EAM2025 XI Conference

23RD - 25TH JULY 2025

Spain Tenerife Canary Islands

European Association of Methodology

Please enter your presentation title here

Pier-Olivier Caron, Université TÉLUQ, Montréal, Canada

A fit index for latent class analysis of dichotomous scale

P.-O. Caron
Université TÉLUQ

EAM2025 – Spain
XI Conference – European Congress of Methodology
July 23th 2025

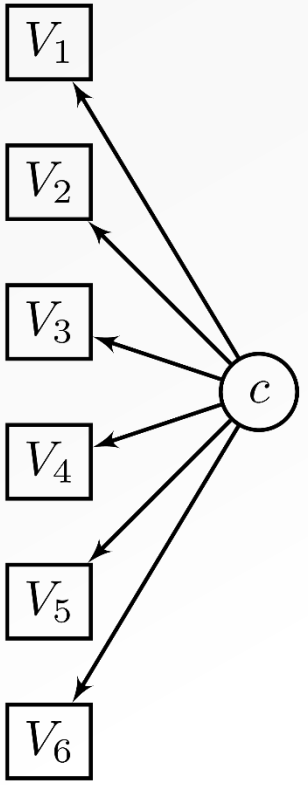
Thank you to the Université TÉLUQ for the support from a FAR3 funding.

UNIVERSITÉ TÉLUQ

2025-07-22

Latent class analysis

- Identifying unobserved groups within a population
- Categorical data
- Part of finite mixture modelling
- Exploratory
- Latent refers to unobserved variables inferred from observed indicators
- In contrast to factor analysis, which attempts to group indicators together, LCA groups participants



Latent class analysis

- Selecting the optimal number of latent classes remains a challenge
- There is no consensus on which fit index to use
- Among fit indices :
  - AIC,
  - BIC,
  - saBIC,
  - χ^2,
  - CAIC,
  - AIC3,
  - Lo-Mendell-Rubin,
  - Vuong-Lo-Mendell-Rubin,
  - Bootstrap LRT
- With a care to group sizes and theoretical justification

The fit index

- Based on the properties of dichotomous variables
- Capitalizes on the recovery of the model implied covariance matrix, $\mathbf{S}(\theta)$, from the response probabilities to measure its discrepancy with the sample covariance matrix $\mathbf{S}$

The fit index

Based on the pattern $\mathbf{X}$ where each row represents one of the 2^I binary I-tuples

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,I} \\ \vdots & \ddots & \vdots \\ x_{2^I,I} & \cdots & x_{2^I,I} \end{bmatrix}$$

where $x_{i,j} \in \{0,1\}\, \forall i \in \{1,2,\ldots,2^I\}, j \in \{1,2,\ldots,I\}$, I is the number of item, the pattern probabilities are

$$\mathbf{P}_i = \sum_{k=1}^{K} \prod_{j=1}^{I} p\left(x_{i,j}^{(k)}\right) c_k,$$

where K is the number of classes and c is the class probability, we derived the implied covariance matrix

$$\mathbf{S}(\theta) = (\mathbf{XP})'\mathbf{X} - \mathbf{MM}',$$

where $\boldsymbol{M}_j = \sum_{i=1}^{2^I} \boldsymbol{X}_{i,j} \boldsymbol{P}_i$.

The fit index

```
probs <- x$probs
 if(is.null(nclass)) {nclass <- length(x$P)}
 n <- x$N
 for(k in 1:nclass){
   pr <- sapply(probs, function(x) x[k,], simplify = FALSE)
   it <- sapply(pr, function(x) 1:length(x), simplify = FALSE)
   egpr <- expand.grid(pr)
   egpr$pr <- apply(egpr, 1, prod) * x$P[k]
   if(k == 1){
     rez <- (cbind(expand.grid(it), pr = egpr$pr))
   } else {
     rez$pr <- rez$pr +  egpr$pr
   }
 }

 rez <- as.matrix(rez)
 nitem <- ncol(rez)
 M <- as.matrix(colSums(rez[,-nitem] * rez[,nitem]))
 ES <- (t(rez[,-nitem] * rez[,nitem])) %*% rez[,-nitem] - M %*% t(M)
```

