Statistical Methods, part 1

Module 2: Latent Class Analysis of Survey Error

LCA

Dan Hedlin
Stockholm University
November 2012

Acknowledgement: **Paul Biemer**

Stockholm University

Latent Variable Analysis

	Latent Variable	Manifest Variable
Factor analysis	Continuous	Continuous
Latent trait analysis	Continuous	Discrete
Latent profile analysis	Discrete	Continuous
Latent class analysis	Discrete	Discrete

Bartholomew, D.J., and Knott, M. (1999). *Latent Variable Models and Factor Analysis*. London: Arnold.

Latent Class Analysis (LCA)

- Developed in 1950's
  - Lazarsfeld, Henry, Green
- Outcome variables can be dichotomous or polytomous (nominal or ordinal)
- Very extensive field
- Advanced use requires knowledge of log-linear models with and without latent variables

Some applications of LCA

Nonsampling Error Analysis

- Identifying flawed questions and other questionnaire problems
- Estimating census undercount in a capture-recapture framework
- Characterizing respondents, interviewers, and questionnaire elements that contribute to survey error
- Adjusting for nonresponse and missing data in surveys
- We will only touch upon LCA applications

Some Other Applications *(cont'd)*

Social Science Analysis

- Causal modeling
- Log-linear analysis compensating for measurement error
- Cluster analysis
- Variable reduction and scale construction

Why in this course?

- Very useful class of models
- Few people seem to know about LCA
- Course plan indicates a content of science and models. LCA does involve scientific issues.

Use of LCA for Investigating Survey Error

- LCA methods and models are prone to misuse
  - E.g., adjustment vs. evaluation
  - Lack of attention to assumptions
  - Weak indicators
- LCA is best use in conjunction with other methods
- Avoid the temptation to believe the model is true
  - Consider the risks if the model is not true
  - Take steps to verify the model assumptions and to validate the estimates

Maximum Likelihood Estimation

Recall that $\hat{\pi}, \hat{\theta}$, and $\hat{\phi}$ contain all the information we need to estimate:

- π, the true population proportion
- *SV*, the sampling variance
- *SRV*, the simple response variance,
- *R*, the reliability ratio
- Measurement bias of the sample proportion, Bias(*p*)
- Total variance, Var(*p*)
- Total MSE(*p*)

We can estimate $\hat{\pi}, \hat{\theta}$, and $\hat{\phi}$ in some cases using only information from remeasurements of the characteristics of interest.

Latent Class Models for Measurement Error

Consider the following:

		Reinterview (y_2)	
		1	2
Interview (y_1)	1	n_{11}	n_{12}
	2	n_{21}	n_{22}

Or more generally: any second measurement that is similar (or next identical in type) to first measurement (i.e. not necessarily a better measurement)

Slight Change in Notation

y_{tr} denotes observation t on unit r

$y_{tr} = 1$ denotes positive response and

$= 2$ denotes a negative response

Assume:

- SRS (simple random sample)
- Two parallel measurements are available
- $\gamma_{\theta\phi} = 0$ (homogeneity, ie constant probabilities for misclassification)

$$\Pr(n_{11}, n_{12}, n_{21}, n_{22}) \text{ is a multinomial}$$

Multinomial Distribution

Consider the distribution of the 2×2 table with cells (11, 12, 21, 22)

Let π_{11} = P(y_i falls in cell 11) and define similarly.

$$\pi_{12}, \pi_{21}, \pi_{22}$$

Then,

$$\mathrm{P}(n_{11}, n_{12}, n_{21}, n_{22}) = \binom{n}{n_{11}\ n_{12}\ n_{21}\ n_{22}} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{21}^{n_{21}} \pi_{22}^{n_{22}}$$

For the interview-reinterview table,

$$\pi_{11} = P(y_{1r} = 1, y_{2r} = 1)$$
$$\pi_{21} = P(y_{1r} = 2, y_{2r} = 1)$$
$$\pi_{12} = P(y_{1r} = 1, y_{2r} = 2)$$
$$\pi_{22} = P(y_{1r} = 2, y_{2r} = 2)$$

$$\pi_{11} = P(y_{1r} = 1,\ y_{2r} = 1 \mid \mu_r = 1) \times P(\mu_r = 1)$$

$$+\ P(y_{1r} = 1,\ y_{2r} = 1 \mid \mu_r = 2) \times P(\mu_r = 2)$$

$$= (1-\theta)(1-\theta)\pi + \phi\phi(1-\pi)$$

$$\boxed{= (1-\theta)^2\pi + \phi^2(1-\pi)}$$

$$\pi_{21} = P(y_{1r} = 2,\ y_{2r} = 1 \mid \mu_r = 1) \times P(\mu_r = 1)$$
$$+\ P(y_{1r} = 2,\ y_{2r} = 1 \mid \mu_r = 2) \times P(\mu_r = 2)$$
$$= \theta(1-\theta)\pi + (1-\phi)\phi(1-\pi)$$

By assumption: $\pi_{12} = \pi_{21}$

$$\pi_{22} = P(y_{1r} = 2,\ y_{2r} = 2 \mid \mu_r = 1) \times P(\mu_r = 1)$$
$$+\ P(y_{1r} = 2,\ y_{2r} = 2 \mid \mu_r = 2) \times P(\mu_r = 2)$$
$$= \theta^2\pi + (1-\phi)^2(1-\pi)$$

Summarising:

$$\text{Likelihood} = \mathrm{L}(\pi, \theta, \phi \mid n_{11}, n_{12}, n_{21}, n_{22})$$

$$\mathrm{P}(n_{11}, n_{12}, n_{21}, n_{22}) = \binom{n}{n_{11}\ n_{12}\ n_{21}\ n_{22}} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{21}^{n_{21}} \pi_{22}^{n_{22}}$$

$$= \binom{n}{n_{11}\ n_{12}\ n_{21}\ n_{22}} \left[\begin{array}{l} [(1-\theta)^2 \pi + \phi^2 (1-\pi)]^{n_{11}} \\ \times [\theta(1-\theta)\pi + \phi(1-\phi)(1-\pi)]^{n_{12}+n_{21}} \\ \times [\theta^2 \pi + (1-\theta)^2 (1-\pi)]^{n_{22}} \end{array} \right]$$

Maximum Likelihood Estimation

The best (ML) estimates of

$$\pi,\ \theta,\ \phi$$

are those that maximize *L* as a function of $\pi,\ \theta,\ \phi$
Unfortunately, no unique solution exists (ie model is not identifiable).

