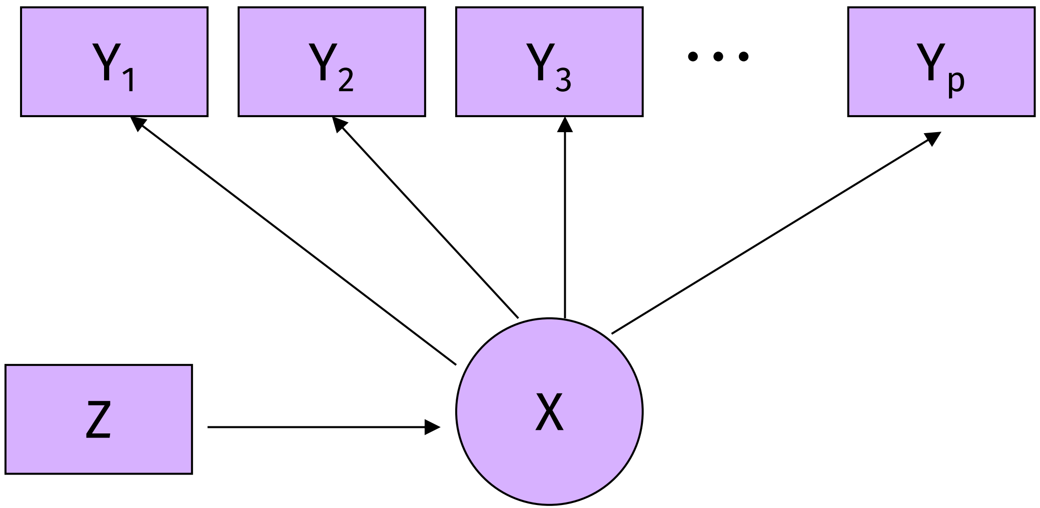
Latent class analysis

More latent class analysis

DL Oberski & L Boeschoten

The Latent Class Model



- Observed (continuous or) **categorical** Items
- **Categorical** Latent Class Variable (X)
- Continuous or Categorical Covariates (Z)

Small example: data from GSS 1987

Y1: "allow anti-religionists to speak" (1 = allowed, 2 = not allowed),
Y2: "allow anti-religionists to teach" (1 = allowed, 2 = not allowed),
Y3: "remove anti-religious books from the library" (1 = do not remove, 2 = remove).

Y1	Y2	Y3	Observed frequency (n)	Observed proportion (n/N)
1	1	1	696	0.406
1	1	2	68	0.040
1	2	1	275	0.161
1	2	2	130	0.076
2	1	1	34	0.020
2	1	2	19	0.011
2	2	1	125	0.073
2	2	2	366	0.214

N = 1713

Profile for 2-class model

```
$Y1
           Pr(1)  Pr(2)
class 1:  0.9601 0.0399
class 2:  0.2284 0.7716

$Y2
           Pr(1)  Pr(2)
class 1:  0.7424 0.2576
class 2:  0.0429 0.9571

$Y3
           Pr(1)  Pr(2)
class 1:  0.9166 0.0834
class 2:  0.2395 0.7605
```

```
Estimated class population shares
 0.6205 0.3795
```

Model equation
for 2-class LC model for 3 indicators

Model for

$$P(y_1, y_2, y_3)$$

the probability of a particular response pattern.

For example, how likely is someone to hold the opinion
“allow speak, allow teach, but remove books from library:

P(Y1=1, Y2=1, Y3=2) = ?

