Comprehensive Examination of Zero-Frequency Cell Correction Strategies in Tetrachoric Correlation Estimation

Jeongwon Choi Hao Wu
Vanderbilt University

Introduction

- Tetrachoric correlation estimates the relationship between two underlying continuous variables from observed binary variables, but zero-frequency cells in contingency tables can cause estimation issues, often resulting in estimates approaching boundary values of 1 or -1.
- Although many ways exist to correct zero-frequency cells, they have not been thoroughly investigated, with most studies focusing on adding 0.5 to zero cells when estimating a single correlation (e.g., Savalei, 2011; Yang & Weng, 2024).
- In this study, different zero-cell correction strategies were examined in the context of estimating single tetrachoric correlations, a correlation matrix, and factor loadings in confirmatory factor analysis.

Study 1: Estimating single tetrachoric correlations

- Zero-cell Correction Options

No Correction	
Correction:	
Added value: **0.1, 0.25, 0.5**	
Way to add the value:	
K	Add to zero cell and **K**eep margins
OA	Add **O**nly to zero cell and use **A**djusted thresholds
OU	Add **O**nly to zero cell and use **U**nadjusted thresholds
AA	Add to **A**ll cells when zero cells exist and use **A**djusted thresholds
AU	Add to **A**ll cells when zero cells exist and use **U**nadjusted thresholds
RA	Add to all cells **R**egardless of their existence and use **A**djusted thresholds
RU	Add to all cells **R**egardless of their existence and use **U**nadjusted thresholds

- Simulation Design
  Sample size: 50, 100, 200 | **Correlation:** 0.3, 0.5, 0.7, 0.9 | **Thresholds:** -1.5, -1.0, -0.8, 0.8, 1.0, 1.5
- **Evaluation Criteria:** RMSE, MAE, MAE of Fisher's z-transformed correlations, Non-coverage rates of the 95% Wald CI
- **Efficiency in Simulations**
  1. To eliminate simulation error, we considered all possible tables for each sample size (e.g., 23,426 tables for a sample of 50) and used their sampling probabilities as weights.
  2. To reduce the number of estimations, we grouped tables identical after reordering cells, and analyzing only one "prototype" table, as switching categories or variables leads to predictable changes (e.g., sign flips)
- **Results**: The choice between adjusted or unadjusted thresholds has minimal impact, and smaller added values tend to be more effective as correlations increase and thresholds move farther apart.

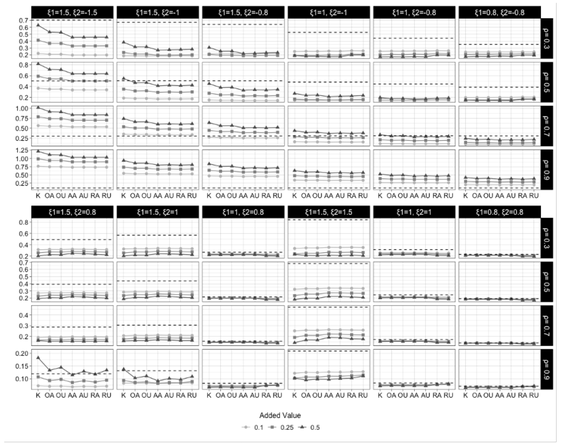

Figure 1. Mean Absolute Error (MAE) for Point Estimates of the Correlation

Study 2: Estimating Correlation Matrices

- Zero-cell Correction Options

No Correction	
Correction:	
Added value: **0.1, 0.25, 0.5**	
Way to add the value:	
K	Add to zero cell and **K**eep margins
O	Add **O**nly to zero cell
A	Add to **A**ll cells when zero cells exist
R	Add to all cells **R**egardless of their existence

- Simulation Design
  Sample size: 50 | **Number of Variables**: 6 | **Number of replications**: 30,000
  Correlation: Uniformly 0.4, Uniformly 0.8, Mix of both
  Thresholds: (1.5, 1.0, 0.8, 1.5, 1.0, 0.8), (1.5, 1.0, 0.8, -1.5, -1.0, -0.8)
  Evaluation Criteria: the number of positive definite correlation matrices, the average weighted squared error
- **Results**

1. A positive definite matrix is uncommon without correcting zero-frequency cells, and adding larger values generally increases the count of positive-definite matrices, with 0.5 added to the zero cell while keeping margins being the most effective strategy.

