Latent class analysis

Latent class analysis extensions

DL Oberski & L Boeschoten

Extension topics

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

Local dependence models

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.

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

Modeling local dependence (loglinear)

Example:

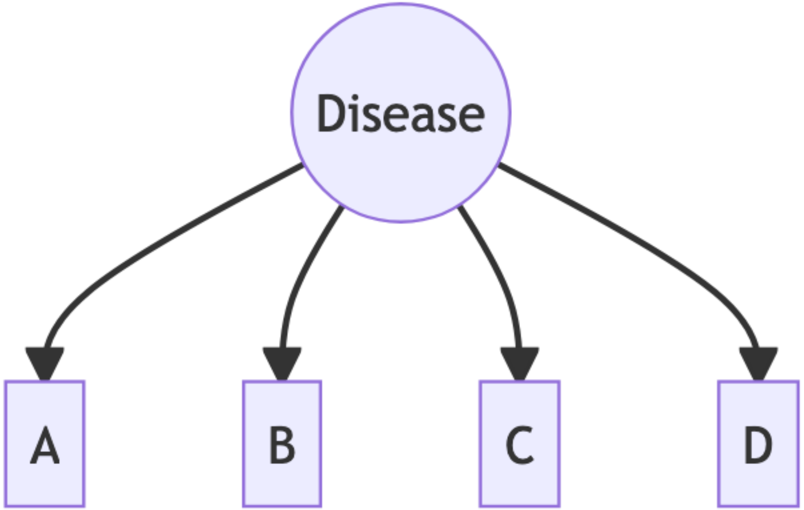
$$P(\mathbf{Y}_i = \mathbf{y}) = \sum_{x=1}^{K} P(x)\, P(Y_{i1} = y_1, Y_{i2} = y_2 \mid x)\, P(Y_{i3} = y_3 \mid x)$$

With:

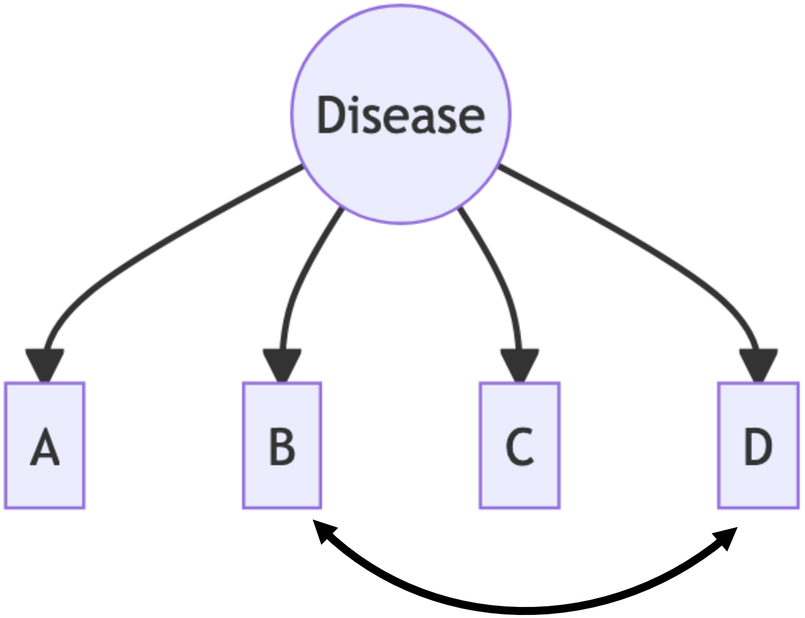
$$P(Y_{i1} = y_1, Y_{i2} = y_2 \mid x) = \frac{\exp(\beta^{1}_{0\,y_1} + \beta^{2}_{0\,y_2} + \beta^{12}_{0\,y_1 y_2} + \beta^{1}_{x y_1} + \beta^{2}_{x y_2})}{\sum_{y_1 y_2} \exp(\beta^{1}_{0\,y_1} + \beta^{2}_{0\,y_2} + \beta^{12}_{0\,y_1 y_2} + \beta^{1}_{x y_1} + \beta^{2}_{x y_2})}$$