3 parameters
2 degrees of freedom (d.f.)

Why only 2 d.f. and not 3 since it is a 2x2 table?

param > # d.f. and model is not identifiable.

Suppose we let

Notation for common probability

$$\theta = \phi = \varepsilon, \text{ say}$$

[This is not realistic. Why not? Nevertheless…]

Then, plugging into the previous formula,

$$L(n_{11}, n_{12}, n_{21}, n_{22} | \pi, \varepsilon)$$

$$= \text{constant} \, [(1-\varepsilon)^2 \pi + \varepsilon^2 (1-\pi)]^{n_{11}} \times [\varepsilon(1-\varepsilon)]^{n_{12}+n_{21}}$$

$$\times [\varepsilon^2 \pi + (1-\varepsilon)^2 (1-\pi)]^{n_{22}}$$

Two parameters π, ε and two d.f.,

so now $\Rightarrow$ π, ε estimable

But, no d.f. left over to estimate model fit (no chi-square test).
Solution fits the data perfectly

==> model is saturated.

No d.f. test model validity or fit.

Example

		Reinterview	
		1	2
Interview	1	1000	50
	2	100	500

Find MLE of π, ε

Numerical methods to solve equations, e.g.:

- grid search
- Newton-Raphson
- Iterative Proportional Fitting (IPF)
- EM-algorithm

More General Model (Hui-Walter, 1980)

Interview π, $\theta(\text{int.}), \phi(\text{int.})$
$\theta(\text{reint.}), \theta(\text{reint.})$

Assumes that error distributions for interview and reinterview are not equal.

As before, we assume

- Single latent variable
- Local independence (we do not assume parallel measurements)

Need at least 5 d.f. to estimate these five parameters:

$$\pi, \theta(\text{int.}), \phi(\text{int.}), \theta(\text{reint.}), \phi(\text{reint.})$$

Trick: Introduce a Bivariate Grouping Variable, *G*

Group 1 (*G*=1)

		Reinterview	
		1	2
Interview	1	n_{111}	n_{112}
	2	n_{121}	n_{122}

Group 2 (*G*=2)

		Reinterview	
		1	2
Interview	1	n_{211}	n_{212}
	2	n_{221}	n_{222}

How many d.f. are there?

3 from G=1 table

3 from G=2 table

==> up to 6 parameters can be estimated

10 Parameters in the unrestricted model

Group 1 π_1, $\theta_1(\text{int.})$ $\theta_1(\text{reint.})$ $\phi_1(\text{int.})$ $\phi_1(\text{reint.})$

Group 2 π_2, $\theta_2(\text{int.})$ $\theta_2(\text{reint.})$ $\phi_2(\text{int.})$ $\phi_2(\text{reint.})$

Need to constrain the model to save 4 df's

Hui-Walter constraints

$$\theta_1(\text{int.}) = \theta_2(\text{int.}) = \theta_{\text{int}}$$
$$\phi_1(\text{int.}) = \phi_2(\text{int.}) = \phi_{\text{int}}$$
$$\theta_1(\text{reint.}) = \theta_2(\text{reint.}) = \theta_{\text{reint}}$$
$$\phi_1(\text{reint.}) = \phi_2(\text{reint.}) = \phi_{\text{reint}}$$

In words, assume classification probabilities do not differ by group.

Now, 6 parameters and 6 d.f. so model can be estimated

Completely saturated, i.e., 0 d.f. for error

Likelihood

$$\text{Likelihood} = \text{P}(\pi_1, \pi_2 \theta_{\text{int}}, \phi_{\text{int}}, \theta_{\text{reint}}, \phi_{\text{reint}} \mid n_{gij}, g = 1,2,\ i = 1,2,\ j = 1,2)$$

$$\propto (\pi_{111})^{n_{111}} (\pi_{121})^{n_{121}} (\pi_{112})^{n_{112}} (\pi_{122})^{n_{122}} (\pi_{211})^{n_{211}} (\pi_{221})^{n_{221}} (\pi_{212})^{n_{212}} (\pi_{222})^{n_{222}}$$

$$= \prod_{g} \prod_{i} \prod_{j} (\pi_{gij})^{n_{gij}}$$

where

$$\pi_{gij} = \text{P}(G = g, y_1 = i, y_2 = j)$$

Likelihood

These probabilities can be rewritten in terms of $\pi_1, \pi_2 \theta_{\text{int}}, \phi_{\text{int}}, \theta_{\text{reint}}, \phi_{\text{reint}}$

$$\pi_{111} = \pi_{g=1}\left[(1-\theta_{\text{int}})(1-\theta_{\text{reint}})\pi_1 + \phi_{\text{int}}\phi_{\text{reint}}(1-\pi_1)\right]$$

$$\pi_{211} = \pi_{g=2}\left[(1-\theta_{\text{int}})(1-\theta_{\text{reint}})\pi_2 + \phi_{\text{int}}\phi_{\text{reint}}(1-\pi_2)\right]$$

and so on.

