Latent class analysis

Latent class analysis extensions

DL Oberski & L Boeschoten

Extension topics

- Local dependence models
- Multiple latent variables
- **Ordinal indicators**
- **Tree-step modelling**

Ordinal indicators

Ordinal indicators

- So far: dichotomous or unordered polytomous indicators
- When indicators are **ordinal**, we can restrict their relationship with the latent variable(s)
- Several options:

1. Different types of logits: **adjacent-category**, cumulative, continuation-ratio
2. Inequality restrictions
3. Binomial count (k out of K)

- LG approach: **adjacent-category logit** models with fixed scores for item categories, for example,
- Assumption: local odds-ratios are category independent

Another “classic” example

General Social Survey 1982 (see McCutcheon, 1987; Magidson & Vermunt, 2001, 2004)

- Evaluation of surveys by respondent (2 questions)
    - Purpose
    - Accuracy
- Evaluation of respondent by interviewer (2 questions)
    - Understanding
    - Cooperation
- Are there different types of survey respondents?

Profiles

	Cluster 1	Cluster 2	Cluster 3
	Ideal	*Believers*	*Skeptics*
Class Sizes	0.6169	0.2038	0.1793
PURPOSE			
Good	0.8905	0.9157	0.1592
Depends	0.0524	0.0706	0.2220
Waste	0.0570	0.0137	0.6189
ACCURACY			
Mostly True	0.6148	0.6527	0.0426
Not True	0.3852	0.3473	0.9574
UNDERSTAND			
Good	0.9957	0.3241	0.7532
Fair, poor	0.0043	0.6759	0.2468
COOPERATE			
Interested	0.9452	0.6879	0.6432
Cooperative	0.0547	0.2583	0.2559
Impatient/ Hostile	0.0001	0.0538	0.1009

Fit measures GSS example

Model	L^2	prop. red. L^2	df	*p*-value	BIC	AIC	class. errors
1-Cluster	257.26	0.00	29	0.00	51.60	199.26	0.00
2-Cluster	79.51	0.69	22	0.00	-76.51	35.51	0.08
3-Cluster	22.09	0.91	15	0.11	-84.29	-7.91	0.13
4-Cluster	6.61	0.97	8	0.58	-50.12	-9.39	0.20

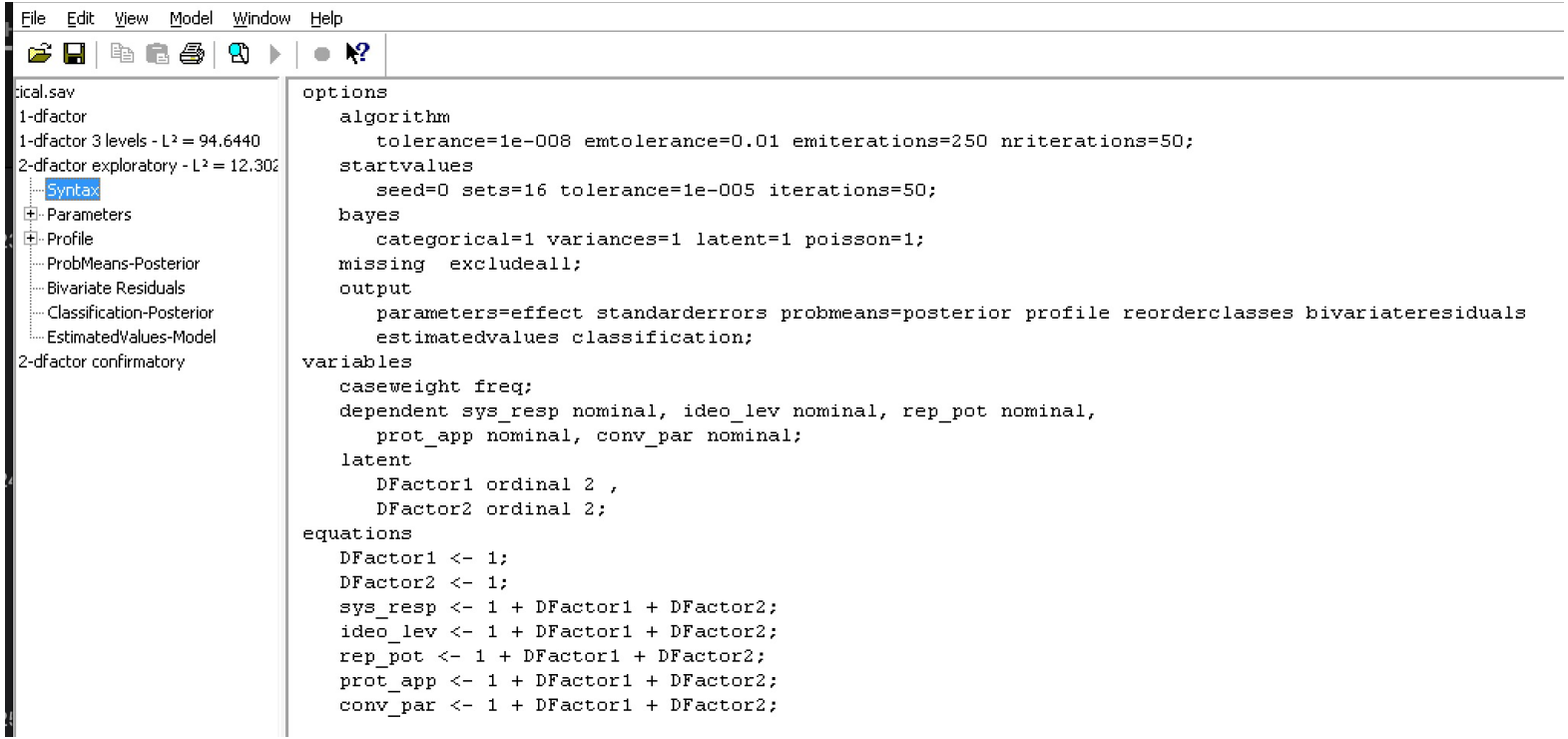
Example GSS (cont.)