Interpretation: two items have a stronger association than can be explained by clusters/DFactors

Independence model



Local *dependence* model



Diagnostic tests

Research | Open Access | Published: 10 March 2023

Estimating sensitivity and specificity of diagnostic tests using latent class models that account for conditional dependence between tests: a simulation study

Suzanne H. Keddie, Oliver Baerenbold, Ruth H. Keogh & John Bradley

BMC Medical Research Methodology **23**, Article number: 58 (2023) | Cite this article

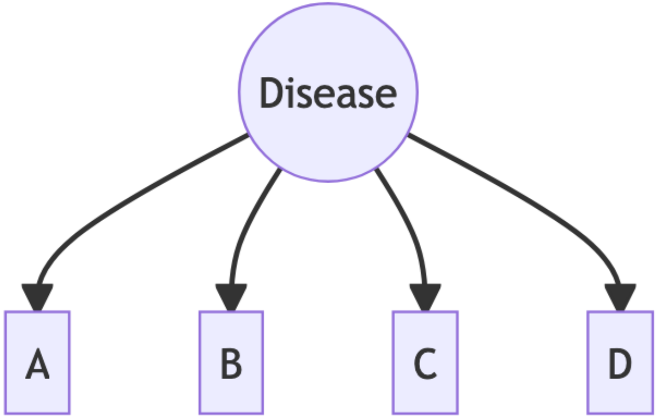
Local dependence methods

- Loglinear modeling (preferred option)
  - Easy to specify (once the loglinear LCA is up & running)
  - Can test nested models for fit
- Combining the two variables into one (“joint item method”)
  - Easy to do & understand
  - Can use ”plain vanilla” LCA software
  - Trouble when there is another local dependence
  - Inflexible and prone to overfitting with polytomous items
- Direct effect method
  - Conceptually different, but may be what you wanted
  - Main advantage is that you can use flexmix

Alvord data

Test	Description
A	Radioimmunoassay of antigen ag121
B	Radioimmunoassay of HIV p24
C	Radioimmunoassay of HIV gp120
D	Enzyme-linked immunosorbent assay

Local dependence example




```r
f_ueber <- cbind(A, B, C, D) ~ 1
fit_ueber_polca <- poLCA(f_ueber, data = uebersax_fulldata)
```

Local dependence example

```
pvals_boot <- bootstrap_bvr_pvals(f_ueber,
                                  data = uebersax_fulldata,
                                  fit_polca = fit_ueber_polca,
                                  nclass = 2, nrep = 3)

pvals_boot
```

```
      A     B     C
B 0.610
C 0.520 0.000
D 0.265 0.560 0.240
```

Using loglinear formula

```
~ X * (A + B + C + D)
```

	coef
(Intercept)	−12.8057
X1	−3.3879
A1	−4.0673
B1	0.7533
C1	4.3829
D1	−4.2866
X1:A1	−5.8097
X1:B1	−0.8964
X1:C1	−5.5574
X1:D1	−5.5040

