige bevegelser viser en motsatt tendens hos menn og kvinner. Det er større tilbøyelighet til tunge løft enn til ensidige bevegelser blant menn i fysisk anstrengende arbeid. Hos kvinner er det omvendt.

Tabell 7. Resultatet av estimering av latent model for paneldata for tre arbeidsmiljøspørsmål: Menn og kvinner.

	p	$p^{(1)}_2$	$p^{(1)}_{11}$	$p^{(1)}_{22}$	$p^{(2)}_2$	$p^{(2)}_{11}$	$p^{(2)}_{22}$	G^2
Menn:								
Tunge løft	0.49	0.55	0.54	0.75	0.70	1.00	1.00	2.31
Belastende arbeids stilling	0.52	0.51	0.55	0.62	0.66	1.00	1.00	3.70
Ensidige bevegelser	0.39	0.47	0.56	0.57	0.85	1.00	1.00	0.31
Kvinner:								
Tunge løft	0.58	0.68	0.59	0.65	0.84	1.00	1.00	4.24
Belastende arbeids stilling	0.40	0.68	0.55	0.65	0.74	1.00	1.00	0.16
Ensidige bevegelser	0.01	0.52	0.72	0.70	0.99	1.00	1.00	0.13

Når vi betrakter spørsmålene over tid, er det mest påfallende hvor liten overrepresentasjonen er av kvinner som svarer det samme alle tre ganger på spørsmålet om ensidige bevegelser. Liten overrepresentasjon tas her i betydning at en ren Markovkjede ser ut til å føye dataene tilfredsstillende. Det er ikke tilfellet for noen av de andre spørsmålene.

5. En mer kompleks modell.

Figur 1 illustrerer en mulig modell for latente variable over tid. Her representerer X den latente variabel på tidspunkt 1 og Y den samme variabel på tidspunkt 2. Gitt verdien av variabelen X er indikatorene på tidspunkt 1 uavhengig fordelt, og også uavhengige av indikatorene på tidspunkt 2. Tilsvarende gjelder for variabelen Y. Hvis X og Y hver er dikotome, kan fordelingen til X og Y fremstilles i følgende tabell.

X \ Y		
	1	2
	3	4

Ved å kombinere de mulige verdier av X og Y vil vi altså ha en latent model der den latente variabelen kan anta 4 verdier som kan skrives

$$p_{i\,j\,k\,i\,j\,k\;\;1\,1\,1\,2\,2\,2} = \sum_{t=1}^{4} p_{i\,|t\;1}\, p_{j\,|t\;1}\, p_{k\,|t\;1}\, p_{i\,|t\;2}\, p_{j\,|t\;2}\, p_{k\,|t\;2}\, p_{t} \,.$$

Dette gir en modell med 27 ukjente parametre.

Nå antar vi i tilegg at indikatorene på tidspunkt 1 bare avhenger av verdien til X og ikke hvilken verdi Y har. Det betyr med andre ord at

$$p_{i\,|1\;1} = p_{i\,|2\;1} \qquad p_{j\,|1\;1} = p_{j\,|2\;1} \qquad p_{k\,|1\;1} = p_{k\,|2\;1}$$

$$p_{i\,|3\;1} = p_{i\,|4\;1}\,, \qquad p_{j\,|3\;1} = p_{j\,|4\;1} \quad \text{og} \quad p_{k\,|3\;1} = p_{k\,|4\;1}\,.$$

Tilsvarende lar vi indikatorene på tidspunkt 2 bare avhenge av Y,

$$p_{i\,|1\;2} = p_{i\,|3\;2} \qquad p_{j\,|1\;2} = p_{j\,|3\;2} \qquad p_{k\,|1\;2} = p_{k\,|3\;2}$$

$$p_{i\,|2\;2} = p_{i\,|4\;2}\,, \qquad p_{j\,|2\;2} = p_{j\,|4\;2} \quad \text{og} \quad p_{k\,|2\;2} = p_{k\,|4\;1}\,.$$