Model		L^2	BIC	df	p-value
Nominal	1-Cluster	257.26	51.60	29	0.00
	2-Cluster	79.51	-76.51	22	0.00
	3-Cluster	22.09	-84.29	15	0.11
Ordinal	1-Cluster	257.26	51.60	29	0.00
	2-Cluster	82.60	-87.60	24	0.00
	3-Cluster	30.67	-104.07	19	0.04

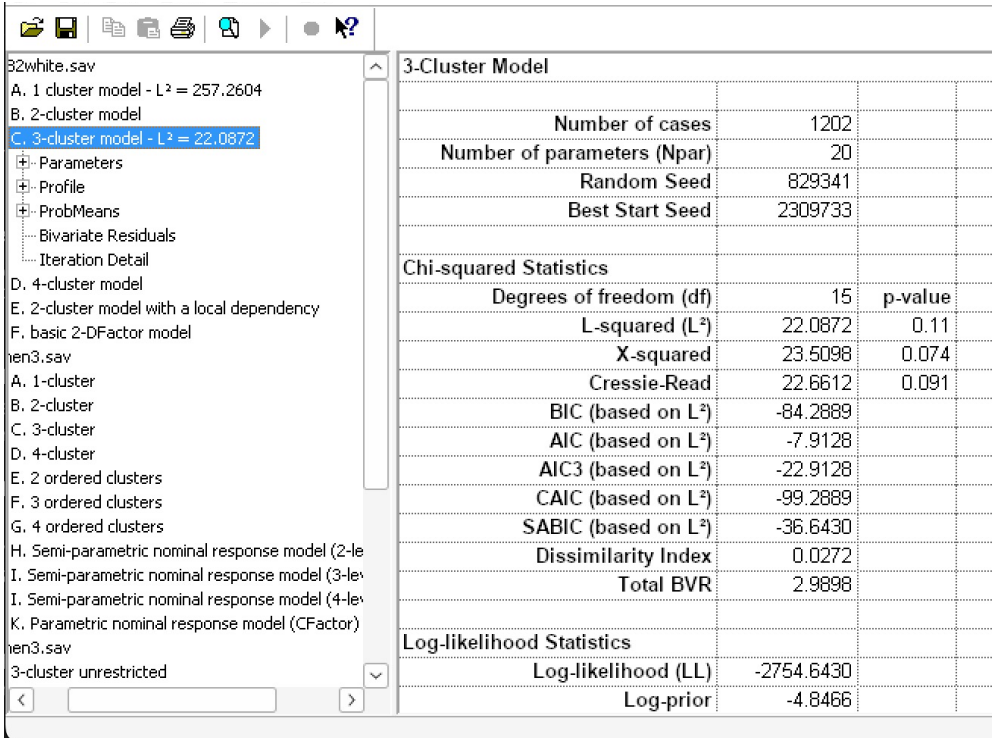
Profiles ordinal

	Cluster1	Cluster2	Cluster3
Cluster Size	0.6407	0.1925	0.1668
PURPOSE			
Good	0.8792	0.2512	0.9167
Depends	0.0756	0.1475	0.0579
waste	0.0451	0.6013	0.0254
ACCURACY			
mostly true	0.6233	0.0415	0.6753
not true	0.3767	0.9585	0.3247
UNDERSTA			
good	0.9949	0.7483	0.2027
fair/poor	0.0051	0.2517	0.7973
COOPERAT			
interested	0.9362	0.6365	0.6971
cooperative	0.0609	0.2754	0.2412
impatient/hostile	0.0029	0.0881	0.0617