Local dependence example

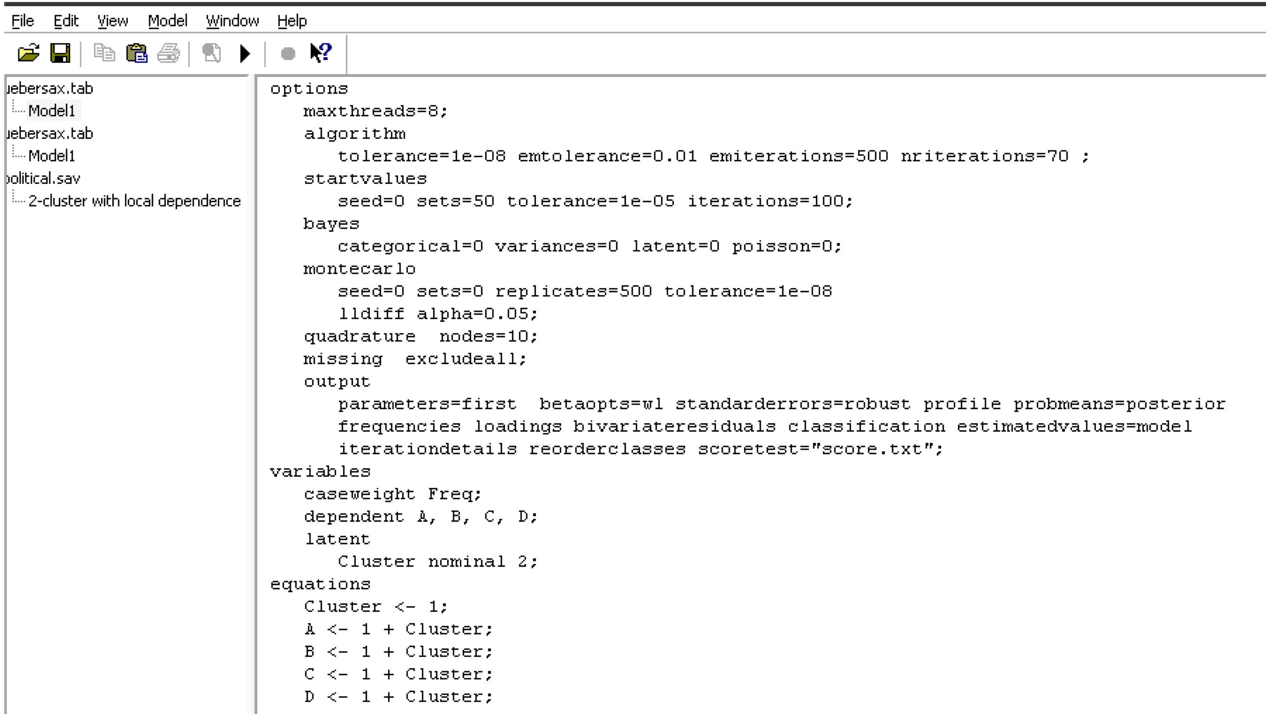
```r
formula_ld <- update(formula, ~ . + B:C)

system.time(
  fit_cvam_ld <-
    cvam(formula_ld, data = df_freq, freq = Freq,
         control = list(startValJitter = 0.05)
))
```

Local dependence example

```r
anova(fit_cvam, fit_cvam_ld)
```

```
Model 1: ~ X * (A + B + C + D)
Model 2: ~ X + A + B + C + D + X:A + X:B + X:C + X:D + B:C
  resid.df -2*loglik df change
1        6   -3070.8
2        5   -3084.0  1 13.171
```

Comparison to Latent GOLD



Comparison to Latent GOLD

Comparison to Latent GOLD

2-Cluster Syntax Model					
Estimation Warnings! See Iteration Detail					
Number of cases	428				
Number of parameters (Npar)	9				
Robustness Effect	0.6700				
Random Seed	270649				
Best Start Seed	939431				
Monte Carlo Seed	270649				
Chi-squared Statistics			Bootstrap		
Degrees of freedom (df)	6	p-value	p-value	s.e.	CV
L-squared (L^2)	16.2272	0.013	0.0020	0.0020	6.3428
X-squared	17.1146	0.0089			
Cressie-Read	16.4174	0.012			
BIC (based on L^2)	-20.1275				
AIC (based on L^2)	4.2272				
AIC3 (based on L^2)	-1.7728				
CAIC (based on L^2)	-26.1275				
SABIC (based on L^2)	-1.0872				
Dissimilarity Index	0.0398				
Total BVR	4.6230				
Log-likelihood Statistics					
Log-likelihood (LL)	-629.8827				
Log-prior	0.0000				
Log-posterior	-629.8827				
BIC (based on LL)	1314.2975				
AIC (based on LL)	1277.7654				
AIC3 (based on LL)	1286.7654				
CAIC (based on LL)	1323.2975				
SABIC (based on LL)	1285.7369				