Plug these into L and maximize to obtain ML estimates

$$\hat{\pi}_1, \hat{\pi}_2 \hat{\theta}_{\text{int}}, \hat{\phi}_{\text{int}}, \hat{\theta}_{\text{reint}}, \hat{\phi}_{\text{reint}}$$

Results for 1996 US Current Population Survey Reinterview

Input Data

Males				Females			
Interview Response	Reinterview Response			Interview Response	Reinterview Response		
	EMP	UNEMP	NLF		EMP	UNEMP	NLF
EMP	2372	14	29	EMP	2087	6	60
UNEMP	10	90	27	UNEMP	10	75	41
NLF	75	18	974	NLF	87	33	1639

Are the Hui-Walter assumptions plausible for these data?

Hui-Walter Model Estimates of Classification Probabilities, fictitious numbers

Interview			
True Status	Observed Status		
	EMP	UNEMP	NLF
EMP	98.6 (0.1)	1.4 (0.1)	0.0 (n/a)
UNEMP	5.6 (15.2)	61.6 (11.1)	27.9 (5.3)
NLF	2.6 (1.5)	0.0 (n/a)	97.4 (1.1)

Standard errors in brackets

Now: Three Indicator LC Model

DATA

	C=1		C=2	
	B=1	B=2	B=1	B=2
A=1	n_{111}	n_{121}	n_{112}	n_{122}
A=2	n_{211}	n_{221}	n_{212}	n_{222}

Classical Latent Class Model

- Requires a minimum of three indicators of *X* for unrestricted model
  - Say, A, B, C
- Assumes local independence
  - i.e., P(A and B and C|X) = P(A|X)P(B|X)P(C|X)
- Notation

$$\pi_{X=x} \text{ or } \pi_x^X = P(X = x),$$

$$\pi_{A=a|X=x} \text{ or } \pi_{a|x}^{A|X} = P(A = a \mid X = x)$$

$$\text{Also, } \pi_x^X, \ \pi_{abc}^{ABC}$$

Interpretations of Local Independence

$$\pi_{abc|x}^{ABC|X} = \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X}$$

$\Rightarrow$ classification errors are independent

i.e., A, B, C represent independent selections from an individual's response distribution

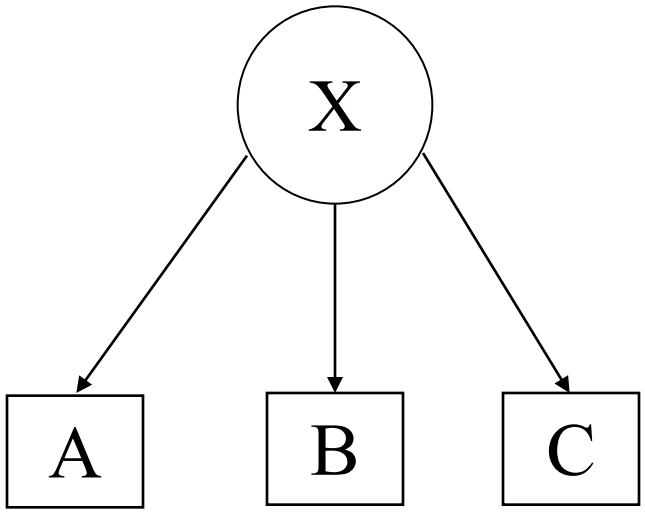
We can refer to this model as $\{A|X\, B|X\, C|X\}$ or $\{AX\, BX\, CX\}$ (borrowed from log-linear modeling)

Latent Class Model for Three Indicators

$$L(\boldsymbol{\pi} \mid \mathbf{n}) \approx \prod_{a}\prod_{b}\prod_{c}(\pi_{abc}^{ABC})^{n_{abc}}$$

$$\pi_{abc}^{ABC} = \sum_{x=1}^{2}\pi_{x}^{X}\pi_{abc|x}^{ABC|X} = \sum_{x=1}^{2}\pi_{x}^{X}\pi_{a|x}^{A|X}\pi_{b|x}^{B|X}\pi_{c|x}^{C|X}$$

Path Diagram



What *is not* shown is as important as what *is* shown

MLE Methods of Estimation

- π-probability model
  - Our focus in this course
- log-linear model with latent variable
- modified path model

The latter two methods require knowledge of log-linear models which is not assumed in this course

LCA Software include:

- Free software
  - ℓEM
  - "SAS" by PennState University
- Commercial software
  - Mplus
  - Latent Gold

Others: http://www.john-uebersax.com/stat/soft.htm

Design of the NHSDA

- National, multistage, household survey
- 1994, 1995, 1996 data
- 43,825 total interviews
- data are weighted
- drug questions are repeated
- See Biemer and Wiesen (2002)

Three Indicators of Past-Year Marijuana Use

The Recency Question (Indicator A)

How long has it been since you last used marijuana or hashish?

A = 1 if either “Within the past 30 days” or “More than 30 days but within past 12 months”

A = 2 if otherwise

Three Indicators of Past-Year Marijuana Use

The Frequency Question (Indicator B)

Now think about the past 12 months from your 12-month reference date through today. On how many days in the past 12 months did you use marijuana or hashish?

B = 1 if response is 1 or more days;

B = 2 if otherwise

Three Indicators of Past-Year Marijuana Use

The Composite Question (Indicator C)

7 questions such as

- used in last 12 months?
- spent a great deal of time getting it, using it, or getting over its effects?
- used drug much more often or in larger amounts than intended?