Two key model assumptions

(X is the latent class variable)

1. (MIXTURE ASSUMPTION)
Joint distribution mixture of 2 class-specific distributions:

$$P(y_1, y_2, y_3) = P(X=1)P(y_1, y_2, y_3 \mid X=1) + P(X=2)P(y_1, y_2, y_3 \mid X=2)$$

2. (LOCAL INDEPENDENCE ASSUMPTION)
Within class X=x, responses are independent:

$$P(y_1, y_2, y_3 \mid X=1) = P(y_1 \mid X=1)P(y_2 \mid X=1)P(y_3 \mid X=1)$$

$$P(y_1, y_2, y_3 \mid X=2) = P(y_1 \mid X=2)P(y_2 \mid X=2)P(y_3 \mid X=2)$$

Example: model-implied proprtion

	X=1	X=2
P(X)	0.620	0.380
P(Y1=1|X)	0.960	0.229
P(Y2=1|X)	0.742	0.044
P(Y3=1|X)	0.917	0.240

P(Y1=1, Y2=1, Y3=2) =

(Mixture assumption)

$$P(Y1=1, Y2=1, Y3=2 \mid X=1)\, P(X=1) +$$
$$P(Y1=1, Y2=1, Y3=2 \mid X=2)\, P(X=2)$$

Example: model-implied proprtion

	X=1	X=2
P(X)	0.620	0.380
P(Y1=1|X)	0.960	0.229
P(Y2=1|X)	0.742	0.044
P(Y3=1|X)	0.917	0.240

P(Y1=1, Y2=1, Y3=2) =

(Mixture assumption)

P(Y1=1, Y2=1, Y3=2 | X=1) 0.620 +

P(Y1=1, Y2=1, Y3=2 | X=2) 0.380 =

(Local independence assumption)

P(Y1=1|X=1) P(Y2=1|X=1) P(Y2=2|X=1) 0.620 +

P(Y1=1|X=2) P(Y2=1|X=2) P(Y2=2|X=2) 0.380

Example: model-implied proprtion

	X=1	X=2
P(X)	0.620	0.380
P(Y1=1|X)	0.960	0.229
P(Y2=1|X)	0.742	0.044
P(Y3=1|X)	0.917	0.240

P(Y1=1, Y2=1, Y3=2) =

(Mixture assumption)

$P(Y1=1, Y2=1, Y3=2 \mid X=1)\ 0.620\ +$

$P(Y1=1, Y2=1, Y3=2 \mid X=2)\ 0.380\ =$

(Local independence assumption)

$(0.960)\,(0.742)\,(1-0.917)\,(0.620)\ +$

$(0.229)\,(0.044)\,(1-0.240)\,(0.380)\ \approx$

≈ 0.0396

The model again

Mixture of K classes

$$P(\boldsymbol{y}) = \sum_{x=1}^{K} P(\boldsymbol{y}\,|X = x)\, P(X = x)$$

Local independence of p variables

$$P(\boldsymbol{y}\,|X = x) = \prod_{j=1}^{p} P\left(y_j\,|X = x\right)$$

Both together gives the likelihood of the observed data:

$$P(\boldsymbol{y}) = \sum_{x=1}^{K} \prod_{j=1}^{p} P\left(y_j\,|X = x\right) P(X = x)$$

"Categorical data" notation

- In some literature an alternative notation is used

- Instead of Y1, Y2, Y3, variables are named A, B, C
- We define a model for the joint probability

$$P(A = i, B = j, C = k) := \pi_{ijk}^{ABC}$$

$$\pi_{ijk}^{ABC} = \sum_{t=1}^{T} \pi_{t}^{X} \pi_{ijkt}^{ABC|X} \quad \text{with} \quad \pi_{ijkt}^{ABC|X} = \pi_{it}^{A|X} \pi_{jt}^{B|X} \pi_{kt}^{C|X}$$

Loglinear parameterization

$$\pi_{ijk\,t}^{ABC|X} = \pi_{i\,t}^{A|X}\pi_{j\,t}^{B|X}\pi_{k\,t}^{C|X}$$

$$\ln(\pi_{ijk\,t}^{ABC|X}) = \ln(\pi_{i\,t}^{A|X}) + \ln(\pi_{j\,t}^{B|X}) + \ln(\pi_{k\,t}^{C|X})$$

$$:= \lambda_{i\,t}^{A|X} + \lambda_{j\,t}^{B|X} + \lambda_{k\,t}^{C|X}$$

The parameterization actually used in most LCM software

$$P(y_k \mid X = x) = \frac{\exp(\beta^k_{0y_k} + \beta^k_{1y_k x})}{\sum_{m=1}^{M_k} \exp(\beta^k_{0m} + \beta^k_{1mx})}$$