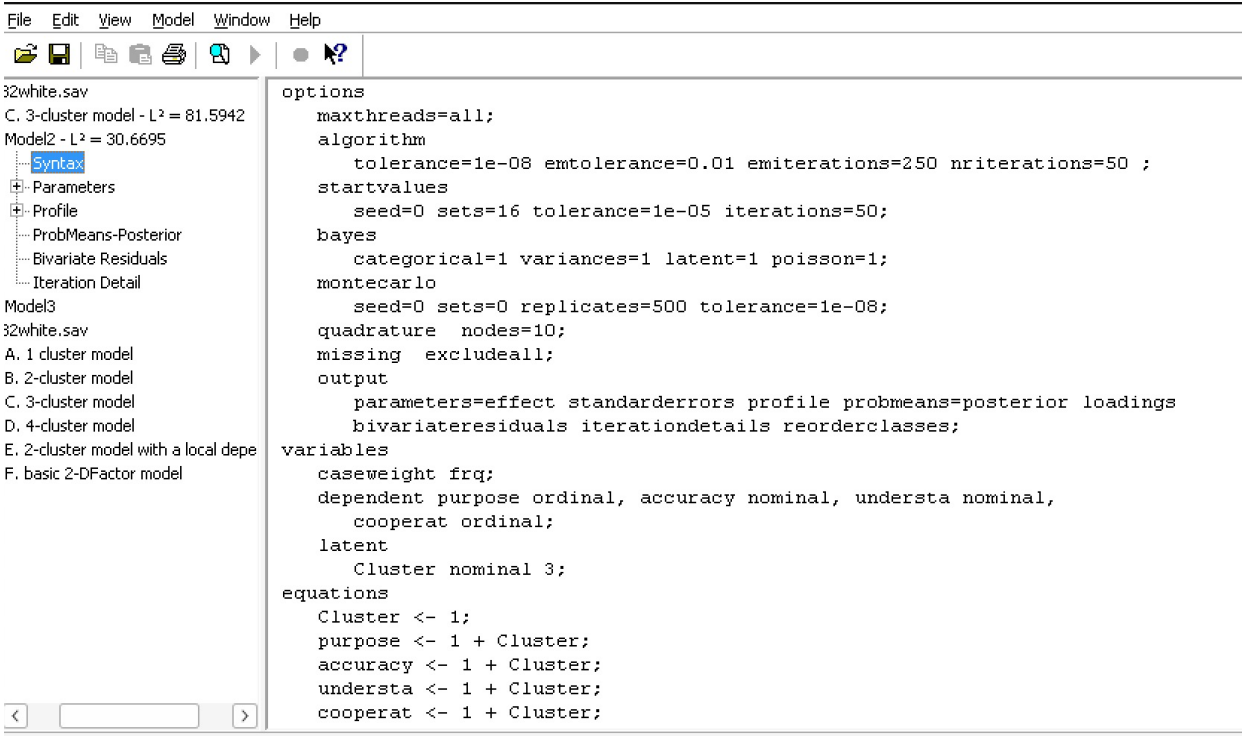
In Latent GOLD - nominal



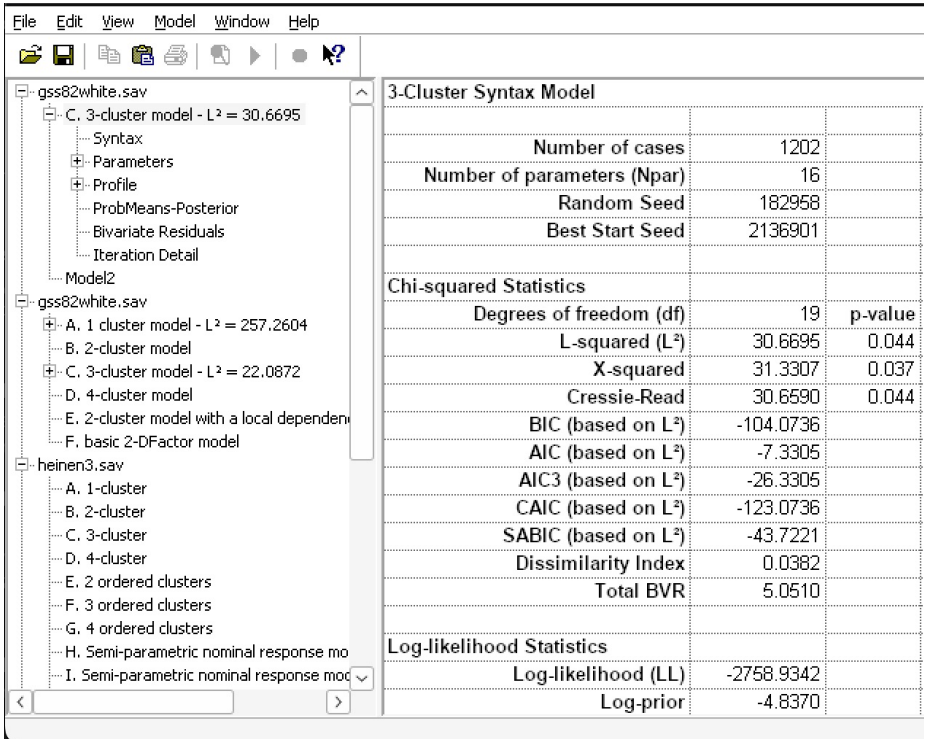
In Latent GOLD - nominal



In Latent GOLD - ordinal



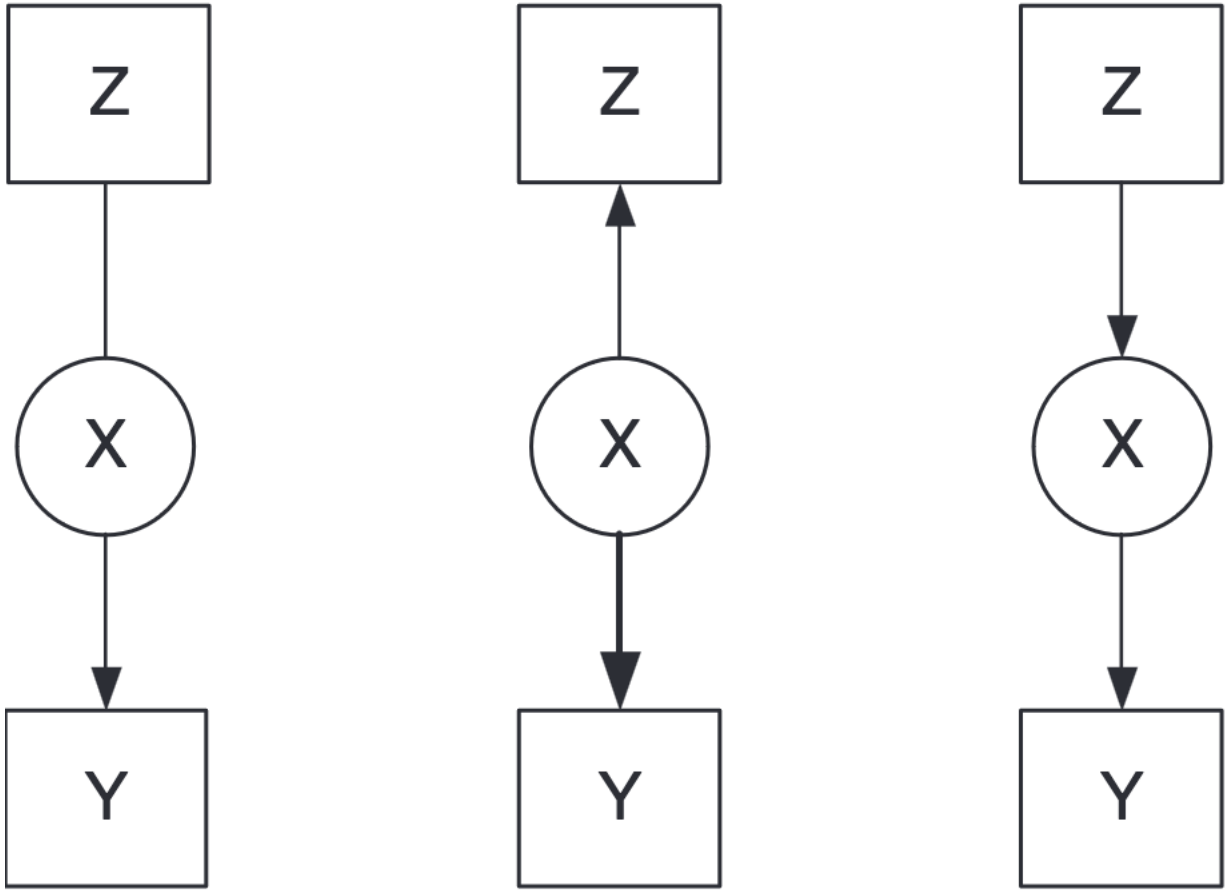
In Latent GOLD - ordinal



Three step modeling

Analyzing the assigned class memberships

Correction for misclassification


Z
X
Y
Z
X
Y
Z
X
Y

Three step approach *without* correction ("classical three step")