C = 1 if response is positive to any question;

C = 2 if otherwise

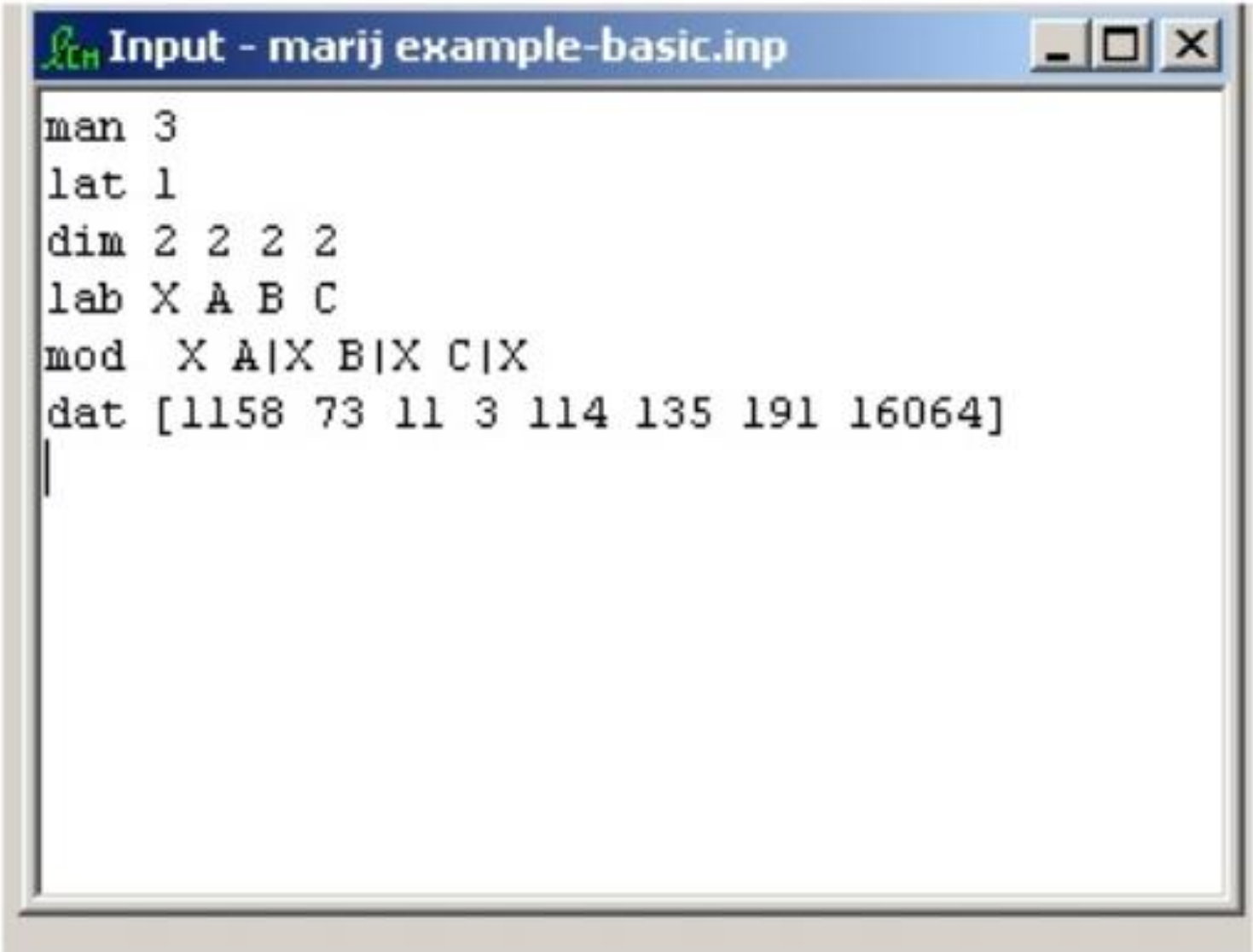
Evidence of Reporting Error: Inconsistency in A, B, and C

Comparison	1994	1995	1996
A vs. B	1.35	1.48	1.61
A vs. C	4.80	2.14	2.48
B vs. C	4.96	2.31	2.69
A vs. B vs. C	5.55	2.96	3.39

1995 Marijuana Data

	C=1		C=2	
	B=1	B=2	B=1	B=2
A=1	1158	11	73	3
A=2	114	191	135	16064

Introduction to ℓEM



Introduction to ℓEM (cont'd)

```
man 3
lat 1
dim 2 2 2 2
lab X A B C
mod  X A|X B|X C|X
dat [1158 73 11 3 114 135 191
  16064]
```