$\beta^k_{0y_k}$ Is a logistic intercept parameter

$\beta^k_{1y_k x}$ Is a logistic slope parameter (loading)

So just a series of **logistic regressions**, with X as independent and Y dep’t!
Similar to CFA/EFA (but logistic instead of linear regression)

A more realistic example

(showing how to evaluate the model fit)

One form of political activism

61.31%

38.69%

Another form of political activism

Relate to covariate?


ESTAMOS DESPIERTOS ¡ QUE HORA ES ?
YA ES HORA DE QUE SE VAYAN !

There are different ways of trying to improve things in [country] or help prevent[9] things from going wrong. During the last 12 months, have you done any of the following?
Have you…**READ OUT…**

		Yes	No	(Don't know)
B13	…contacted a politician, government or local government official?	1	2	8
B14	…worked in a political party or action group?	1	2	8
B15	…worked in another organisation or association?	1	2	8
B16	…worn or displayed a campaign badge/sticker?	1	2	8
B17	…signed a petition?	1	2	8
B18	…taken part in a lawful public demonstration?	1	2	8
B19	…boycotted certain products?	1	2	8

Data from the European Social Survey round 4 Greece

contplt	wrkprty	wrkorg	badge	sgnptit	pbldmn	bctprd	clsprty
2	2	2	2	2	2	1	2
2	2	2	2	2	2	1	1
2	2	2	2	2	1	1	1
2	2	2	2	2	2	2	1
2	2	2	2	2	2	2	1
2	2	2	2	2	2	1	2
2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	1
2	2	2	2	2	2	2	2

```
ess_greece <- read_csv("https://daob.nl/files/lca/ess_greece.csv.gz")

K <- 4        # Change to 1,2,3,4,..

fit_K <- poLCA(cbind( contplt, wrkprty, wrkorg,
                      badge, sgnptit, pbldmn, bctprd) ~ 1,
                      ess_greece, nclass = K)
```

Evaluating model fit

In the previous small example, you calculated the model-implied (expected) probability for response patterns and compared it with the observed probability of the response pattern:

```
observed - expected
```

The small example had $2^3 - 1 = 7$ unique patterns and 7 unique parameters, so df = 0 and the model fit perfectly.

$$\text{observed} - \text{expected} = 0 \iff df = 0$$

Evaluating model fit

Current model (with 1 class, 2 classes, ...)

Has $2^7 - 1 = 128 - 1 = 127$ unique response patterns

But much fewer parameters

So the model can be **tested**.

Different models can be compared with each other.

Evaluating model fit

- Global fit
- Local fit
- Substantive criteria

Global fit

Goodness-of-fit "chi-squared" statistics

$$\chi^2 = \sum_{\text{patterns}} \frac{(\text{observed } - \text{expected})^2}{\text{expected}}$$

$$L^2 = 2 \sum_{\text{patterns}} \text{observed} \cdot \ln\left(\frac{\text{observed}}{\text{expected}}\right)$$

- L^2 is sometimes called G2 and χ^2 is sometimes written as X2
- df = number of patterns - 1 - Npar
- Sparseness: bootstrap *p*-values