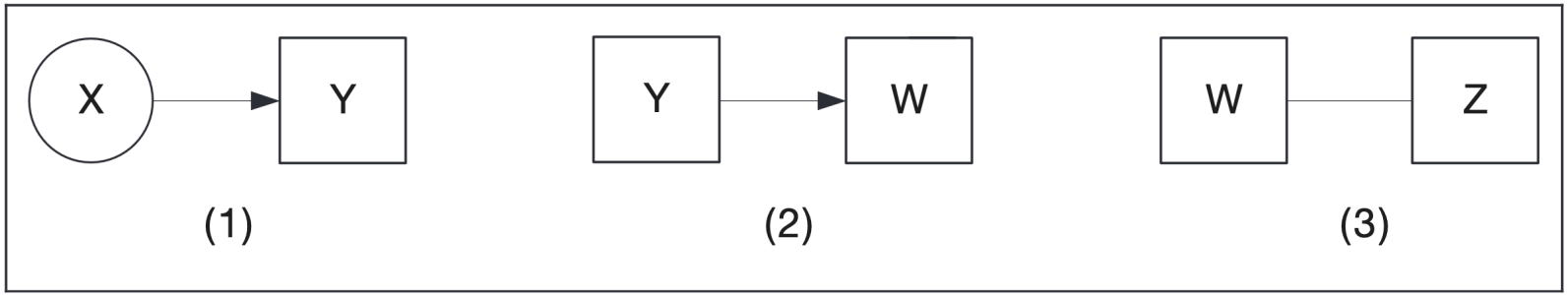


Figure 2. The steps of the standard three-step approach.

Source: Bakk & Kuha 2017

Classify-analyze does not work!

- You might think that after classification it is easy to model people's latent class membership

- "Just take assigned class and analyze it as though observed"

- Unfortunately, -> **biased estimates** and wrong se's

  (Still relatively little-known among practicioners?)

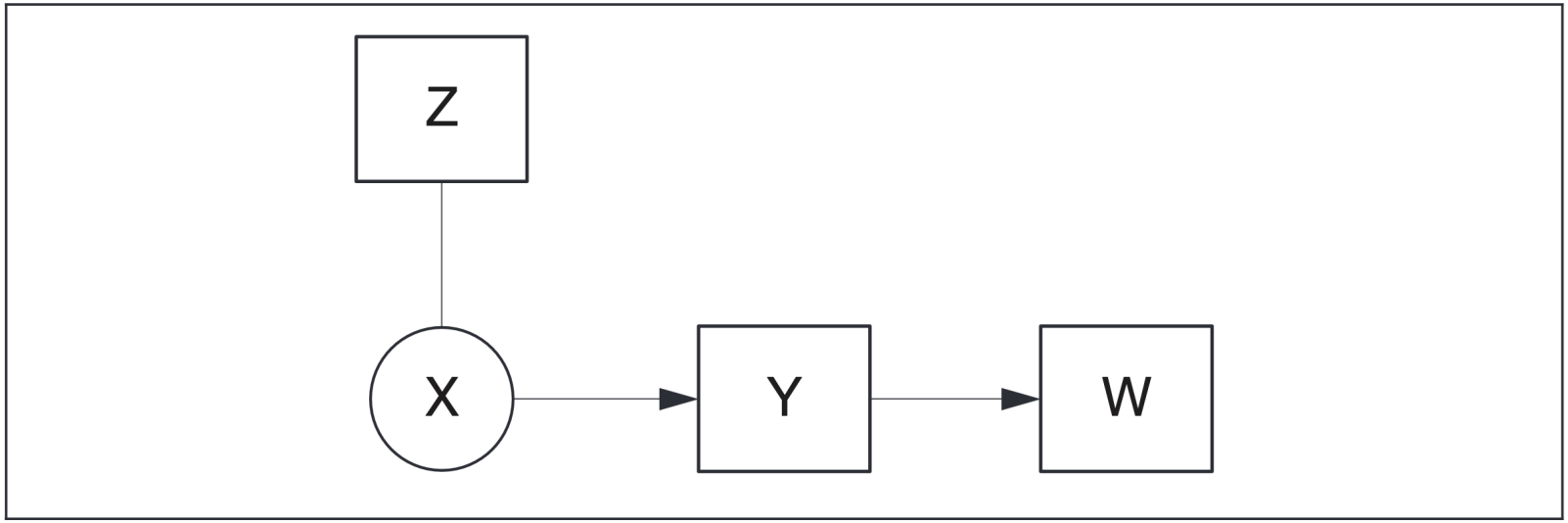


Figure 3. The relationship between variables *W, X,* **Y**, and *Z* in the three-step approach.

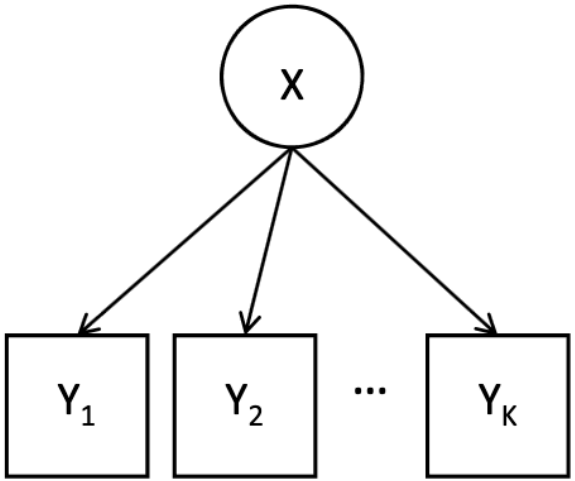
“Classical 3step” does not work!