*Number of manifest variables

*Number of latent variables

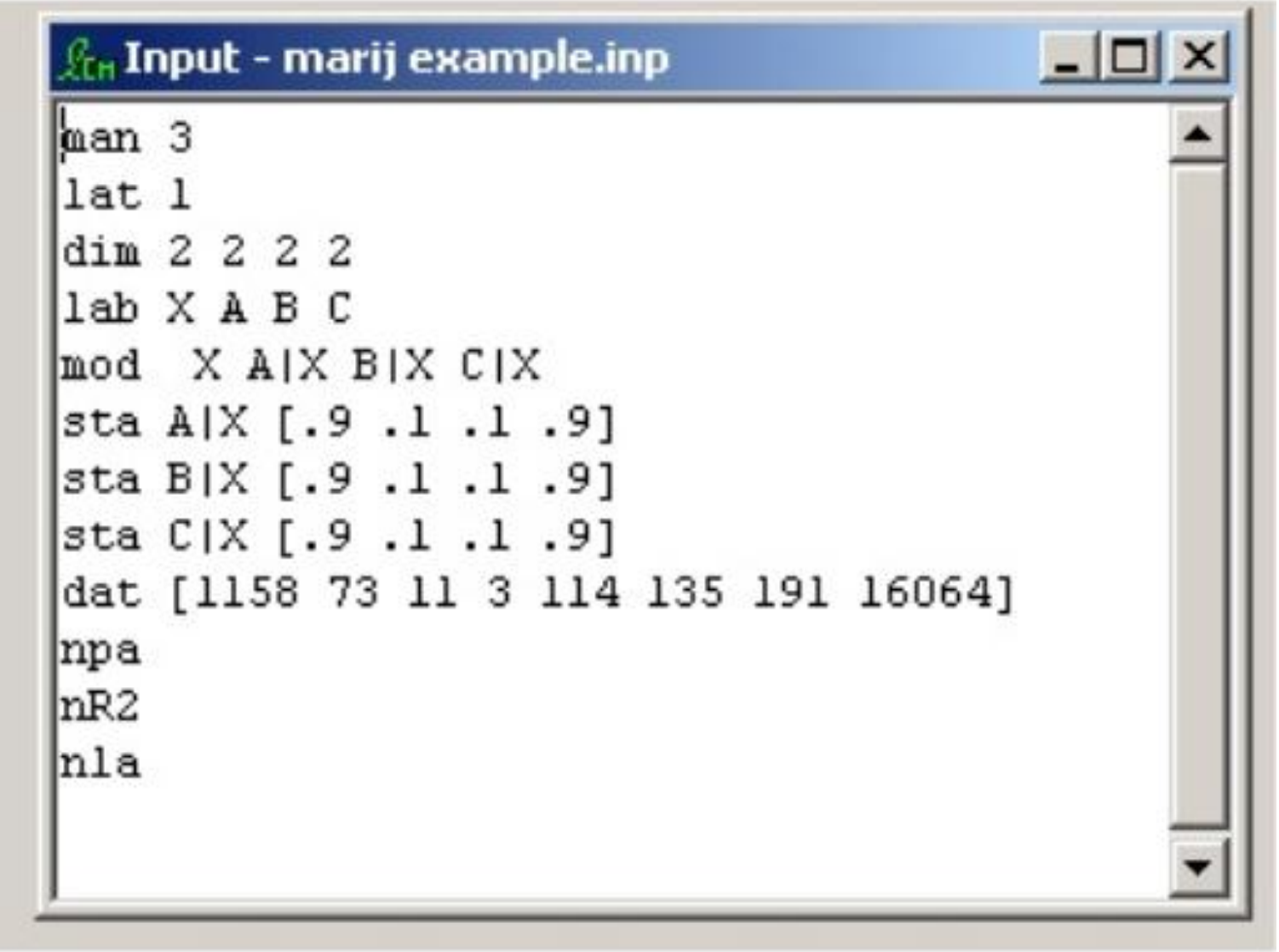
*Dimensions

*Labels

*Model

*Data

Introduction to ℓEM (cont'd)



Introduction to ℓEM (cont'd)

```
man 3
lat 1
dim 2 2 2 2
lab X A B C
mod  X A|X B|X C|X
sta A|X [.9 .1 .1 .9]
sta B|X [.9 .1 .1 .9]
sta C|X [.9 .1 .1 .9]
dat [1158 73 11 3 114 135 191 16064]
npa
nR2
nla
```

*Number of manifest variables
*Number of latent variables
*Dimensions
*Labels
*Model
*Starting values for A|X
*Starting values for B|X
*Starting values for C|X
*Data
*Output control
*Output control
*Output control

ℓEM Output (cont'd)

```
LEM: log-linear and event history analysis with missing data.
Developed by Jeroen Vermunt (c), Tilburg University, The
Netherlands.
Version 1.0b (September 18, 1997).

*** INPUT ***

  man 3
  lat 1
  dim 2 2 2 2
  lab X A B C
  mod  X A|X B|X C|X
  sta A|X [.9 .1 .1 .9]
  sta B|X [.9 .1 .1 .9]
  sta C|X [.9 .1 .1 .9]
  dat [1158 73 11 3 114 135 191 16064]
  npa
  nR2
  nla
```

ℓEM Output (cont'd)

```
*** STATISTICS ***

  Number of iterations = 8
  Converge criterion   = 0.0000008181
  Seed random values   = 5307

  X-squared            = 0.0000 (0.0000)
  L-squared            = 0.0000 (0.0000)
  Cressie-Read         = 0.0000 (0.0000)
  Dissimilarity index  = 0.0000
  Degrees of freedom   = 0
  Log-likelihood       = -7371.28926
  Number of parameters = 7 (+1)
  Sample size          = 17749.0
  BIC(L-squared)       = 0.0000
  AIC(L-squared)       = 0.0000
  BIC(log-likelihood)  = 14811.0671
  AIC(log-likelihood)  = 14756.5785

WARNING: no information is provided on identification of
parameters
```

ℓEM Output (cont'd)

```
*** FREQUENCIES ***

  A B C      observed  estimated  std. res.
  1 1 1     1158.000   1157.999      0.000
  1 1 2       73.000     73.001     -0.000
  1 2 1       11.000     11.001     -0.000
  1 2 2        3.000      2.999      0.000
  2 1 1      114.000    114.001     -0.000
  2 1 2      135.000    134.999      0.000
  2 2 1      191.000    191.000      0.000
  2 2 2    16064.000  16064.000     -0.000
```

ℓEM Output (cont'd)

```
*** (CONDITIONAL) PROBABILITIES ***

* P(X) *

  1                0.0768
  2                0.9232

* P(A|X) *

  1 | 1            0.9115
  2 | 1            0.0885
  1 | 2            0.0001
  2 | 2            0.9999
```

ℓEM Output (cont'd)

```
* P(B|X) *

  1 | 1          0.9906
  2 | 1          0.0094
  1 | 2          0.0079
  2 | 2          0.9921

* P(C|X) *

  1 | 1          0.9407
  2 | 1          0.0593
  1 | 2          0.0117
  2 | 2          0.9883
```

Fit Statistics

- X-squared
- L-squared
- Dissimilarity index
- Degrees of freedom
- Log-likelihood
- Number of parameters
- Sample size
- BIC(L-squared)
- AIC(L-squared)

Pearson Statistic

$$X^2 = \sum_{abc} \frac{(n_{abc} - \hat{m}_{abc})^2}{\hat{m}_{abc}}$$

$\hat{m}_{abc} =$ model estimated frequency in cell (a, b, c)

Distributed approximately as a chi-square random variable if the model is true. Poor approximation if average cell size is less than 5.