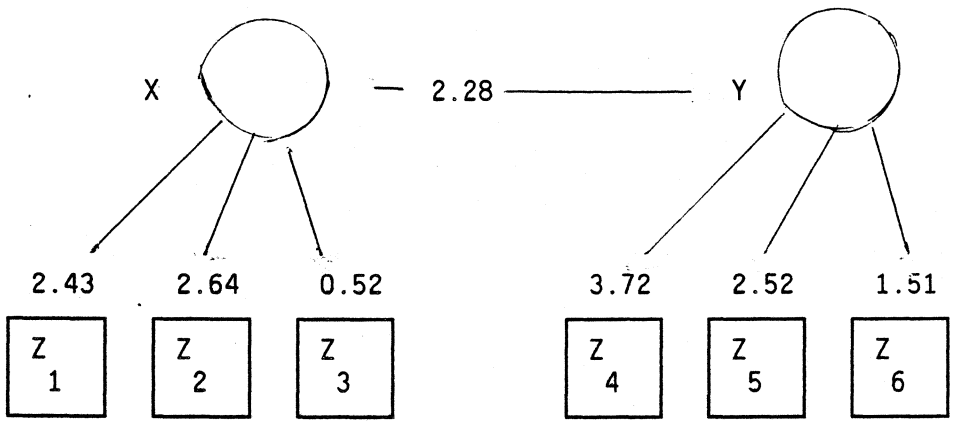
Restriksjonene ovenfor gjør at parameterantallet reduseres til 15.

Første linje i Tabell 8 gir overgangstallene for de mulige transisjonene fra 1980 til 1983. Ialt kan indikatorene anta 8 = 2x2x2 verdier på hvert tidspunkt, slik at det dreier seg om 64 mulige overganger.

Fordelingen til X og Y kan nå estimeres og gir følgende tabell

		Y
X	0.41	0.02
	0.09	0.48

Som vi ser er det en klar sammenheng mellom de to variablene. Ved å beregne halvparten av logaritmen til kryssproduktforholdet for tabellen på forrigfe side og i tabellene som angir forholdet mellom X og Y og indikatorene, får en følgende figur analogt til Goodman (1973b):



Den andre raden i tabell 8 angir den tilpassede modellen. Som vi ser er tilpassningen ikke særlig god. Dette vises også av en stor verdi av X^2, 549.0. Det er 15 ukjente parametre under modellen slik at den riktige tilnærmelsen vil være X^2 med 63 - 15 = 48 frihetsgrader. 549 ligger langt utenfor verdier som kan anses som akseptable. Ved å se på tabellen er det tydelig at det særlig er diagonalelementene som er dårlig tilpasset.

I første rekke av pedagogiske grunner kan det være instruktivt å se hva en reduksjon av den latente variabel til svaret på spørsmålet om belastende arbeidsstilling innebærer. Det betyr at vi setter:

$$p_{j_1|1} = p_{j_1|2} = p_{j_2|1} = p_{j_2|3} = 0.0$$

$$p_{j_1|2} = p_{j_1|4} = p_{j_2|2} = p_{j_2|4} = 1.0$$

Resultatet av en slik tilpassning er gitt i tredje linje i tabell 8. Nå er X^2 = 700.6. Differensen 700.6 - 549.0 = 151.6 skal nå sammenlignes med X^2-fordelingen med 4 frihetsgrader, som gir inntrykk av hvor dårlig tilpassning det er.

Tabell 8. Observerte tabeller for overgang mellom åtte svarkategorier fra 1980 til 1983.