Information criteria

- For model comparison
- Parsimony versus fit

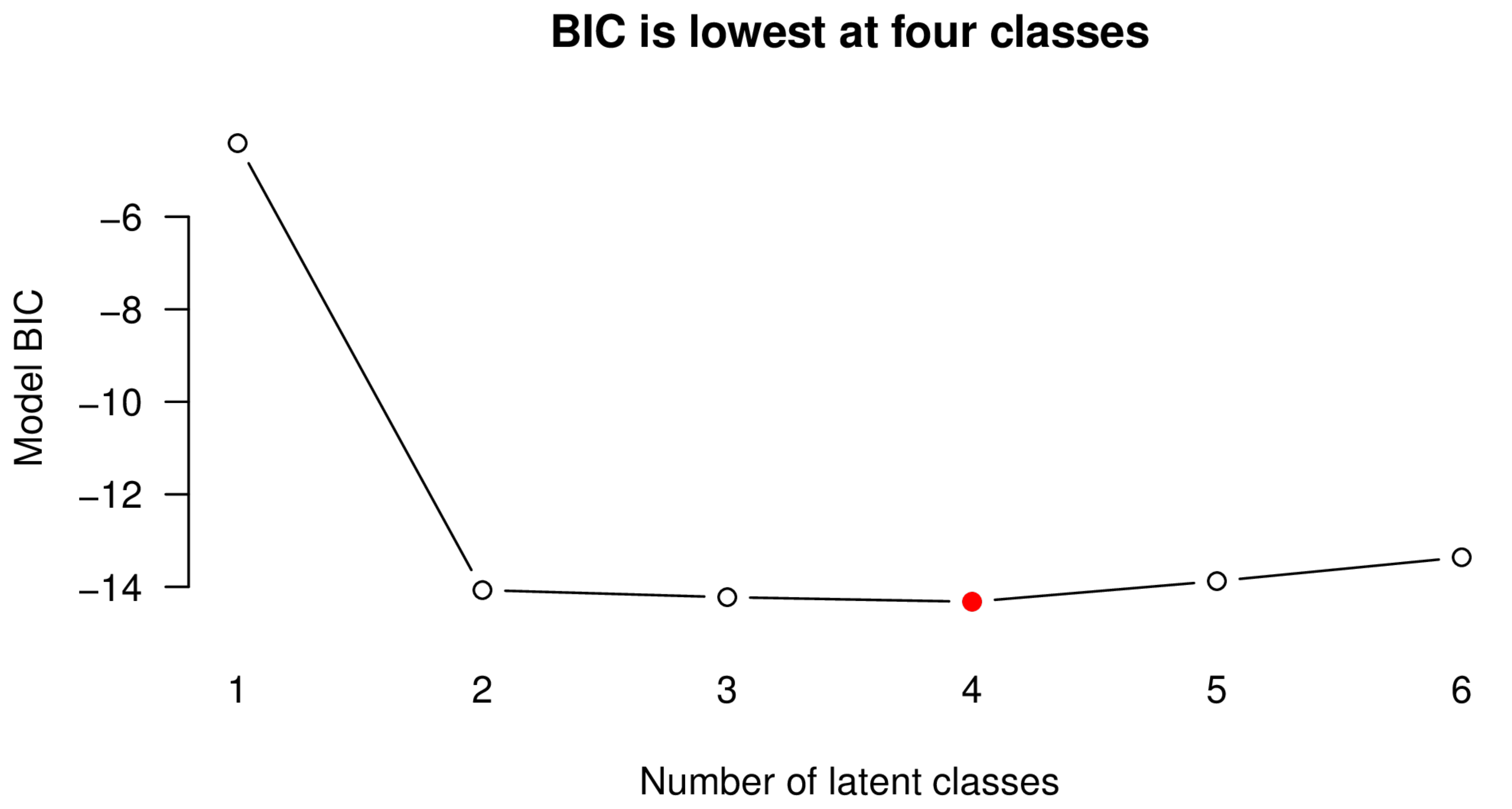
Common criteria:

$$BIC(L^2) = L^2 - \mathrm{df} \cdot \ln N \qquad BIC(LL) = -2LL + \; q \cdot \ln N$$
$$AIC(L^2) = L^2 - 2\,\mathrm{df} \qquad AIC(LL) = -2LL - 2\,q$$
$$AIC3(L^2) = L^2 - 3\,\mathrm{df} \qquad AIC3(LL) = -2LL - 3\,q$$