2-Cluster Syntax Model					
Estimation Warnings! See Iteration Detail					
Number of cases	428				
Number of parameters (Npar)	10				
Robustness Effect	0.7016				
Random Seed	81233				
Best Start Seed	306966				
Monte Carlo Seed	81233				
Chi-squared Statistics			Bootstrap		
Degrees of freedom (df)	5	p-value	p-value	s.e.	CV
L-squared (L^2)	3.0560	0.69	0.1120	0.0141	4.1716
X-squared	4.4875	0.48			
Cressie-Read	3.7095	0.59			
BIC (based on L^2)	-27.2396				
AIC (based on L^2)	-6.9440				
AIC3 (based on L^2)	-11.9440				
CAIC (based on L^2)	-32.2396				
SABIC (based on L^2)	-11.3726				
Dissimilarity Index	0.0038				
Total BVR	0.1648				
Log-likelihood Statistics					
Log-likelihood (LL)	-623.2971				
Log-prior	0.0000				
Log-posterior	-623.2971				
BIC (based on LL)	1307.1854				
AIC (based on LL)	1266.5941				
AIC3 (based on LL)	1276.5941				
CAIC (based on LL)	1317.1854				
SABIC (based on LL)	1275.4515				

Multiple latent variables

Voting in NL

https://www.dataarchive.lissdata.nl/data_variables/view/5115

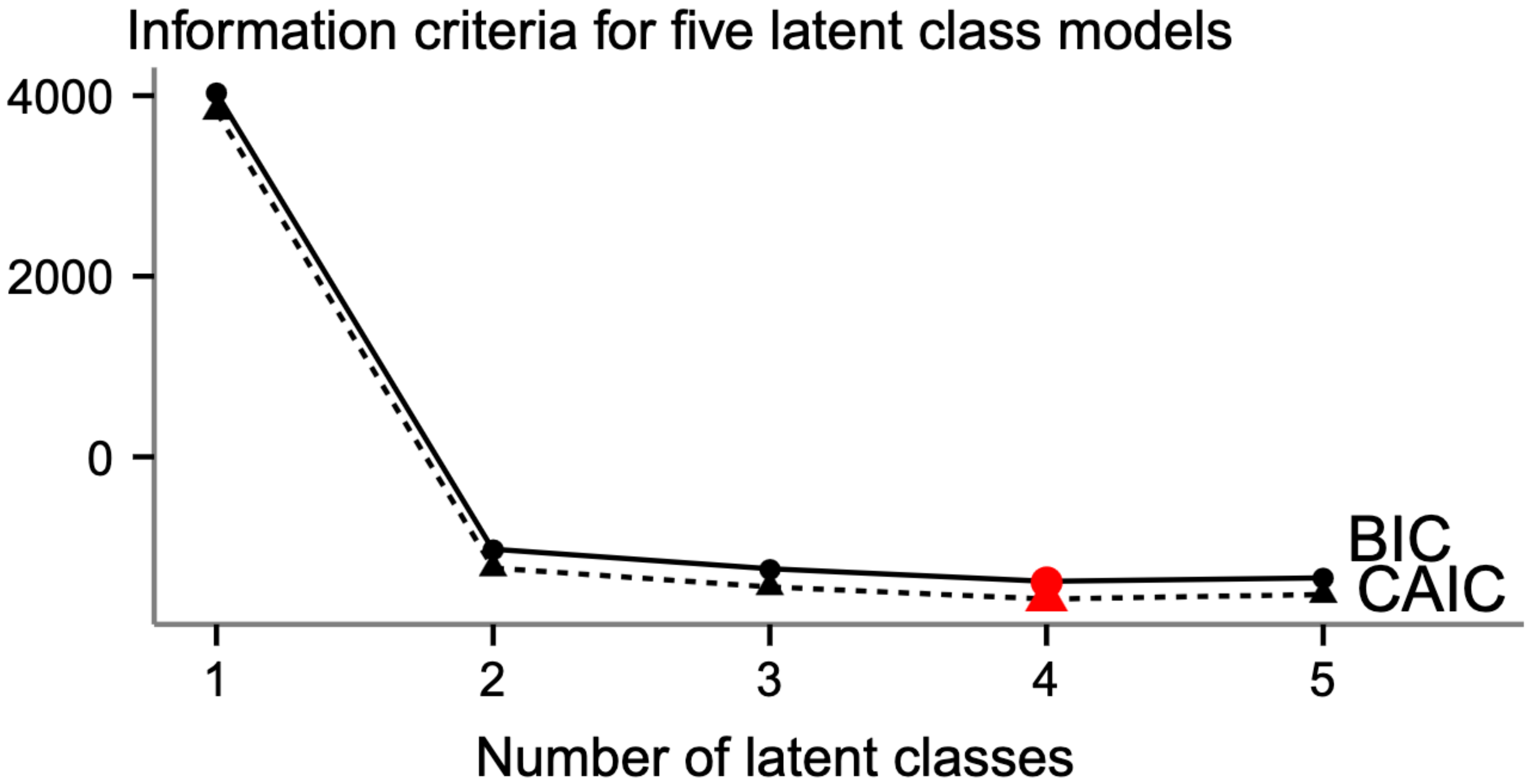
"Did you vote in the most recent parliamentary elections, held on 22 November 2006 / … ?"