				1983							
		Tunge løft:				Ja				Nei	
		Belastende arbeidsstilling:		Ja		Nei		Ja		Nei	
		Ensidige bevegelser:		Ja	Nei	Ja	Nei	Ja	Nei	Ja	Nei
	Tunge løft	Bel.arb. stilling	Ens. bev.	65	20	8	8	5	2	8	5
				34.2	31.7	9.9	9.2	10.5	9.7	3.4	8.9
			Ja	36.5	27.6	0.2	2.1	9.5	7.1	2.4	27.1
		Ja	Nei	17	34	2	5	2	6	6	10
				34.1	31.7	9.9	9.2	10.5	9.8	3.4	9.0
				36.4	27.5	0.2	2.1	9.4	7.1	2.3	26.9
			Ja	8	4	4	1	2	1	2	3
				5.6	5.2	1.6	1.5	1.7	1.7	0.7	4.1
				1.0	0.8	0.1	0.7	0.3	0.2	0.8	9.0
1980	Ja	Nei	Nei	4	9	2	8	3	0	1	9
				5.9	5.5	1.7	1.6	1.8	1.9	1.0	9.0
				2.7	2.1	0.2	1.8	0.7	0.5	2.0	23.4
			Ja	12	3	1	0	24	0	11	8
				8.6	8.0	2.5	2.3	2.6	2.5	1.0	4.9
				12.0	9.1	0.1	0.7	3.1	2.3	0.8	8.9
		Ja	Nei	4	5	1	1	8	7	1	9
				8.9	8.2	2.6	2.4	2.7	2.7	1.3	9.9
				11.9	9.0	0.1	0.7	3.1	2.4	0.8	8.8
			Ja	6	0	4	0	14	2	54	45
				6.4	6.0	1.9	1.9	2.1	3.7	5.1	85.6
				9.9	7.5	0.6	6.6	2.6	1.9	7.4	85.4
	Nei	Nei	Nei	7	10	2	5	9	8	52	205
				15.1	14.6	4.6	4.7	5.1	9.9	14.2	243.0
				25.7	19.4	1.5	17.1	6.6	5.0	19.2	221.2

6 Forholdet mellom Goodmans og Habermans parametriseringer.

Den parametrisering som er benyttet for tverrsnittsdata i kapittel 1, er den vanlige som benyttes blant annet av Lazarsfeld og Henry (1968) og Goodman (1973a og b). Haberman (1979) bruker en annen parametrisering. Noen ord om forholdet mellom dem er på sin plass: Særlig fordi Haberman i sin bok har et lett tilgjengelig dataprogram, mens hans parametrisering ikke er standard i denne typen modeller.

La oss ta utgangspunkt i en 2x2x2 tabell med to latente klasser og vise hvordan parametrene i de to parametriseringene forholder seg til hverandre. Eksemplet skulle være tilstrekkelig generelt til å illustrere de mer almenne ideer som ligger bak. Den tradisjonelle parametriseringen som ble brukt i kapittel 1, er

$$p_{ijkt} = p_{i|t}\, p_{j|t}\, p_{k|t}\, p_t$$

hvor $\sum_{i=1}^{2} p_{i|t} = \sum_{j=1}^{2} p_{j|t} = \sum_{k=1}^{2} p_{k|t} = \sum_{t=1}^{2} p_t = 1$ for t=1,2. Det gir ialt sju frie parametre. Siden $\Sigma\, x_{ijk} = n$, er det like mange ukjente parametre som observasjoner.

Haberman tar utgangspunkt i formuleringen som er gjengs i log-lineære modeller og skriver:

$$\begin{aligned} \log n\, p_{ijkt} &= \log n\, p_{i|t}\, p_{j|t}\, p_{k|t}\, p_t \\ &= \log n + \log p_{i|t} + \log p_{j|t} + \log p_{k|t} + \log p_t \\ &= \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_t^T + \lambda_{it}^{AT} + \lambda_{jt}^{BT} + \lambda_{kt}^{CT}. \end{aligned} \qquad (1)$$

der T betegner den latente dimensjonen og A, B og C betegner de tre dimensjonene i 2x2x2 tabellen. Nå skal λ'ene tilfredsstille

$$\sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_k \lambda_k^C = \sum_t \lambda_t^T = \sum_i \lambda_{it}^{AT} = \sum_t \lambda_{it}^{AT} = \ldots = 0.$$

som ialt gir sju parametre som vi kan la være $\lambda_1^A, \lambda_1^B, \ldots, \lambda_{11}^{BT}, \lambda_{11}^{CT}$. Her kan vi sløyfe fotskriften for enkelhets skyld og skriver altså parametrene $\lambda^A, \lambda^B, \lambda^C, \lambda^T, \lambda^{AT}, \lambda^{BT}$ og λ^{CT}.