Degrees of freedom

- Using the square difference of the Fisher transform of both covariance matrices, we derived a pseudo χ_v^2 statistic
- Three degree of freedom tested :

$$v = \frac{p(p-1)}{2}$$

$$v = \frac{(p-k)((p-k)-1)}{2} - k$$

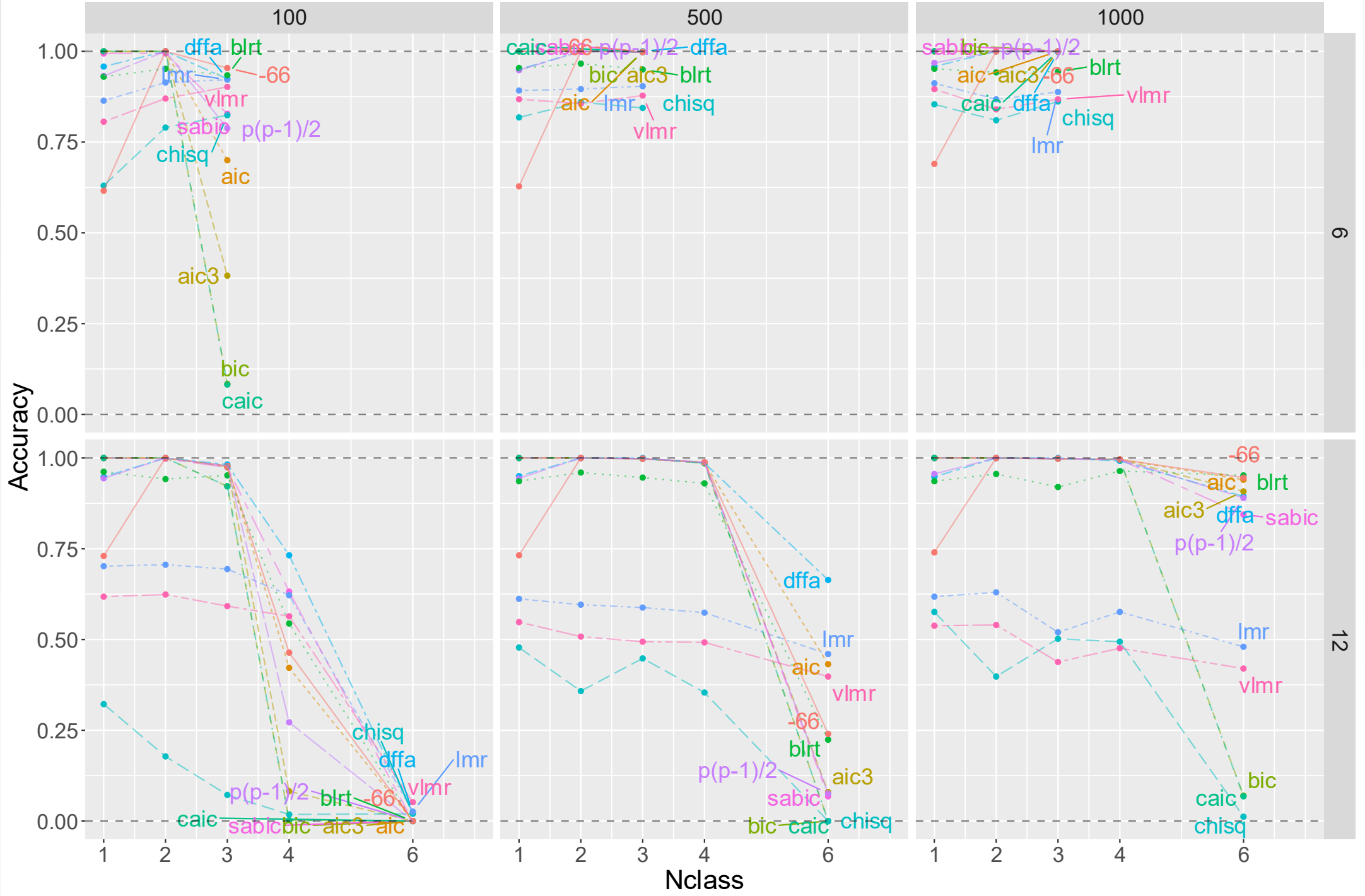
$$\chi_1^2 \sim \chi^2 - \frac{p(p-1)}{2}$$

