The Impact of Item Scale Misspecification and Dichotomization on Class and Parameter Recovery in LCA of Count Data[1]

Kathryn S. Macia & Robert E. Wickham, Ph.D.
Palo Alto University, Palo Alto, CA

Simulation Study Overview

Research Question: How do misspecification of Poisson distributed indicators as continuous and normally distributed and different methods of dichotomization impact class enumeration, parameter estimates, and recovery of class assignments, and does the impact vary as a function of class separation and sample size?

Model: Latent class model with Poisson distributed indicators.

Simulation Conditions:
- Class Separation: High and Low
- Sample Size: *N* = 500, 1000, 2000

Analysis Conditions:
- Poisson (true distribution)
- Misspecified as continuous, normally distributed
- Dichotomized by median split
- Dichotomized by presence/absence of endorsement (0 vs. 1+)

Outcomes:
- Identification of the true number of classes
- Recovery of class assignments
- Parameter estimate bias

Background

- Finite mixture modeling is often used by substance use researchers to identify classes of individuals with distinct patterns of substance use (e.g., Haas, Wickham, Macia, Shields, & Macher, 2015; Tomczyk, Isensee, & Hanewinkel, 2015).
- A relatively common practice in these studies is to dichotomize indicators prior to mixture analysis.
  - A systematic review of articles published in addiction journals which conducted LCA/LPA on substance use count variables (*n* = 44) found that 65% dichotomized count indicators.
  - In all studies that did not dichotomize, there was no indication that the indicators were specified as following a count distribution in the analysis.
- Numerous articles in the GLM literature have warned against dichotomization (e.g., Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002).
- In addition, simulation studies have shown that ignoring skewness in continuous indicators can bias class enumeration and parameter estimates in mixture analysis (e.g., Asparouhov & Muthén, 2016).
- However, no research has examined how misspecification and dichotomization of count indicators may bias results of a mixture analysis.
- The purpose of this study was to examine how these practices (dichotomization and misspecification as continuous and normally distributed) may impact results obtained from a cross-sectional mixture analysis on Poisson distributed indicators.

Method

Data Simulation: 3 (sample size) x 2 (class separation) simulation design with 1000 replications requested for each condition.

- *Population model*: Latent class model with 7 Poisson distributed indicators and 4 classes.
- *Sample sizes*: *N* = 500, 1000, and 2000
- *Population parameter values*:
  - Latent class proportions were the same for all conditions.
  - Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) was used to create 2 sets of item parameter values (class means) with high and low class separation.

Item	Polysub (10%)	Pref Sub 1 (25%)	Pref Sub 2 (30%)	Low/No (35%)
	High Class Separation			
1	11.7	11.0	9.0	4.0
2	8.1	6.0	5.0	2.0
3	4.6	4.0	3.5	0.5
4	2.4	2.0	0.05	0.005
5	2.4	2.0	0.05	0.005
6	2.4	0.05	2.0	0.005
7	2.4	0.05	2.0	0.005
	Low Class Separation			
1	9.0	8.6	7.3	4.0
2	5.8	4.6	4.0	2.0
3	2.9	2.6	2.3	0.5
4	1.2	1.0	0.03	0.005
5	1.2	1.0	0.03	0.005
6	1.2	0.03	1.0	0.005
7	1.2	0.03	1.0	0.005

Data Analysis: Each simulated dataset was analyzed using LCA with 2-5 classes extracted under 4 item conditions:

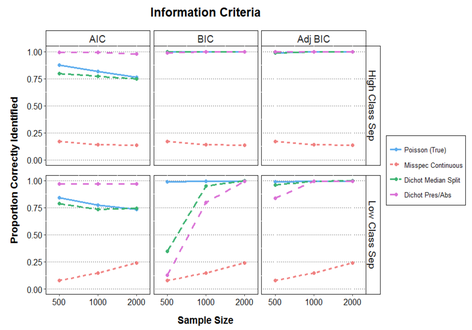
1) Specified as Poisson distributed (true distribution)
2) Misspecified as continuous and normally distributed
3) Dichotomized by median split, specified as ordered categorical
4) Dichotomized by presence/absence of endorsement, specified ordered categorical

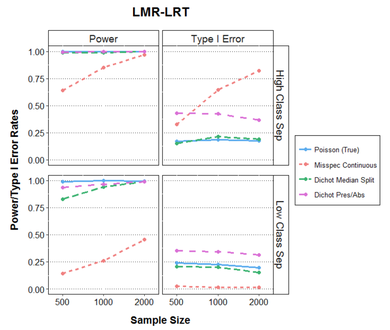
Outcomes:

- *Class Enumeration*:
  - Lo, Medell, and Rubin (2001) Likelihood Ratio Test (LMR-LRT) was used to calculate power and Type I error rates.
  - AIC, BIC, and Adjusted BIC were used to calculate proportion of replications that information criteria correctly identified the true number of classes.
- *Recovery of Class Assignments*:
  - Hubert and Arabie (1985) Adjusted Rand Index (ARI) was used to compare true class assignments to LCA-estimated class assignments. ARI is a chance-corrected measure where 1 = perfect agreement and 0 = chance level.
- *Parameter Estimate Bias*:
  - Percent bias was calculated as: $\left|\frac{\hat{\theta}-\theta}{\theta}\right| \times 100$