Det er ikke vanskelig å se bakgrunnen i faktorielle loglineære modeller for formen (1). Husk at vi opererer med modeller der de observerte faktorene er uavhengige gitt den latente dimensjonen. Det betyr i et loglineært oppsett at vi har en modell med første ordens samspill med alle samspillsledd som ikke indikerer den latente dimensjonen satt lik null. Dette gir akkurat formen (1).

Det er også verdt å merke seg at (1) kan skrives på matriseform.

$$\{ \log n \, p_{ijkt} \} = \lambda \underline{e} + A\underline{\lambda}$$

der $\underline{\lambda} = (\lambda^{A}, \lambda^{B}, \lambda^{C}, \lambda^{T}, \lambda^{AT}, \lambda^{BT}, \lambda^{CT})^{t}$, $\underline{e} = (1,\ldots,1)^{t}$ og

$$A = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & -1 & -1 & -1 & -1 \\
1 & 1 & -1 & 1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 & 1 \\
1 & -1 & 1 & 1 & 1 & -1 & 1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & -1 & -1 & 1 & 1 \\
-1 & 1 & 1 & 1 & -1 & 1 & 1 \\
-1 & 1 & 1 & -1 & 1 & -1 & -1 \\
-1 & 1 & -1 & 1 & -1 & 1 & -1 \\
-1 & 1 & -1 & -1 & 1 & -1 & 1 \\
-1 & -1 & 1 & 1 & -1 & -1 & 1 \\
-1 & -1 & 1 & -1 & 1 & 1 & -1 \\
-1 & -1 & -1 & 1 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & 1 & 1 & 1
\end{bmatrix}$$

Dette gir en sammenheng med det vanlige variansanalyseoppsettet.

A er nå designmatrisen.

Sammenhengen mellom parametrene $p_{i|t}$, $p_{j|t}$ og $p_{j|t}$ og $\lambda^{A}, \lambda^{B}, \lambda^{C}, \lambda^{AT}, \lambda^{BT}$ og λ^{CT} er som følger. $p_{i|t}$, i=1,2 kan for for eksempel skrives:

$$p_{2|1} = \frac{\sum_{k,j} p_{2kj1}}{\sum_{k,j} p_{1kj1} + \sum_{k,j} p_{2kj1}} = \frac{\exp(-\lambda^{A} - \lambda^{AT})}{\exp(\lambda^{A} + \lambda^{AT}) + \exp(-\lambda^{A} - \lambda^{AT})} = \tag{2}$$

$$= \frac{1}{1 + \exp(2\lambda^{A} + 2\lambda^{AT})}$$

Tilsvarende er

$$p_{2|2} = \frac{1}{1 + \exp(2\lambda^{A} - 2\lambda^{AT})} . \tag{3}$$

p_2^T er derimot vanskeligere å uttrykke, men størrelsen kan skrives

$$p_2^T = \sum_{i,j,k} p_{ijk2} = \frac{\sum_{i,j,k} n\, p_{ijk2}}{n}$$

$$= \frac{\sum_{q_A,q_B,q_C} \exp(q_A\lambda^{A} + \ldots + q_C\lambda^{C} + \lambda^{T} + q_A\lambda^{AT} + \ldots + q_C\lambda^{CT})}{\sum_{q_A,q_B,q_C,q_T} \exp(q_A\lambda^{A} + \ldots + q_C\lambda^{C} + q_T\lambda^{T} + q_Aq_T\lambda^{AT} + \ldots + q_Cq_T\lambda^{CT})}$$

der q_A, q_B, q_C og q_T antar vedien -1 eller 1, slik at tellerer har åtte ledd og nevneren seksten.