Solutions:

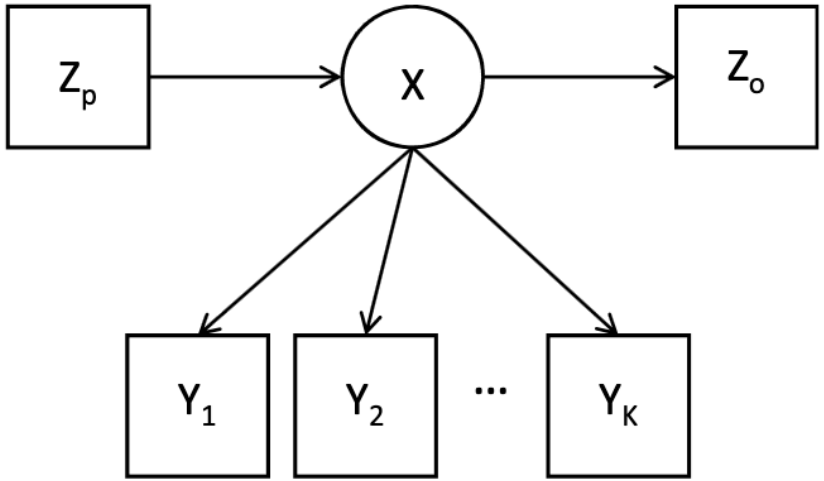
- *Ignore the problem.*
  - This will lead to bias and wrong standard errors.
  - But bias might not be so bad compared with additional variance from alternatives (tradeoff)
  - Hard to know when this will be the case…

- *Model covariates directly* (as in `poLCA` or loglinear `cvam`)
  - “circularity” : Feels like “cheating” (even though from model-based perspective nothing is wrong)
  - Can change the class solution sometimes to something less useful
  - Can be tricky when the intended use of the class assignments is not a multinomial regression

“One step approach”



Measurement model



One-step approach

Source: Bakk & Kuha 2020

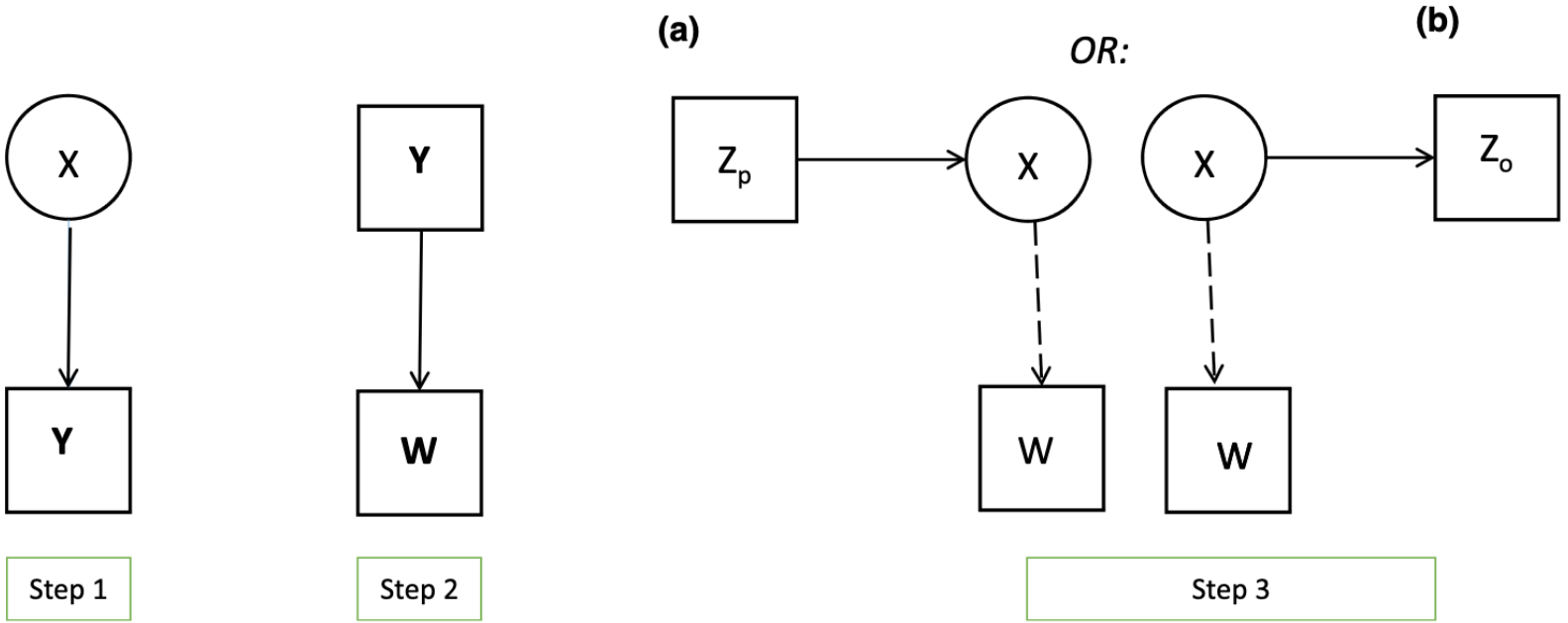
How to model (X, Z)?