The L2 and LL versions are equivalent
(but give different numbers)

Model fit comparisons

	L^2	BIC(L^2)	AIC(L^2)	df	p-value
1-Cluster	1323.0	-441.0	861.0	120	0.000
2-Cluster	295.8	-1407.1	-150.2	112	0.001
3-Cluster	219.5	-1422.3	-210.5	104	0.400
4-Cluster	148.6	-1432.2	-265.4	96	1.000
5-Cluster	132.0	-1387.6	-266.0	88	1.000
6-Cluster	122.4	-1336.1	-259.6	80	1.000


BIC is lowest at four classes
Model BIC
−6
−8
−10
−12
−14
1
2
3
4
5
6
Number of latent classes

Local fit

Why doesn’t an LC model fit?

→ because **local independence assumption** is violated

Local fit: bivariate residuals (BVR)

Pearson “chi-squared” comparing observed and estimated frequencies in 2-way tables.

Expected frequency in two-way table:

$$N \cdot P(y_k, y_{k'}) = N \cdot \sum_{x=1}^{C} P(X = x) P(y_k \mid X = x) P(y_{k'} \mid X = x)$$

Observed:

Just make the bivariate cross-table from the data!

Example calculating a BVR

Observed

	No	Yes
No	3250	280
Yes	123	216

Expected

	No	Yes
No	3217	313
Yes	156	183

Bivariate residuals

	No	Yes
No	32.6	-32.6
Yes	-32.6	32.6

$$\mathrm{BVR}_{1,3} = r_{11}{}^2 \sum_{k,l} \hat{\mu}_{kl}^{-1} = (32.6)^2 \sum_{k,l} \hat{\mu}_{kl}^{-1} \approx 1063(0.0154) \approx 16.3$$

1-class model BVR's

	contplt	wrkprty	wrkorg	badge	sgnptit	pbldmn	bctprd
contplt	.						
wrkprty	342.806	.					
wrkorg	133.128	312.592	.				
badge	203.135	539.458	396.951	.			
sgnptit	82.030	152.415	372.817	166.761	.		
pbldmn	77.461	260.367	155.346	219.380	272.216	.	
bctprd	37.227	56.281	78.268	65.936	224.035	120.367	.

2-class model BVR's

	contplt	wrkprty	wrkorg	badge	sgnptit	pbldmn	bctprd
contplt	.						
wrkprty	15.147	.					
wrkorg	0.329	2.891	.				
badge	2.788	12.386	8.852	.			
sgnptit	2.402	1.889	9.110	0.461	.		
pbldmn	1.064	1.608	0.108	0.945	3.957	.	
bctprd	1.122	2.847	0.059	0.717	18.025	4.117	.