Når det gjelder modeller for data på panelform, vil det for modeller av Markovtypen være nødvendig med samspillsledd mellom et år, det foregående og den latente dimensjon. Det inngår altså andre ordens samspill i formuleringen (1). Det springende punkt for latente modeller av denne typen er imidlertid at en av Markovkjedene tas som kjent. Det ser ikke ut til at Habermans program tillalter slike restriksjoner på parametrene.

Eksempel I appendixet er det gjengitt resultatet av kjøring av Habermans program på data fra tabell 1 (arbeidsmiljøspørsmal i 1987). Innsetting i (2) og (3) gir

$$p_{2|1} = 1/(1 + \exp(\ 2(\ -0.46751 + 0.87320)) = 1/(\ 1 + \exp(0.80938))$$

$$= 0.3080$$

og

$$p_{2|2} = 1/(1 + \exp(\ 2(\ -0.46751 - 0.87320)) = 1/(\ 1 + \exp(-2.67942))$$

$$= 0.9358$$

som nøyaktig er de tilsvarende verdiene i tabel 2.

7. Estimeringsmetode.

Ikke i noen tilfeller resulterer de latente modellene i lukkede former for sannsynlighetssetimatorer. Vi er derfor avhengige av gode iterasjonsprosedyrer. Haberman (1979) inneholder et program som er basert på en modifisert skåringsalgoritme. Det konvergerer raskt. Av andre positive egenskaper kan nevnes at det gjør bruk av det log-lineære oppsettet som gir stor fleksibilitet i modelformuleringen. Dessuten gir skåringsalgoritmen kovariansmatrisen til estimatorene som et biprodukt. Av mer negative trekk kan nevnes at log-lineære parametriseringer ikke er standard i denne typen modeller, og at restriksjoner på parametrene ikke er mulig. Særlig ser det ut til å være avhengig av gode startverdier. Ellers konvergerer ikke prosedyren og programmet aborteres.

Resultatene i avsnitt 2-6 er framkommet ved bruk av den såkalte EM-algoritmen, se Everitt (1984) for en nærmere beskrivelse, og det er brukt spesielt utviklede programmer for denne algoritmen. Den konvergerer langsomt, men ser ut til å langt mindre avhengig av startverdiene. Problemet er at program ikke ser ut til å være offentig tilgjengelig selv om de kan skaffes Clogg (1976) og Hagenaars (1988). Dessuten får en ikke kovariansmatrisen til estimatene uten ekstra beregninger.

8. Oppsummering og konklusjon.

Vi har i dette notatet estimert noen enkle latente modeller for kategoriske data. Resultatet er ikke udelt vellykket. Det viktigste problemet er knyttet til rettferdiggjøring av modellen. Det er nødvendig å gi den latente variabelen en substansiell forklaring. Det kan med gode grunner reises tvil om dette er tilfelle for de data som har vært nyttet. Dessuten gir ikke annet enn de enkleste modeller en rimelig god tilpassning.

Ideen er imidlertid besnærende og kan utvilsomt finne anvedelse i mer velegnede situasjoner enn de vi har sett på. Det er imidlertid viktig å være klar over at modeller av denne typen har en rekke problematiske numeriske og statistiske sider, slik at heller ikke av denne grunn er det modeller som kan brukes rutinemessig.

I tillegg til de problemene som er nevnt ovenfor, vil vi til slutt nevne ett som vi har sett bort fra, spørsmålet om identifiserbarhet. Er det slik at forskjellige verdier av parametrene leder til ulike fordelinger for observatorene. Hvis den latente variabelen er diskret og kan anta mange verdier, er det ikke tilfellet for kategoriske data. Da har det ingen mening å estimere modellen. Spørsmålet om hvilke restriksjoner som må innføres ser det ikke ut å være noe kjent svar på.

Referenser

Bartholomew, D.J. (1982). Stochastic Models for Social Processes. Thrd. ed., Wiley, New York.