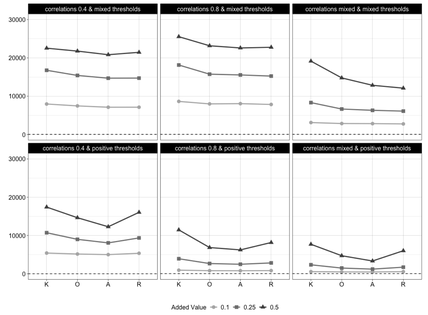

Figure 2. Number of Positive Definite Matrices

2. The average weighted squared error decreased with larger added values for positive thresholds. For mixed-signed thresholds, the lowest loss was achieved with the addition of 0.25 at correlations of 0.4, no correction at correlations of 0.8, and the addition of 0.5 for mixed correlations.

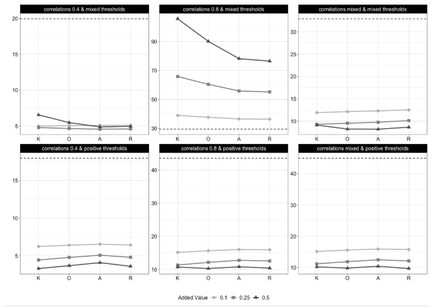

Figure 3. Average Weighted Squared Error

Study 3: Factor loadings in confirmatory factor analysis

- We examined the same zero-cell correction methods from Study 2, except for the 'No correction' option, in the context of a one-factor confirmatory factor analysis (CFA) for binary data.
- Simulation Design
  Sample size: 50 | **Number of Variables**: 4 | **Population Loadings**: 0.4, 0.7
  Thresholds: All positive or mixed signed values of 1.5, 1.0, 0.8 (6 in total)
  Number of replications: 1,000
  Estimation: Diagonally Weighted Least Square (DWLS)
  Evaluation Criteria: RMSE, bias of factor loadings
- **Results**

1. For most conditions with extreme thresholds, smaller values minimize bias. For less extreme thresholds, the optimal added value varies, with medium and large values being more effective.
2. Larger added values typically lower RMSE for positive thresholds, while smaller values generally reduce RMSE for mixed-sign thresholds.

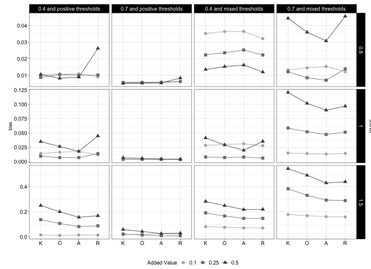

Figure 4. Bias of Loadings

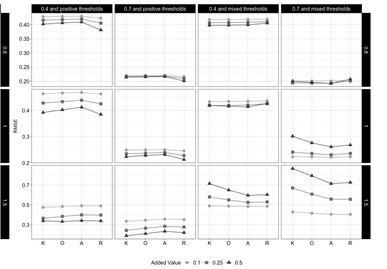

Figure 5. Root Mean Square Error (RMSE) of Loadings

Conclusion

- Different correction methods perform differently under distinct conditions and the size of the correlations, and the pattern of the thresholds moderate the results.
- Our examination across three scenarios suggests that no single approach works best in every situation and highlights the importance of considering the specific situation in which these corrections are applied.

References

- Savalei, V. (2011). What to do about zero frequency cells when estimating polychoric correlations. Structural Equation Modeling: A Multidisciplinary Journal, 18(2), 253–273.
- Yang, T.-R., & Weng, L.-J. (2024). Revisiting Savalei's (2011) research on remediating zero-frequency cells in estimating polychoric correlations: A data distribution perspective. Structural Equation Modeling: A Multidisciplinary Journal, 31(1), 81-96.

jeongwon.choi@vanderbilt.edu