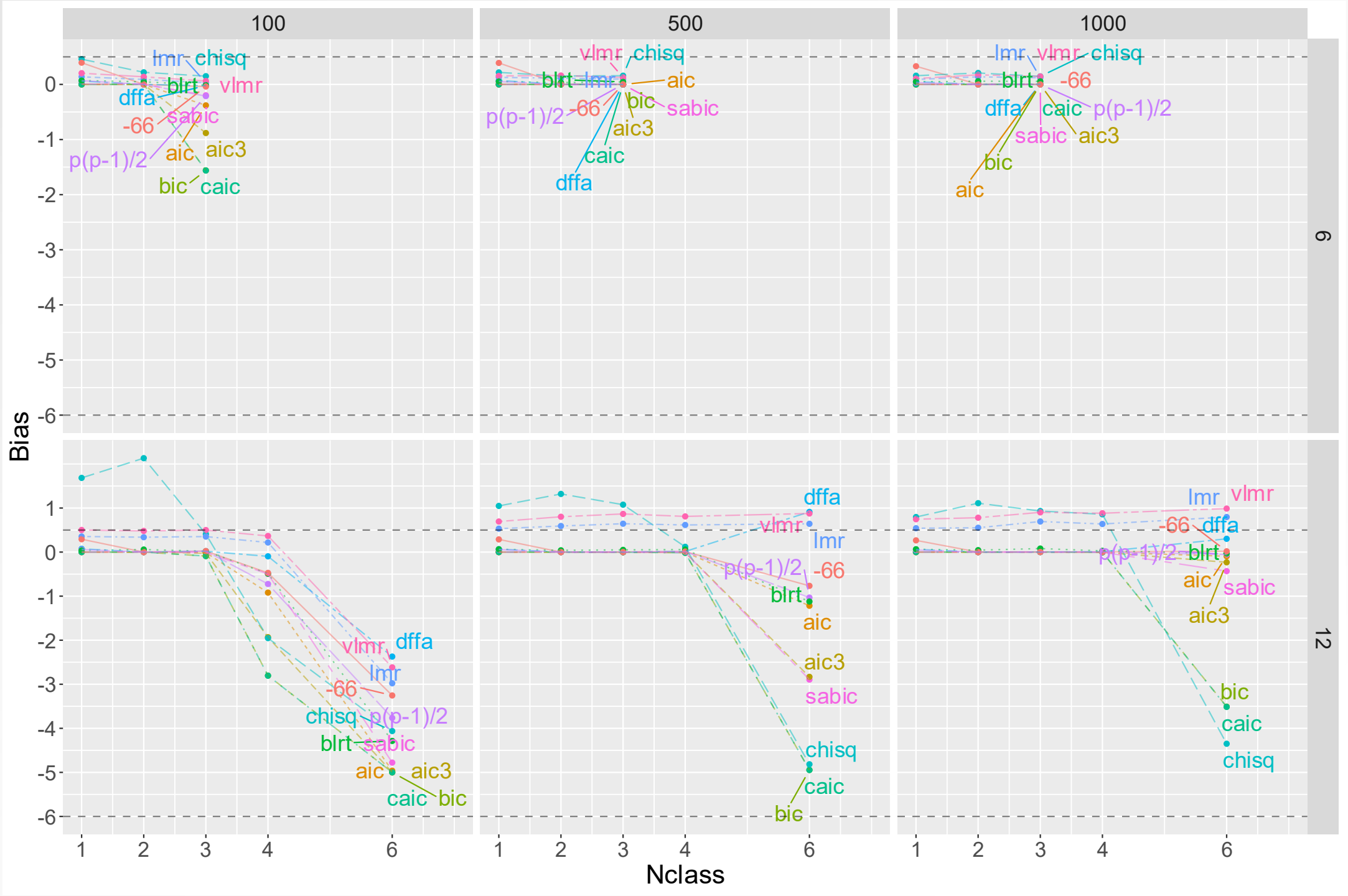
METHOD

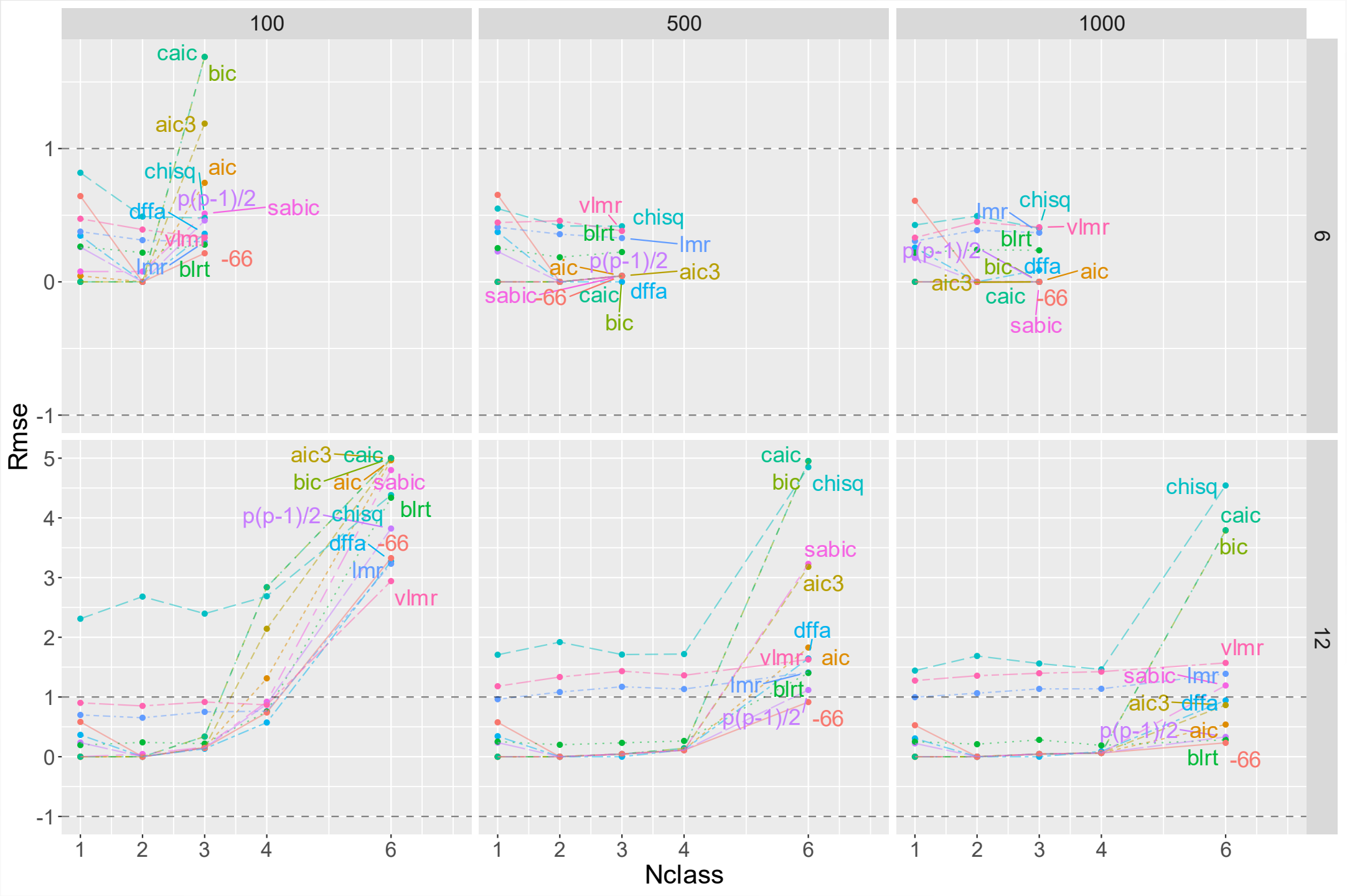
Simulations

- A Monte Carlo simulation was carried out to compare the accuracy, bias, and RMSE of three versions of this fit index with nine usual fit indices (AIC, BIC, saBIC, χ^2, CAIC, AIC3, Lo-Mendell-Rubin, Vuong-Lo-Mendell-Rubin, and the bootstrap LRT).
- Sample size (n) : 100, 500, and 1000
- Number of items : 6 and 12
- Equal class sizes
- Class' response probabilities = .8
- Each condition is repeated 500 times (blrt repeated 250 times)

RESULTS


100
500
1000
6
12
Accuracy
Nclass
1.00
0.75
0.50
0.25
0.00
1
2
3
4
6
dffa
blrt
lmr
-66
vlmr
sabic
p(p-1)/2
chisq
aic
aic3
bic
caic


100
500
1000
6
12
Bias
Nclass
lmr
chisq
vlmr
blrt
dffa
-66
sabic
p(p-1)/2
aic
aic3
bic
caic


100
500
1000
6
12
Rmse
Nclass
caic
bic
aic3
chisq
aic
p(p-1)/2
sabic
dffa
vlmr
lmr
blrt
-66

Results

- The simulation shows new among the three versions tested, two (dffa and -66) had very good properties: less bias and more accurate than other indices.
- The other one had very good accuracy but tended to narrowly miss the correct number of classes leading excessive over-extraction when it failed
- Overall, the new fit inex with dffa degree of freedom is promising

Conclusion

- Investigating the asymptotic properties of the underlying pseudo-χ^2 distribution, improving the current criteria and extending the index for ordinal scales.
- More simluations are needed to investigate a further ranges of conditions

Availability

Available in R

poLCAExtra package

with the function polca.cov()

Caron, P.-O. (2024). *poLCAExtra : New and Convenient Functions for the Package 'poLCA'*. https://github.com/quantmeth/poLCAExtra

References

- Bellemare-Lepage, A., Chatelois, M., & Caron, P.-O. (2023). Une introduction à l'analyse de classes latentes avec R. *The Quantitative Methods for Psychology, 19*(2), 217-229. https://doi.org/10.20982/tqmp.19.2.p217
- Caron, P.-O. (2024). *poLCAExtra : New and Convenient Functions for the Package 'poLCA'*. https://github.com/quantmeth/poLCAExtra
- Linzer, D. A., & Lewis, J. B. (2011). poLCA : An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*(10), 1-29. https://doi.org/ 10.18637/jss.v042.i10