ADVANCEMENTS IN LATENT CLASS ANALYSIS: A PRACTICAL ORIENTATION

Jacques Hagenaars
Marcel Croon
Jeroen Vermunt

Department of Methodology/WORC
Faculty of Social Sciences
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands

email jacques.a.hagenaars@kub.nl
phone +31.13.4662086

The Sixth Annual Lecture in Social Statistics
ISP0M, K.U. Brussels
June 8-11, 1999

LATENT VARIABLE MODELS

Latent $\Rightarrow$ Not (directly) Observed

- **true value, scaling model, latent structure model**
- **clustering, underlying typologies**
- **unobserved heterogeneity, random coefficient models, multi-level models**
- **random and systematic misclassifications and errors**
- **missing data**

LATENT STRUCTURE MODEL

NOTATION

W, X, Y, Z: latent variables

A, B, C, D: manifest variables

BASIC MODEL

	A	e_A
	B	e_B
X		
	C	e_C
	D	e_D

- **latent and manifest variables**
- **local independence**
- **error term, probabilistic relation latent-manifest**

TYPOLOGY OF LSA

		LATENT		
		nominal	*ordinal*	*interval*
	nominal	**lca**		**ltr.**
MANIFEST	*ordinal*		**olca**	
	interval	**lprof**		**fa**

in addition: continuous versus discrete

estimated versus fixed scores

link between latent/manifest

‘general’ (categorical) lca: in all cells

BASIC LATENT CLASS ANALYSIS

Essential responsibility of government towards (1 = yes 2 = no):

A - equal rights men/women

B - good education

X

C - good medical care

D - equal rights guest workers

ABCD		ABCD		ABCD		ABCD	
1 1 1 1	*59*	1 2 1 1	*7*	2 1 1 1	*75*	2 2 1 1	*8*
1 1 1 2	*56*	1 2 1 2	*15*	2 1 1 2	*161*	2 2 1 2	*68*
1 1 2 1	*14*	1 2 2 1	*4*	2 1 2 1	*22*	2 2 2 1	*22*
1 1 2 2	*36*	1 2 2 2	*23*	2 1 2 2	*115*	2 2 2 2	*145*

N = 808

X is categorical and unobserved; one has to determine number of latent categories (classes)

A - equal rights men/women

B - good education

X

C - good medical care

D - equal rights guest workers

$$\pi_{abcdx}^{ABCDX} = \pi_{x}^{X}\pi_{ax}^{A|X}\pi_{bx}^{B|X}\pi_{cx}^{C|X}\pi_{dx}^{D|X}$$

$$\pi_{abcd}^{ABCD} = \sum_{x} \pi_{abcdx}^{ABCDX}$$

$$\pi_{abcdx}^{ABCDX} = \eta\tau_{a}^{A}\tau_{b}^{B}\tau_{c}^{C}\tau_{d}^{D}\tau_{x}^{X}\tau_{ax}^{AX}\tau_{bx}^{BX}\tau_{cx}^{CX}\tau_{dx}^{DX}$$

$$\ln\pi_{abcdx}^{ABCDX} = \phi + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_d^D + \lambda_x^X + \lambda_{ax}^{AX} + \lambda_{bx}^{BX} + \lambda_{cx}^{CX} + \lambda_{dx}^{DX}$$

{AX,BX,CX,DX}

A- equal rights men/women
B - good education
C- good medical care
D- equal rights guest workers

X		A		B		C		D	
		1	2	1	2	1	2	1	2
1	.41	.40	.60	.95	.05	.85	.15	.46	.54
2	.59	.17	.83	.47	.53	.35	.65	.12	.88

$\lambda_{11}^{AX} = 0{,}303$ $\lambda_{11}^{BX} = 0{,}772$ $\lambda_{11}^{CX} = 0{,}590$ $\lambda_{11}^{DX} = 0{,}464$

$L^2 = 13.99$ df = 6 p = .03
$X^2 = 13.97$

- **INTERPRETATION**

- **TESTING ‘usual’ chi-square statistics (L^2, X^2,** BIC); also $L^2_{r/u}$ (NOT for X=2 vs X=3)

- **MODIFYING MODELS** inspect residuals; inspect 'loadings'; more latent classes; more latent variables

ESTIMATION

- Maximum Likelihood Estimates; (product)multinomial sampling

- Table ABCDX is not observed (X is latent)

- EM, NR, etc.

- EM: - initial estimates $\hat{\pi}^{ABCDX}_{abcdx}$

  - E-step: $\hat{p}^{ABCDX}_{abcdx} = \hat{\pi}^{X|ABCD}_{x\,abcd} p^{ABCD}_{abcd}$

  - M-step: 'fit' appropriate model (AX,BX,CX,DX} to $\hat{p}$

  - repeat E- and M-step, using latest 'updates' untill convergence

PROBLEMS: - local/global maxima

- ‘terminal’ estimates (boundary -)

- identifiability: necessary and sufficient conditions

RESTRICTED MODELS

$$\Pi_x^X = c$$

$$\pi_{ax}^{A|X} = 1 \quad = 0 \quad = c$$

$$\pi_{11}^{A|X} = \pi_{22}^{A|X}$$

$$\pi_{ix}^{A|X} = \pi_{ix}^{B|X}$$

$$\tau_{ix}^{AX} = \tau_{ix}^{BX}$$

linear/ordinal restrictions on λ_{ax}^{AX}

TWO LATENT VARIABLE MODEL

Y	**A - equal rights m/w**
	D - equal rights g.w.
Z	**B - good education**
	C - good medical care

$$\lambda_{11}^{YZ} = 1.136$$

$$\lambda_{11}^{AY} = 0{,}393 \quad \lambda_{11}^{DY} = 0{,}663$$

$$\lambda_{11}^{BY} = 0{,}776 \quad \lambda_{11}^{CY} = 0{,}619$$

YZ		**A**		**B**		**C**		**D**	
		1	**2**	**1**	**2**	**1**	**2**	**1**	**2**
1 1	***.26***	***.51***	***.49***	***.95***	***.05***	***.85***	***.15***	***.65***	***.35***
1 2	***.01***	***.51***	***.49***	***.45***	***.55***	***.33***	***.67***	***.65***	***.35***
2 1	***.18***	***.18***	***.82***	***.95***	***.05***	***.85***	***.15***	***.12***	***.88***
2 2	***.56***	***.18***	***.82***	***.45***	***.55***	***.33***	***.67***	***.12***	***.88***

$L^2 = 5.76$ df = 4 p = .22
$X^2 = 5.75$

LATENT (CLASS) SCORES

Determine modal latent class x*, given observed scores

$$\pi_{x*\,abcd}^{X|ABCD}$$

Assign all respondents to their modal class

Error measures

$$\epsilon_{abcd}^{ABCD} = 1 - \pi_{x\,abcd}^{X|ABCD}$$

$$E = \sum_{a,b,c,d} (\pi_{abcd}^{ABCD} \epsilon_{abcd}^{ABCD})$$

$$\lambda_{X|ABCD} = [(1 - \pi_{x*}^{X}) - E] / (1 - \pi_{x*}^{X})$$

Identifiability of latent scores
Biased latent scores

COMPARATIVE ANALYSES

Nation, Time, Group are just (independent) variables

Main Issue: Comparability of Measurement Model

Relevant Models

{GAX,GBX,GCX,GDX}

{GX,AX,BX,CX,DX,GA,GB,GC,GD}

{GX,AX,BX,CX,DX}

{G,AX,BX,CX,DX}

Differences between parametrizations in terms of the loglinear parameters or the response probabilities