1	yes
2	no

Voting in NL

ELECTION			ELECTION		
2006	2008	2009	2010	2011	2012
	`cv08a053`	`cv09b053`	`cv10c053`	`cv11d053`	`cv12e053`

Oberski, D.L. Beyond the number of classes: separating substantive from non-substantive dependence in latent class analysis. *Adv Data Anal Classif* **10**, 171–182 (2016). https://doi.org/10.1007/s11634-015-0211-0



Information criteria for five latent class models
4000
2000
0
1
2
3
4
5
BIC
CAIC
Number of latent classes

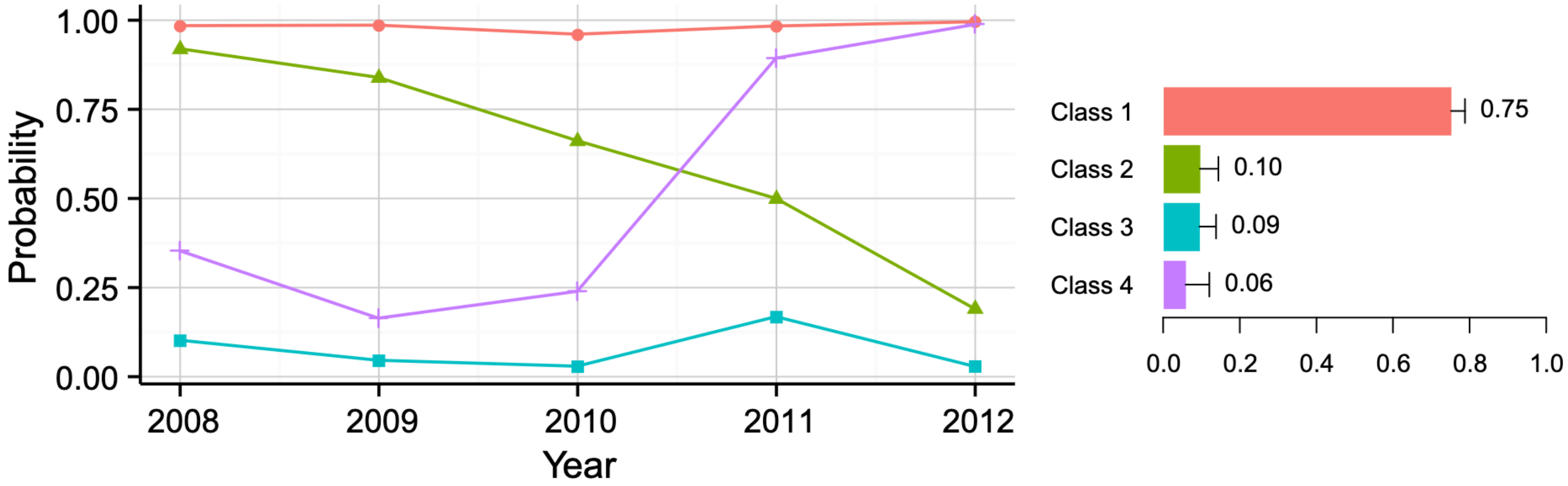
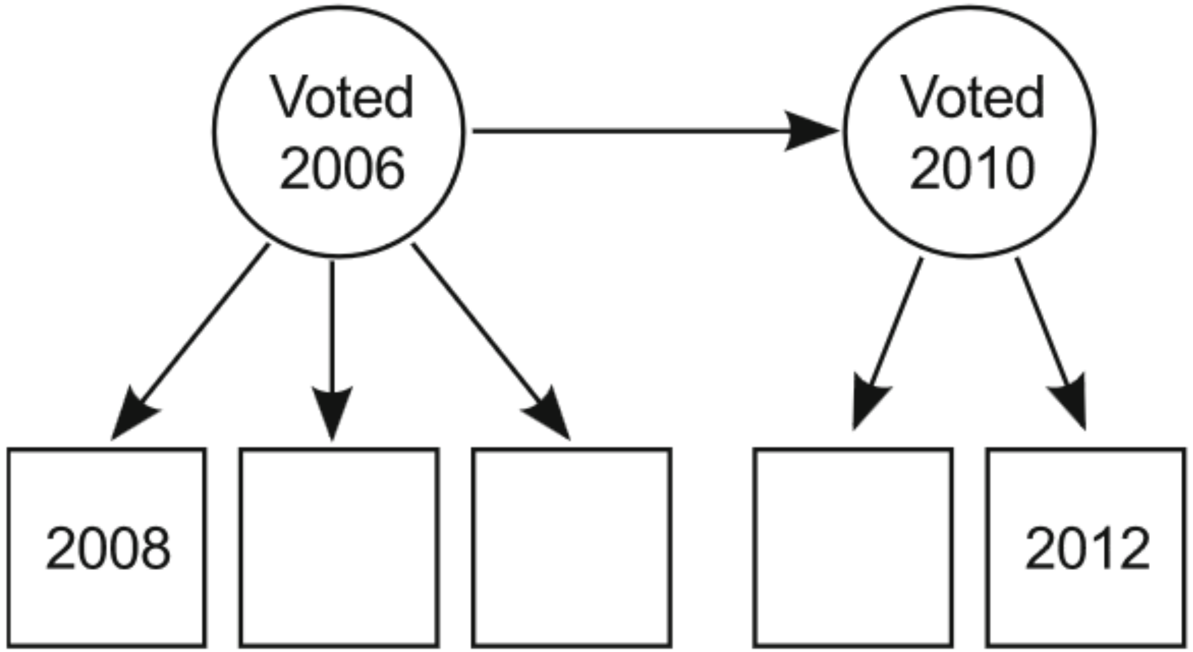


Fig. 2 *Left*: probability profile plot for the four-class solution. *Right*: legend with estimated class sizes and 2 s.e. error bars