Likelihood Ratio Chi-square Statistic

$$L^2 = 2\sum_{abc} n_{abc} \log\left(\frac{n_{abc}}{\hat{m}_{abc}}\right)$$

Distributed approximately as a chi-square random variable if the model is true. Poor approximation if average cell size is less than 5.

Dissimilarity Index

$$D = 2\sum_{abc} \frac{| n_{abc} - \hat{m}_{abc} |}{2n}$$

Smallest proportion of observations
that would need to be reallocated to other cells
to make the model fit perfectly.
Should be less than 0.05 for a well-fitting
model.

Degrees of Freedom

df = number of cells -

number of estimated model parameters

Log-Likelihood

$$\log(L) = \sum_{abc} n_{abc} \log(\hat{\pi}_{abc})$$

BIC, *AIC*, *BIC-L*^{2} and *AIC-L*^{2}

$$BIC = -2\log L + (\log n) \times npar$$

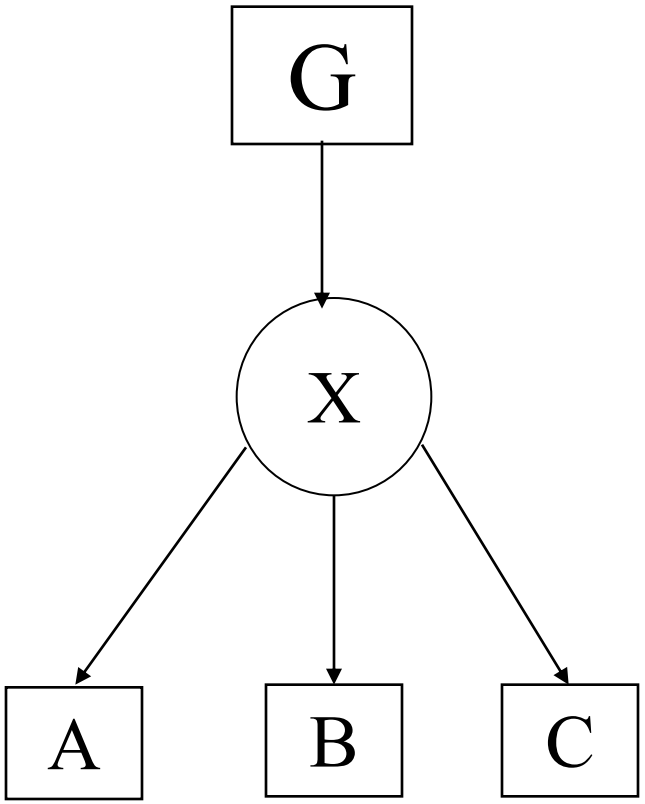
$$AIC = -2\log L + 2npar$$

$$BIC(L^2) = L^2 - df \log(n)$$

$$AIC(L^2) = L^2 - df\ 2$$

Three Indicator Model with a Grouping Variable

Model: GX A|X B|X C|X

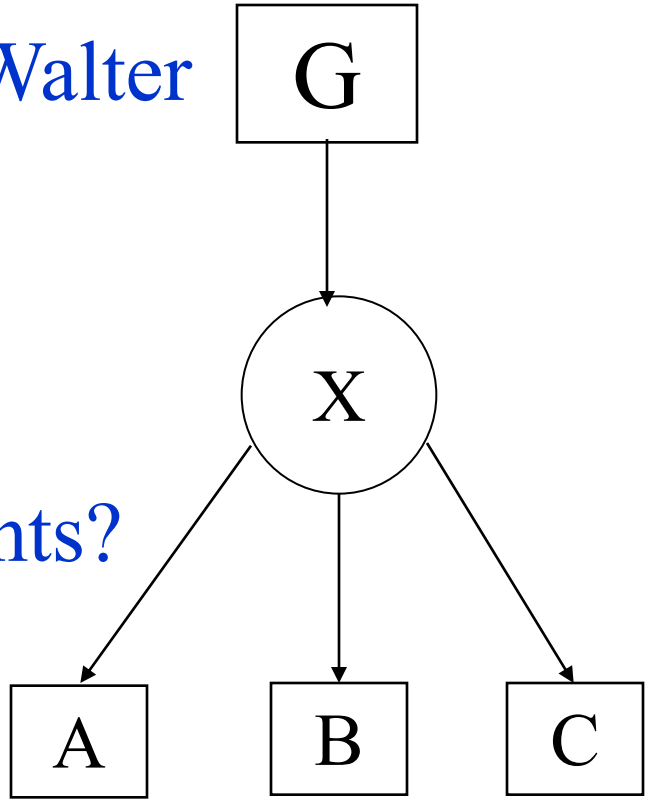


Three Indicator Model with a Grouping Variable

Model: GX A|X B|X C|X

This is the Hui-Walter model for three indicators.

Do you see the implied constraints?