Solutions:

- Three-step modeling: (BCH, 2002; Vermunt 2010; Bakk et al 2013)
    1. Run the latent class model with Y and X only
    2. Compute things on the LC model
        a. Assign cases to their suspected class W
        b. Compute the classification matrix E
    3. Correct (W|Z) for misclassification in W as a measure of X using E

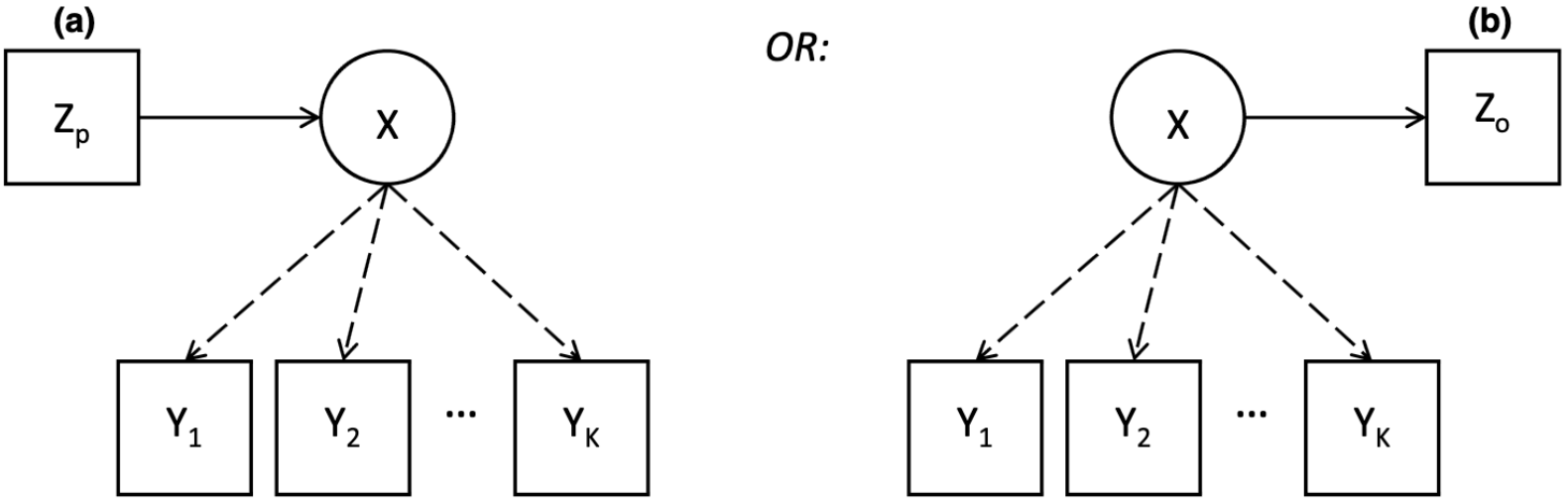
- Two-step modeling: (Bakk & Kuha 2018)
    1. Run the latent class model with Y and X only
    2. Run another latent class model, now with covariate Z, while fixing all "measurement parameters", (Y|X), to estimates from step 1

Three-step approach *with correction*



Source: Bakk & Kuha 2020

Two-step approach



Step 2 of the two-step approach

Source: Bakk & Kuha 2020

Table 1. Recommended modelling approaches for different circumstances

Approach	Distal outcome	Covariate	Large models
One-step approach	Not recommended	Recommended with reservations	Recommended with reservations
ML	Recommended only if underlying model assumptions are met for models with continuous and count outcomes. Suitable for categorical outcomes	Recommended	Extendable to models with multiple latent variables, multiple outcomes and covariates. Not possible to (easily) model DIF[a]
BCH	Most robust for continuous and count outcomes. Negative weights can be counterintuitive	Recommended	Currently not possible to extend to complex models, or to model DIF
Two-step approach	Not recommended with continuous or count outcome. Recommended with categorical outcome	Recommended	Recommended, most flexible stepwise approach. Flexible for modelling DIF
LTB	Recommended if no heteroscedasticity is present for continuous and count outcome. Recommended for categorical outcome	Not appropriate	Not appropriate
Classical three-step	Not recommended	Not recommended	Not recommended

Note. [a]By DIF we mean differential item functioning, also known as direct effect between the covariate and (some) indicator(s).

Source: Bakk & Kuha 2020

Correction for misclassification

Which is of these is true?

Measurement error/misclassification:

- Always causes estimates of relationships between variables to be *underestimated*;
- Is something you can only try to *prevent* as much as possible. You can't really do anything about it once it's there;
- Is not relevant when you're only looking at *trends* (over time);
- Is not relevant when you're only looking at *averages*;
- Is taken care of by using assigned classes/factor scores;
- Is a problem with self-reports, but not with administrative registers or other behavioral variables.

Which is of these is true?

Measurement error/misclassification:

- ~~Always causes estimates of relationships between variables to be *underestimated*;~~
- ~~Is something you can only try to *prevent* as much as possible. You can't really do anything about it once it's there;~~
- ~~Is not relevant when you're only looking at *trends* (over time);~~
- ~~Is not relevant when you're only looking at *averages*;~~
- ~~Is taken care of by using assigned classes/factor scores;~~
- ~~Is a problem with self-reports, but not with administrative registers or other behavioral variables.~~

Correction for misclassfication

- Standard BCH
- Generalized BCH
- ML

Standard/classic BCH

- Named after Bolck, Croon & Hagenaars (2002)
- Invented by Fuller (1987)
- (Example of Stigler’s law of eponymy)

Real example with simple categorical data analysis

	True vote (admin)	
Answered	No	Yes
”No”	0.671	0.015
”Yes”	0.329	0.985
Total	1	1

	Ideology	
True vote	Dem.	Rep.
No	0.405	0.312
Yes	0.595	0.688
Total	1	1

	Ideology	
Answered	Dem.	Rep.
”No”	0.305	0.248
”Yes”	0.695	0.752
Total	1	1

	Ideology	
True vote	Dem.	Rep.
No	0.405	0.312
Yes	0.595	0.688
Total	1	1

	Ideology	
Answered	Dem.	Rep.
”No”	0.305	0.248
”Yes”	0.695	0.752
Total	1	1

A little numerical example

- Proportion Democrats who say ”No” is a *mixture* of
  - 0.671 of true No’s say ”No” and
  - 0.015 of true Yes’s who say ”No”
- This leads to:
  - (0.671)(0.405) + (0.015)(0.595) = 0.281.
- So we *observe* about 30% of nonvoters among Democrats,
- when the *truth* is 40% nonvoters among Democrats.