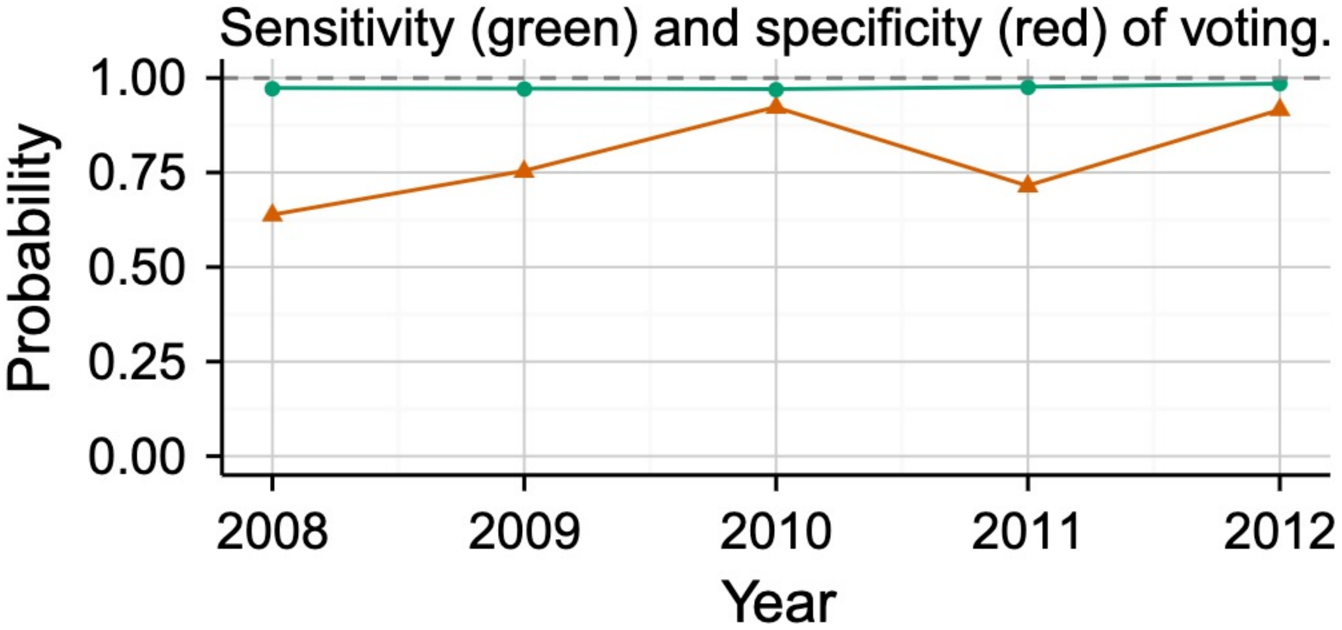
Voted 2006
Voted 2010
2008
2012

```r
df_freq$X1 <- latentFactor(NROW(df_freq), 2)
df_freq$X2 <- latentFactor(NROW(df_freq), 2)
```

```
> head(df_freq)

     A B C D E Freq   X1   X2
1    0 0 0 0 0  125 <NA> <NA>
2    1 0 0 0 0   15 <NA> <NA>
3 <NA> 0 0 0 0   33 <NA> <NA>
4    0 1 0 0 0    7 <NA> <NA>
5    1 1 0 0 0   23 <NA> <NA>
6 <NA> 1 0 0 0    5 <NA> <NA>
```

Loglinear LCA using cvam

```
formula <-
    ~ X1 * (A + B + C) + X2 * (D + E) + X1:X2

fit_cvam <-
    cvam(formula, data = df_freq, freq = Freq)
```


Sensitivity (green) and specificity (red) of voting.
1.00
0.75
0.50
0.25
0.00
Probability
2008
2009
2010
2011
2012
Year

	Voted in 2006		
	No	Yes	
2010			
No	0.713	0.051	18%
Yes	0.287	0.949	82%
	19%	81%	