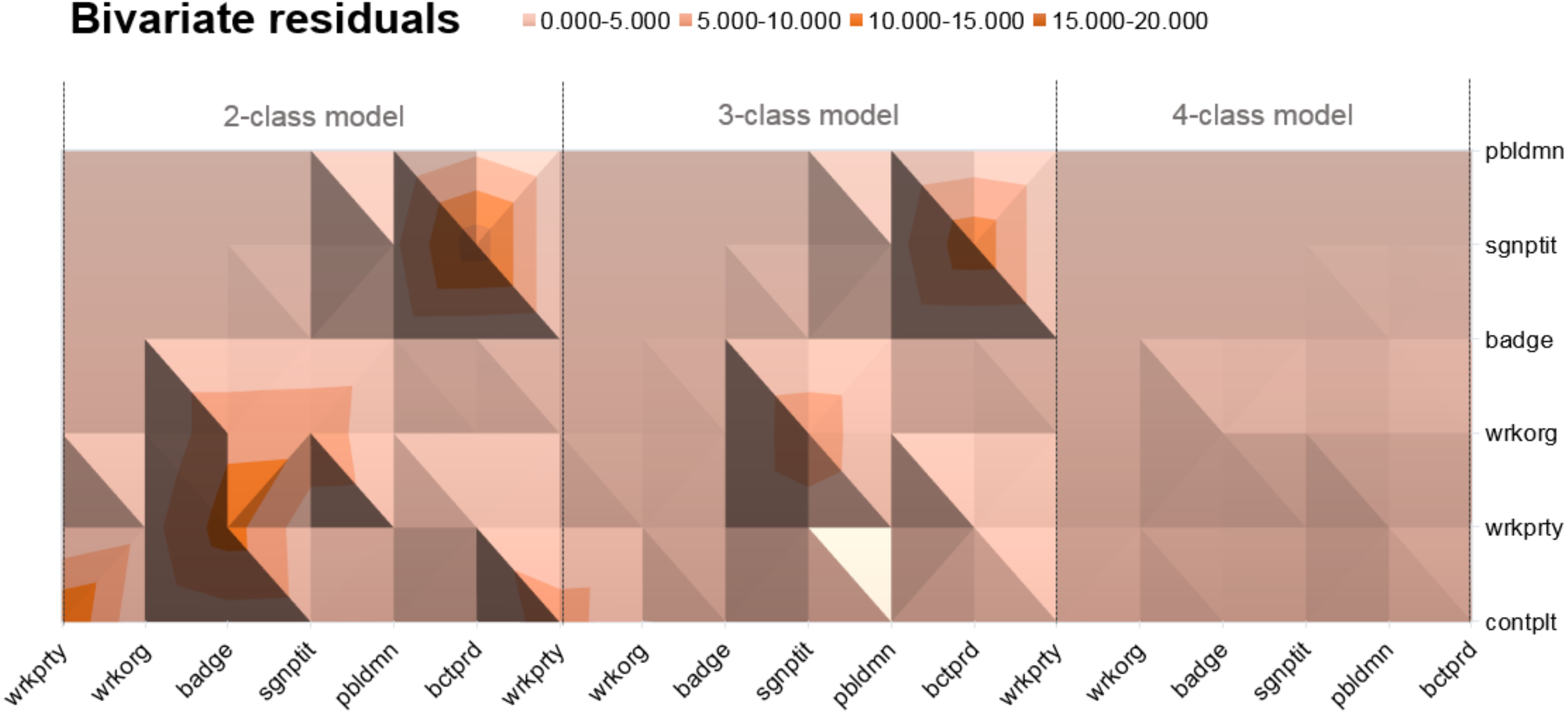
3-class model BVR's

	contplt	wrkprty	wrkorg	badge	sgnptit	pbldmn	bctprd
contplt	.						
wrkprty	7.685	.					
wrkorg	0.048	0.370	.				
badge	0.282	0.054	0.273	.			
sgnptit	2.389	2.495	8.326	0.711	.		
pbldmn	2.691	0.002	0.404	0.086	2.842	.	
bctprd	2.157	2.955	0.022	0.417	13.531	1.588	.

4-class model BVR's

	contplt	wrkprty	wrkorg	badge	sgnptit	pbldmn	bctprd
contplt	.						
wrkprty	0.659	.					
wrkorg	0.083	0.015	.				
badge	0.375	0.001	1.028	.			
sgnptit	0.328	0.107	0.753	0.019	.		
pbldmn	0.674	0.939	0.955	0.195	0.004	.	
bctprd	0.077	0.011	0.830	0.043	0.040	0.068	.