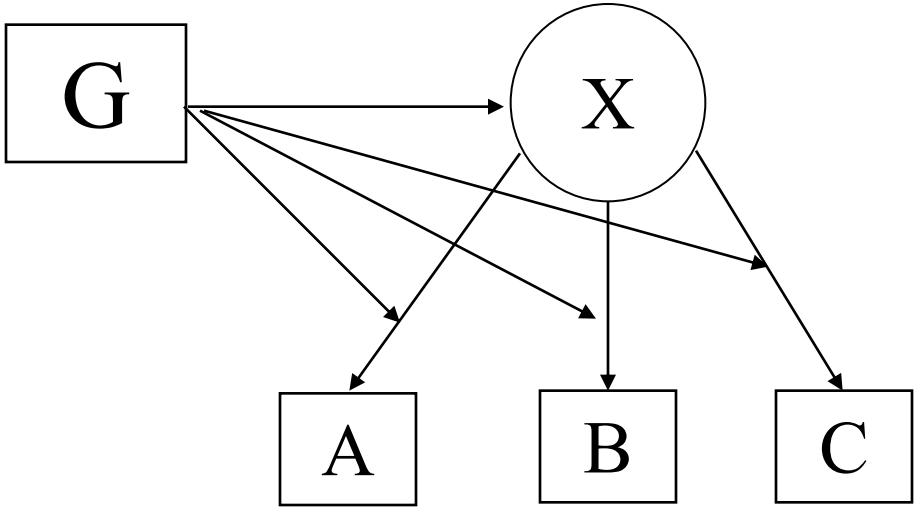
ℓEM Input Code

```
man 4          * 4 manifest variables
lat 1          * 1 latent variable
dim 2 2 2 2 2
lab X S A B C * S=1(Males)
                   =2(Females)
mod  SX A|X B|X C|X
rec 16    * 16 records in data set
rco       * last column is a count
sta A|X [.9 .1 .1 .9]
sta B|X [.9 .1 .1 .9]
sta C|X [.9 .1 .1 .9]
```

ℓEM Input Code

```
dat [                                   *  Data ordering is
1    1       1       1       698        *  Sex    A    B    C    Count
1    1       1       2
1    1       2       1       5
1    1       2       2       2
1    2       1       1       65
1    2       1       2       83
1    2       2       1       121
1    2       2       2       7495
2    1       1       1       460
2    1       1       2       30
2    1       2       1       6
2    1       2       2       1
2    2       1       1       49
2    2       1       2       52
2    2       2       1       70
2    2       2       2       8569]
npa
nR2
nla
```

Three Indicator Model with a Grouping Variable – Fully Saturated



Model: GX A|XG B|XG C|XG

Three Indicator Model with a Grouping Variable – Fully Saturated

$$\pi_{gabc}^{GABC} = \sum_{x=1}^{2} \pi_{gx}^{GX} \pi_{a|xg}^{A|XG} \pi_{b|xg}^{B|XG} \pi_{c|xg}^{C|XG}$$

ℓEM Input Code

```
man 4
lat 1
dim 2 2 2 2 2
lab X S A B C
mod  SX A|SX B|SX C|SX
rec 16
rco
sta A|SX [.9 .1 .9 .1 .1 .9 .1 .9]
sta B|SX [.9 .1 .9 .1 .1 .9 .1 .9]
sta C|SX [.9 .1 .9 .1 .1 .9 .1 .9]
```

Starting values for each group. Order consistent with `lab` statement

ℓEM Results for Grouping Variable Models

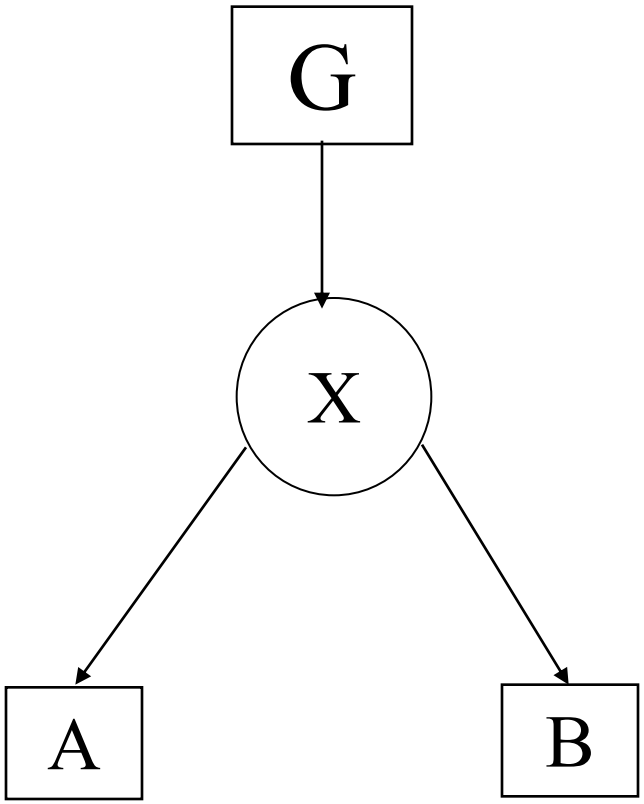
Model	d.f	L^2	p	D	BIC-L^2
X A|X B|X C|X	7	49.0	0.0	0.024	117.4799
SX A|X B|X C|X	6	33.2	0.0	0.006	-25.4787
SX A|XS B|XS C|XS	0	0	n/a	0	0

ℓEM Estimates for Model 2

```
*** (CONDITIONAL) PROBABILITIES ***

* P(XS) *

  1 1             0.0461
  1 2             0.0308
  2 1             0.4335
  2 2             0.4896

* P(A|X) *

  1 | 1           0.9110
  2 | 1           0.0890
  1 | 2           0.0001
  2 | 2           0.9999

* P(B|X) *

  1 | 1           0.9905
  2 | 1           0.0095
  1 | 2           0.0079
  2 | 2           0.9921

* P(C|X) *

  1 | 1           0.9402
  2 | 1           0.0598
  1 | 2           0.0117
  2 | 2           0.9883
```