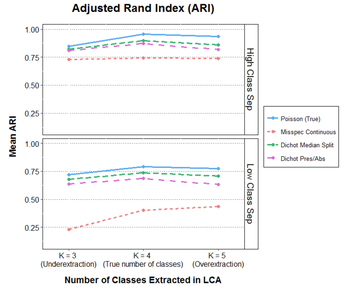
Results

Class Enumeration:





Recovery of Class Assignments:

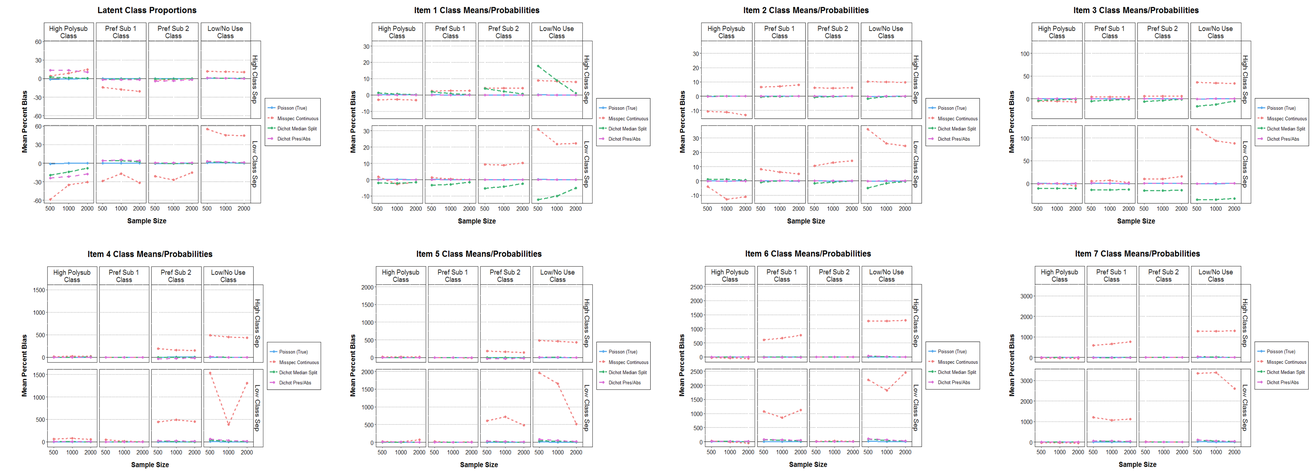


Note: ARI plot is marginalized over sample sizes since mean ARI did not vary by sample size.

ARI	η^2
Item Specif/Dichot	.93
# Classes Extracted	.83
Class Sep	.00
N	.00
Item S/D * Class Extr.	.47

Note: η^2 for interactions only shown if > .01.

Parameter Estimate Bias:



Percent Bias	η^2
Item Specif/Dichot	.15
Class	.17
Parameter	.20
Class Separation	.00
N	.00
Item S/D * Class	.17
Item S/D * Param	.22
Class * Param	.27
Item S/D * Class * Param	.30

Note: η^2 for interactions only shown if > .01.

Summary of Results

- Misspecifying count indicators as continuous and normally distributed had the most severe consequences for all areas of mixture recovery, particularly under low class separation.
- The 2 methods of dichotomizing showed fewer problems in recovery, but overall they showed worse recovery than the Poisson conditions. Most notably:
  - Dichotomizing by median split led to worse parameter recovery particularly for the low/no use class.
  - Dichotomizing by presence/absence of endorsement resulted in more elevated Type I error rates and greater parameter bias for parameters with population values near zero.
  - For both methods of dichotomization, there was worse class assignment recovery compared to the Poisson condition, and BIC was unable to accurately identify the true number of classes under low class separation and small sample size.
- Overall, when there were problems in mixture recovery, the low/no use class tended to have the most biased parameter estimates, followed by the smallest class (high polysub class).

Conclusion and Recommendations

- The distribution of substance use variables measured on a count scale should always be examined prior mixture analysis so that item distributions are specified appropriately.
- For optimal mixture recovery, count data should be analyzed using a count distribution; however, dichotomization may lead to similar results in some circumstances and much better mixture recovery than (incorrectly) assuming items are normally distributed.

References

Asparouhov, T., & Muthén, B. (2016). Structural equation models and mixture models with continuous non-normal skewed distributions. *Structural Equation Modeling*, *23*, 1-19.
Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249-253.
Haas, A.L., Wickham, R., Macia, K., Shields, M., & Macher, R. (2015). Identifying classes of conjoint alcohol and marijuana use in entering freshmen. *Psychology of Addictive Behaviors*, *29*, 620-626.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193-218.
Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79-86.
Lo, Y., Mendell, N.R., & Rubin, D.B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*, 767-778.
MacCallum, R.C., Zhang, S., Preacher, K.J., & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19-40.
Tomczyk, S., Isensee, B., & Hanewinkel, R. (2015). Latent classes of polysubstance use among adolescents–a systematic review. *Drug and Alcohol Dependence*, *160*, 12-29.

Correspondence concerning this poster should be sent by electronic mail to: kmacia@paloaltou.edu

[1]To cite this poster:
Macia, K. S., & Wickham, R. E. (2016). *The Impact of Item Scale Misspecification and Dichotomization on Class and Parameter Recovery in LCA of Count Data.* Poster presented at the sixth annual Modern Modeling Methods Conference, Storrs, CT.