General result

The observed table is the result of a matrix multiplication between the misclassification rates and the true table:

$$\overbrace{\begin{bmatrix} \textit{No} & \textit{Yes} \\ 0.671 & 0.015 \\ 0.329 & 0.985 \end{bmatrix}}^{\text{Misclassification}} \cdot \overbrace{\begin{bmatrix} & \textit{Dem.} & \textit{Rep.} \\ \textit{No} & 0.405 & 0.312 \\ \textit{Yes} & 0.595 & 0.688 \end{bmatrix}}^{\text{True}} \approx \overbrace{\begin{bmatrix} & \textit{Dem.} & \textit{Rep.} \\ \text{"}\textit{No}\text{"} & 0.305 & 0.248 \\ \text{"}\textit{Yes}\text{"} & 0.695 & 0.752 \end{bmatrix}}^{\text{Observed}}$$

K*now*: **Observed table** and **Misclassification table**

W*ant*: **True table**

→To estimate true table, just invert the equation:

$$\overbrace{\begin{bmatrix} & \textit{Dem.} & \textit{Rep.} \\ \textit{No} & 0.405 & 0.312 \\ \textit{Yes} & 0.595 & 0.688 \end{bmatrix}}^{\text{True}} \approx \overbrace{\begin{bmatrix} \textit{No} & \textit{Yes} \\ 0.671 & 0.015 \\ 0.329 & 0.985 \end{bmatrix}^{-1}}^{\text{Misclassification}} \cdot \overbrace{\begin{bmatrix} & \textit{Dem.} & \textit{Rep.} \\ \textit{"No"} & 0.305 & 0.248 \\ \textit{"Yes"} & 0.695 & 0.752 \end{bmatrix}}^{\text{Observed}}$$

- Note the **negative weights**
- Many people have intuition that *M* should be used as weights
- It's actually the inverse!

$$\overbrace{\begin{bmatrix} & \textit{Dem.} & \textit{Rep.} \\ \textit{No} & 0.405 & 0.312 \\ \textit{Yes} & 0.595 & 0.688 \end{bmatrix}}^{\text{True}} \approx \overbrace{\begin{bmatrix} \textit{No} & \textit{Yes} \\ 1.502 & -0.023 \\ -0.502 & 1.023 \end{bmatrix}}^{\text{Misclassification inverse}} \cdot \overbrace{\begin{bmatrix} & \textit{Dem.} & \textit{Rep.} \\ \textit{"No"} & 0.305 & 0.248 \\ \textit{"Yes"} & 0.695 & 0.752 \end{bmatrix}}^{\text{Observed}}$$

- For comparison, in our example we did have the true values
- and the true and estimated tables appear similar:

$$\overbrace{\begin{bmatrix} & \textit{Dem.} & \textit{Rep.} \\ \textit{No} & 0.442 & 0.355 \\ \textit{Yes} & 0.558 & 0.645 \end{bmatrix}}^{\text{Estimated}} \overset{?}{\approx} \overbrace{\begin{bmatrix} & \textit{Dem.} & \textit{Rep.} \\ \textit{No} & 0.405 & 0.312 \\ \textit{Yes} & 0.595 & 0.688 \end{bmatrix}}^{\text{True}} \neq \overbrace{\begin{bmatrix} \textit{Dem.} & \textit{Rep.} \\ 0.305 & 0.248 \\ 0.695 & 0.752 \end{bmatrix}}^{\text{Observed}}$$

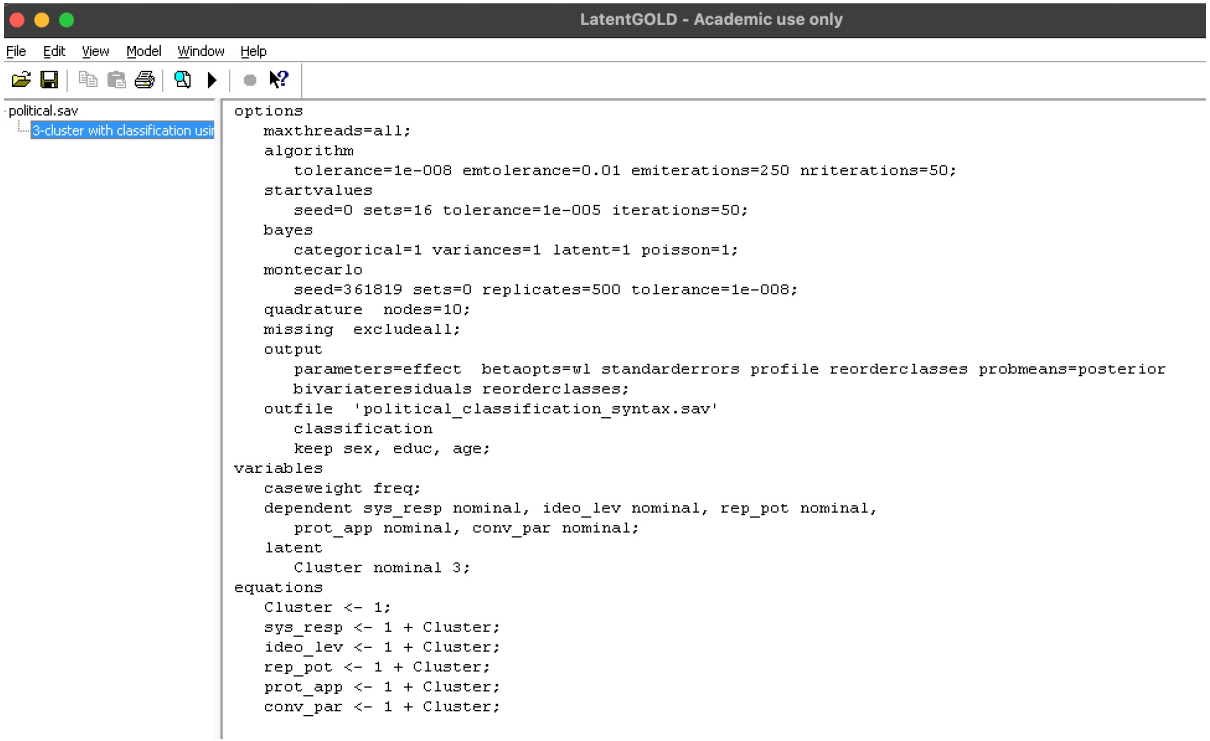
- So in this example the correction of the Turnout-Ideology crosstable for measurement error has worked very well;
- In practice of course we cannot usually do this check!