Bivariate residuals

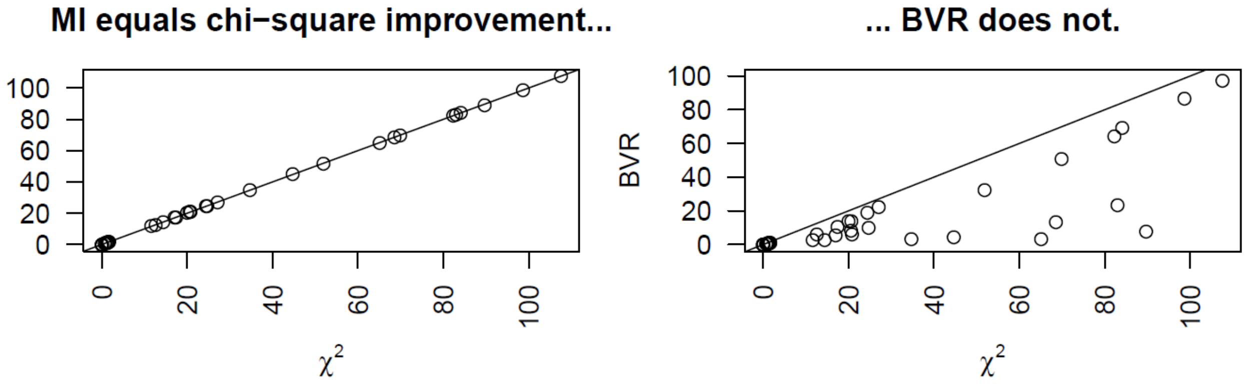


Local fit: bootstrapping BVR

The bivariate residual (BVR) is not actually chi-square distributed! (Oberski, Van Kollenburg & Vermunt 2013)

Solutions:

- Bootstrap p-values of BVR (LG5)
- “Modification indices” (score test) (LG5)



Example of modification index (score test) for 2-class model

Covariances / Associations							
term			coef	EPC(self)	Score	df	BVR
contplt	<->	wrkprty	0	1.7329	28.5055	1	15.147
wrkorg	<->	wrkprty	0	0.6927	**4.3534**	1	**2.891**
badge	<->	wrkprty	0	1.3727	16.7904	1	12.386
sgnptit	<->	bctprd	0	1.8613	37.0492	1	18.025

wrkorg <-> wrkparty is "not significant" according to BVR but is when looking at score test!

(but not after adjusting for multiple testing)

Why doesn't an LC model fit?