Specifying Hui-Walter Model in ℓEM

Model: GX AX BX



Two Parameter, Two Indicator Model

```
lat 1
man 2
dim 2 2 2
lab X A B
mod X
      A|X eq2
      B|X eq2
des[0 1 1 0    *Set   A=2|X=1 = A=1|X=2
    0 1 1 0]   *    = B=2|X=1 = B=1|X=2
sta A|X [.9 .1 .1 .9]
sta B|X [.9 .1 .1 .9]
dat [1000 100 50 500]
```

A Model for Estimating Mode Effects

Study Design (Biemer, 2001)

Draw a sample and randomly divide. Assign one subsample to F-F; the other to CATI. Reinterview both by CATI.		Reinterview by CATI	
		B=1	B=2
Interview by Face to face	A=1	n_{F11}	n_{F12}
	A=2	n_{F21}	n_{F22}
Interview by CATI	A=1	n_{T11}	n_{T12}
	A=2	n_{T21}	n_{T22}

Model Assumptions

- *X* denotes (latent) true characteristic
- *G*=1 denotes F-F sample
- *G*=2 denotes CATI sample $\pi_{x|1}^{X|G} \neq \pi_{x|2}^{X|G}$
- CATI interview and CATI reinterview share a common mode effect; i.e.,

$$\pi_{2|12}^{A|XG} = \pi_{2|12}^{B|XG} = \pi_{2|11}^{B|XG} = \theta_T \text{ and}$$

$$\pi_{1|22}^{A|XG} = \pi_{1|22}^{B|XG} = \pi_{1|21}^{B|XG} = \phi_T$$

Summary of Model Parameters

$$\pi_g^G, \pi_{1|1}^{X|G}, \pi_{1|2}^{X|G}, \theta_T, \theta_F, \phi_T, \phi_F$$

Leaving 0 d.f. for testing fit

Does Anyone Smoke Inside the Home?

		Reinterview by CATI	
		R=1	R=2
Interview by Face to face	F=1	334	70
	F=2	29	1233
Interview by CATI	T=1	282	20
	T=2	9	931

ℓEM Input Code

```
* Example:  Does Anyone Smoke Inside the Home?
lat 1
man 3
dim 2 2 2 2
lab X G A B
mod G
    X|G
    A|GX eq2
    B|GX eq2
sta X|G [.3 .4 .7 .6]
des [0 1 0 2
     3 0 4 0
     0 2 0 2
     4 0 4 0]
dat [ 334 70 29 1233
      282 20  9  931]
```

ℓEM Input Code

```
* Example:  Does Anyone Smoke Inside the Home?
lat 1
man 3
dim 2 2 2 2
lab X G A B                  *    G denotes the split sample
mod G
    X|G                      *    P(X|G) varies across split
    A|GX eq2                 *    For G=1 A = ff and B=CATI
    B|GX eq2                 *    For G=2 A & B are both CATI
sta X|G [.3 .4 .7 .6]
des [0 1 0 2     *X=1 Int: Theta_F and Theta_T
     3 0 4 0     *X=2 Int: Phi_F and Phi_T
     0 2 0 2     *X=1 Reint: Theta_T
     4 0 4 0]    *X=2 Reint: Phi_T
dat [ 334 70 29 1233
       282 20  9  931]
```

ℓEM Fit Statistics

```
X-squared                = 4.1720 (0.0000)
L-squared                = 4.2787 (0.0000)
Cressie-Read             = 4.1952 (0.0000)
Dissimilarity index      = 0.0019
Degrees of freedom       = 0
Log-likelihood           = -4047.33498
Number of parameters     = 7 (+1)
Sample size              = 2908.0
BIC(L-squared)           = 4.2787
AIC(L-squared)           = 4.2787
BIC(log-likelihood)      = 8150.4965
AIC(log-likelihood)      = 8108.6700
```

ℓEM Estimates

```
* P(G) *
  1       0.5729
  2       0.4271

* P(X|G) *
  1 | 1   0.2288
  1 | 2   0.2507
  2 | 1   0.7712
  2 | 2   0.7493
```

```
* P(A|XG) *
  1 | 1 1   0.9207
  2 | 1 1   0.0793
  1 | 1 2   0.9516
  2 | 1 2   0.0484
  1 | 2 1   0.0413
  2 | 2 1   0.9587
  1 | 2 2   0.0002
  2 | 2 2   0.9998
```

```
* P(B|XG) *
  1 | 1 1   0.9516
  2 | 1 1   0.0484
  1 | 1 2   0.9516
  2 | 1 2   0.0484
  1 | 2 1   0.0002
  2 | 2 1   0.9998
  1 | 2 2   0.0002
  2 | 2 2   0.9998
```

Importance of Model Validity Depends on the Application

- In some applications, validity can be supported by ability to identify real questionnaire problems.
- In other applications, this type of validation may be quite difficult
- Further, LCA methodology is being pushed in the US to adjust the reported survey estimates for misclassification bias.
  - Unemployment rate
  - Expenditures
  - Total population size in a census

Analyzing Unequal Weighting and Clustered Samples ("Analysis of survey data")

- Option 1: Ignore the weights and clustering
  - Pros: Often these do not change estimates of classification error
  - Cons: can lead to false inference about classification error (see, for example, Patterson, Dayton and Graubard (2002))

Analyzing Unequal Weighting and Clustered Samples (cont'd)

- Option 2: Use software (Latent Gold, Mplus) that properly account for the weights
  - Pros: Unbiased estimates and asymptotically unbiased standard errors
  - Cons: Convergence problems in some cases