Blumen, I., Kogan, M. og McCarthy, P.J. (1955). The Industrial Mobily of Labor as a Probability Process. Cornell University Press, Itacha, New York.

Clogg, C.C. (1977). Unrestricted and Restricted Maximum Likelihood Latent Structure Analysis: A Manual for Users. Working Paper 1977-09, Pennsylvania State University.

Everitt, B.S. (1984). An Introductin to Latent Variable Models. Chapman and Hall, London.

Goodman, L.A. (1974a). Exploratory latent structure analysis using both identfiable and unidentifiable models. Biometrika 61 215-231.

Goodman, L.A. (1975b). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. The Amer. J. of Sociol. 79 1179-1259.

Hagenaars,J.A. (1986). Latent structure models with direct effects between indicators: Local dependence models. Sociological Meth. and Research 16 374-405.

Haberman,S.J. (1979). Analysis of Qualitative Data. Vol. 1 and 2. Academic Press, New York.

Lazarsfeld,P.F. and Henry,N.W. (1968). Latent Structure Analysis. Houghton Mifflin, Boston.

Appendiks. Kjøreoppsett og utskrift for Habermans program.

```
2   8   1   7   0   0   1   0
(8F4.0)
(8F6.3,/,8F6.3)
ENKEL LATENT MODEL MED TO KLASSER FOR 2**3 TATELL.
   TU.LØ  ARB.ST ENS.BEV      LAT   TU.LØ  ARB.ST. ENS.BEV
 101  57  13  21  48  34  91  251
  1.    1.    1.    1.    1.    1.    1.    1.
 -1.   -1.   -1.   -1.   -1.   -1.   -1.   -1.
  1.    1.    1     1.   -1.   -1.   -1.   -1.
  1.    1.    1     1.   -1.   -1.   -1.   -1.
  1.    1.   -1.   -1.    1.    1.   -1.   -1.
  1.    1.   -1.   -1.    1.    1.   -1.   -1.
  1.   -1.    1.   -1.    1.   -1.    1.   -1.
  1.   -1.    1.   -1.    1.   -1.    1.   -1.
  1.   -1.    1.   -1.    1.   -1.    1.   -1.
 -1.    1.   -1.    1.   -1.    1.   -1.    1.
  1.   -1.    1.   -1.   -1.    1.   -1.    1.
  1.   -1.    1.   -1.   -1.    1.   -1.    1.
  1.   -1.   -1.    1.    1.   -1.   -1.    1.
  1.   -1.   -1.    1.    1.   -1.   -1.    1.
100.80  0.20 56.37  0.63  6.96   6.04  3.89 17.11
 44.82  3.18 25.06  8.94  3.10  87.90  1.73249.27
```

Kommentarer.

Linje 1- 6. Se beskrivelse i Haberman (1979).

linje 7-20. Dette er matrisen gjengitt i avsnitt 6 transponert.

linje 21-23. Initialverdier for de tilpassede hyppigheter.

1 ENKEL LATENT MODEL MED TO KLASSER FOR 2**3 TATELL.
OQOEFFICIENTS AND STANDARD ERRORS

TU.LØ	ARB.ST	ENS.BEV	LAT	TU.LØ	ARB.ST	ENS.BEV
-0.46751	-0.16529	-0.11541	0.29544	0.87220	1.50098	0.40576
0.11534	0.36482	0.05513	0.47225	0.10826	0.33519	0.05358

LIKELIHOOS RATID CHI SQUARE = -0.00013
PEARSON CHI SQUARE = 0.00000
NUMBER OF DEGREES OF FREEDOM = 0

J	K	OBSERVED COUNT	EXPECTED COUNT	ADJUSTED RESIDUAL
1	1	101.000	101.000	0.000
2	1	57.000	57.000	0.000
3	1	13.000	13.000	0.000
4	1	21.000	21.000	0.000
5	1	48.000	48.000	0.000
6	1	34.000	34.000	0.000
7	1	91.000	91.000	0.000
8	1	251.000	251.000	0.000