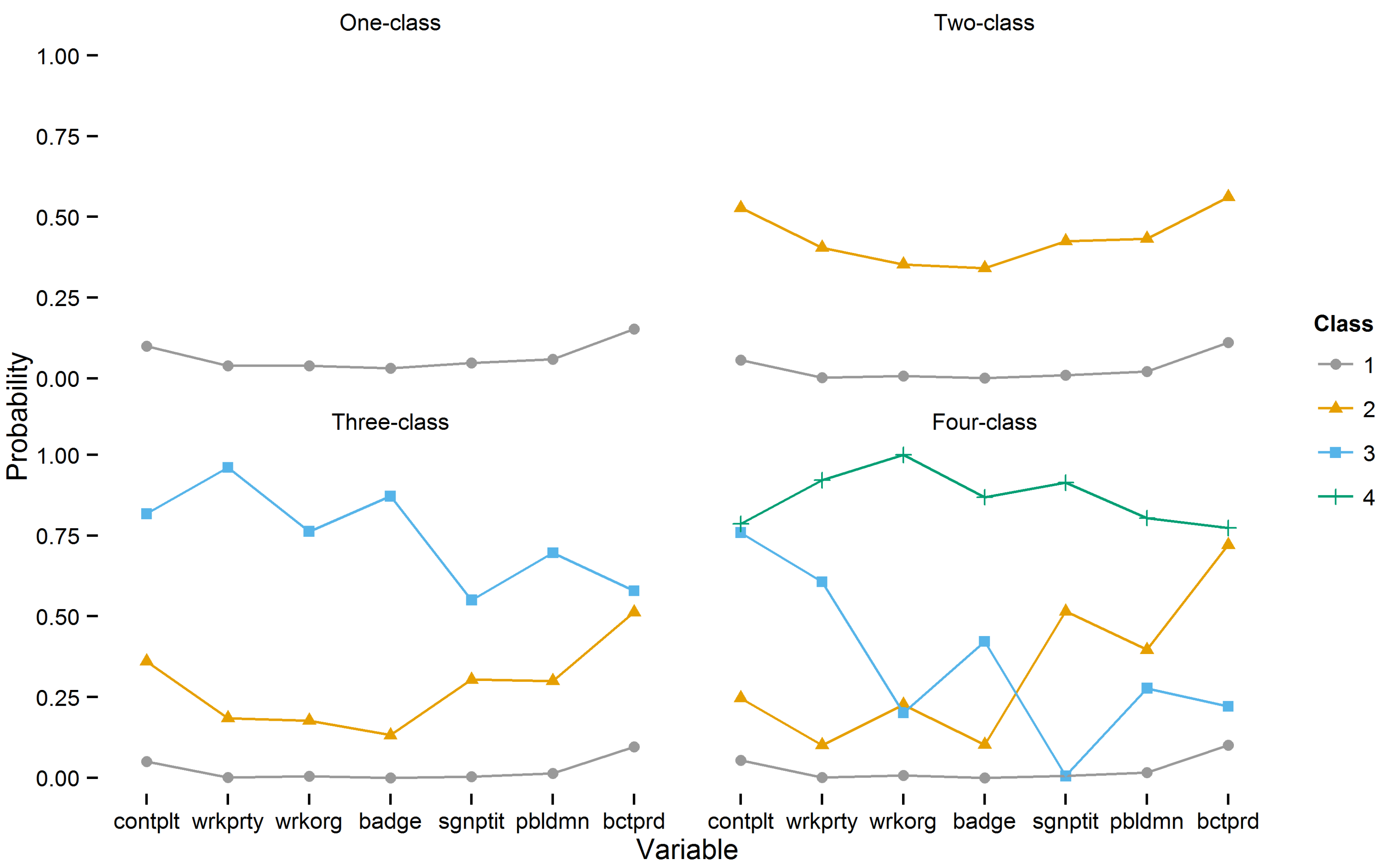
Answer: because **local independence assumption** is violated

Three possible solutions:

1. Increase the number of clusters or latent classes;
2. Increase the number of discrete factors or latent variables;
3. Allow for **local dependencies** or direct relationships between certain items;
4. Ignore the issue.

Option 3 is similar to correlated errors in structural equation models (SEM)

Interpreting the results and using substantive criteria



One-class
Two-class
Three-class
Four-class
1.00
0.75
0.50
0.25
0.00
Probability
contplt
wrkprty
wrkorg
badge
sgnptit
pbldmn
bctprd
Variable
Class
1
2
3
4

EPC-interest for looking at change in substantive parameters

After fitting two-class model, how much would loglinear “loadings” of the items change if local dependence is accounted for?

term			Y1	Y2	Y3	Y4	Y5	Y6	Y7
contplt	<->	wrkprty	-0.44	-0.66	0.05	1.94	0.05	0.02	0.00
wrkorg	<->	wrkprty	0.00	-0.19	-0.19	0.63	0.02	0.01	0.00
badge	<->	wrkprty	0.00	-0.37	0.03	-1.34	0.03	0.01	0.00
sgnptit	<->	bctprd	0.01	0.18	0.05	1.85	-0.58	0.02	-0.48

See Oberski (2013); Oberski & Vermunt (2013); Oberski, Moors & Vermunt (2015)

Model fit evaluation: summary

Different types of criteria to evaluate fit of a latent class model:

- **Global**
  BIC, AIC, L2, X2
- **Local**
  Bivariate residuals, modification indices (score tests), and expected parameter changes (EPC)
- **Substantive**
  Change in the solution when adding another class or parameters

Model fit evaluation: summary

- Compare models with different number of classes using BIC, AIC, bootstrapped L2

- Evaluate overall fit using bootstrapped L2 and bivariate residuals

- Can be useful to look at the profile of the different solutions: if nothing much changes, or very small classes result, fit may not be as useful