“Generalized” BCH (Vermunt 2010)

- BCH impractical for ~more complicated than cross-table
- We saw that the elements of the inverse of the classification matrix are used as weights (some negative)
- It turns out we can extend this idea:

1. Create a “complete-data” set in which each original observation is replicated K times, one for each value of X
2. Set the weight of each observation to the appropriate element of the inverse classification matrix
3. Run any analysis of your choice which allows (neg.) weights.
   (using observation ID as cluster corrects standard errors)

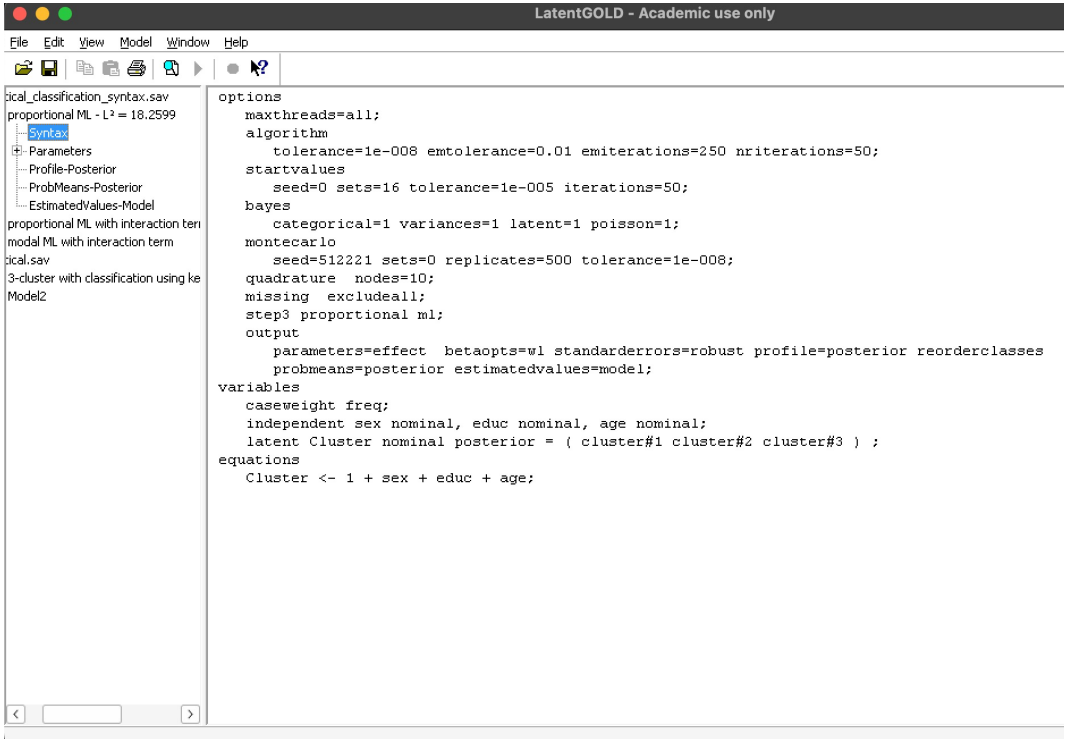
3 step in Latent GOLD – step 1 & 2



3 step in Latent GOLD

rep_pot Repression Potential	prot_app Protest Approval	conv_par Convential Participation	freq cell count	sex Sex	educ Education (training)	age Age (generation)	clu#1 Cluster1	clu#2 Cluster2	clu#3 Cluster3	clu# Cluster modal
2	1	1	1	1	2	3	0.22885148	0.582529165	0.1886193499	2
2	1	1	1	2	1	3	0.22885148	0.582529165	0.1886193499	2
2	1	1	1	2	2	2	0.22885148	0.582529165	0.1886193499	2
2	1	2	1	1	1	2	0.70227264	0.085522685	0.2122046782	1
2	1	2	1	1	2	2	0.70227264	0.085522685	0.2122046782	1
2	1	2	1	2	2	2	0.70227264	0.085522685	0.2122046782	1
2	2	1	4	1	1	1	0.05050353	0.060949501	0.8885469688	3
2	2	1	1	1	2	1	0.05050353	0.060949501	0.8885469688	3
2	2	1	1	1	2	3	0.05050353	0.060949501	0.8885469688	3
2	2	1	1	2	1	1	0.05050353	0.060949501	0.8885469688	3
2	2	1	2	2	2	1	0.05050353	0.060949501	0.8885469688	3
2	2	1	1	2	2	2	0.05050353	0.060949501	0.8885469688	3
2	2	2	6	1	1	1	0.13319179	0.007690198	0.8591180074	3
2	2	2	3	1	1	2	0.13319179	0.007690198	0.8591180074	3
2	2	2	1	1	1	3	0.13319179	0.007690198	0.8591180074	3

3 step in Latent GOLD - ML



3 step in Latent GOLD – ML & BCH

- step3
  - <proportional | modal>
  - <ml | bch | none>
- The command WriteBCH=‘filename’ creates an expanded data file with one record per latent class for each observation and the weights needed for the various types of non-adjusted and bch-adjusted 3-step analysis methods