Pairwise likelihood goodness-of-fit tests for factor models

IMPS 2023 @ University of Maryland

Haziq Jamil
Universiti Brunei Darussalam
https://haziqj.ml

27 July 2023

Joint work with Irini Moustaki (LSE)

Introduction

Context

Employ latent variable models (factor models) to binary data $Y_1, \ldots, Y_p$ collected from surveys via simple random or complex sampling.

(Psychometrics)
Behavioural checklist

(Education)
Maths achievement test

(Sociology)
Intergenerational support

Photo credits: @glenncarstenspeters, @ivalex, @oanhmj (Unsplash).

Introduction (cont.)

- Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top \in \{0, 1\}^p$ be a vector of Bernoulli rvs.
- The probability of observing a response pattern $\boldsymbol{y}_r = (y_{r1}, \ldots, y_{rp})^\top$, for any $r = 1, \ldots, R := 2^p$, is given by the joint distribution

$$\pi_r = \mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}_r) = \mathrm{P}(Y_1 = y_{r1}, \ldots, Y_p = y_{rp}). \tag{1}$$

Introduction (cont.)

- Let $\boldsymbol{Y} = (Y_1, \dots, Y_p)^\top \in \{0,1\}^p$ be a vector of Bernoulli rvs.
- The probability of observing a response pattern $\boldsymbol{y}_r = (y_{r1}, \dots, y_{rp})^\top$, for any $r = 1, \dots, R := 2^p$, is given by the joint distribution

$$\pi_r = \mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}_r) = \mathrm{P}(Y_1 = y_{r1}, \dots, Y_p = y_{rp}). \tag{1}$$

- Suppose $h = 1, \dots, N$ observations of $\boldsymbol{Y} = \boldsymbol{y}^{(h)}$ are recorded, and each unit h is assigned a (normalised) survey weight w_h with $\sum_h w_h = N$.
- Let $\hat{p}_r = \hat{N}_r / N$ be the rth entry of the R-vector of proportions $\hat{\boldsymbol{p}}$ with

$$\hat{N}_r = \sum_h w_h [\boldsymbol{y}^{(h)} = \boldsymbol{y}_r]. \tag{2}$$

Introduction (cont.)

- Let $\boldsymbol{Y} = (Y_1, \dots, Y_p)^\top \in \{0,1\}^p$ be a vector of Bernoulli rvs.
- The probability of observing a response pattern $\boldsymbol{y}_r = (y_{r1}, \dots, y_{rp})^\top$, for any $r = 1, \dots, R := 2^p$, is given by the joint distribution

$$\pi_r = \mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}_r) = \mathrm{P}(Y_1 = y_{r1}, \dots, Y_p = y_{rp}). \tag{1}$$

- Suppose $h = 1, \dots, N$ observations of $\boldsymbol{Y} = \boldsymbol{y}^{(h)}$ are recorded, and each unit h is assigned a (normalised) survey weight w_h with $\sum_h w_h = N$.
- Let $\hat{p}_r = \hat{N}_r / N$ be the rth entry of the R-vector of proportions $\hat{\boldsymbol{p}}$ with

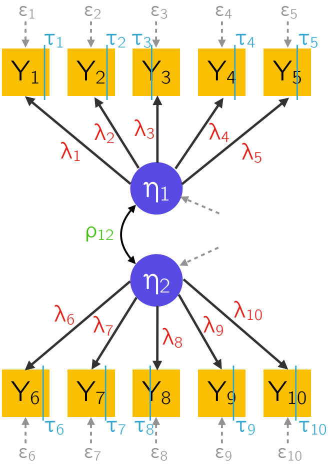
$$\hat{N}_r = \sum_h w_h [\boldsymbol{y}^{(h)} = \boldsymbol{y}_r]. \tag{2}$$

- Denote by $\boldsymbol{\pi}$ the R-vector of joint probabilities. It is widely known (Agresti, 2012) for IID samples that

$$\sqrt{N}(\hat{\boldsymbol{p}} - \boldsymbol{\pi}) \xrightarrow{\mathrm{D}} \mathrm{N}_R(\mathbf{0}, \boldsymbol{\Sigma}), \tag{3}$$

as $N \to \infty$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top$. This also works under complex sampling (Fuller, 2011), but $\boldsymbol{\Sigma}$ may take a different form.

Parametric models



- E.g. binary factor model with underlying variable approach (s.t. constraints)

$$
\begin{aligned}
Y_i &= \begin{cases} 1 & Y_i^* > \tau_i \\ 0 & Y_i^* \leq \tau_i \end{cases} \\
\boldsymbol{Y}^* &= \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon} \\
\boldsymbol{\eta} \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Psi}),\ & \boldsymbol{\epsilon} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Theta}_\epsilon)
\end{aligned} \tag{4}
$$

- The log-likelihood for $\boldsymbol{\theta}^\top = (\boldsymbol{\lambda}, \boldsymbol{\rho}, \boldsymbol{\tau})$ is

$$
\log L(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \sum_{r=1}^{R} \hat{N}_r \log \pi_r(\boldsymbol{\theta}) \tag{5}
$$

where $\pi_r(\boldsymbol{\theta}) = \int \phi_p(\boldsymbol{y}^* \mid \mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^\top + \boldsymbol{\Theta}_\epsilon)\, \mathrm{d}\boldsymbol{y}^*$.

- FIML may be difficult (high-dimensional integral; perfect separation).

Pairwise likelihood estimation

- For a pair of variables Y_i and Y_j, $i, j = 1, \dots, p$ and $i < j$, define

$$\pi^{(ij)}_{y_i y_j}(\boldsymbol{\theta}) = \mathrm{P}_{\boldsymbol{\theta}}(Y_i = y_i, Y_j = y_j), \qquad y_i, y_j \in \{0, 1\}. \tag{6}$$

There are $\tilde{R} = 4 \times \binom{p}{2}$ such probabilities, with $\sum_{y_i, y_j} \pi^{(ij)}_{y_i y_j}(\boldsymbol{\theta}) = 1$.

Pairwise likelihood estimation

- For a pair of variables Y_i and Y_j, $i, j = 1, \ldots, p$ and $i < j$, define

$$\pi^{(ij)}_{y_i y_j}(\boldsymbol{\theta}) = \mathrm{P}_{\boldsymbol{\theta}}(Y_i = y_i, Y_j = y_j), \qquad y_i, y_j \in \{0, 1\}. \tag{6}$$

  There are $\tilde{R} = 4 \times \binom{p}{2}$ such probabilities, with $\sum_{y_i, y_j} \pi^{(ij)}_{y_i y_j}(\boldsymbol{\theta}) = 1$.

- The pairwise log-likelihood takes the form (Katsikatsou et al., 2012)

$$\log \mathcal{L}_{\mathrm{P}}(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \sum_{i<j} \sum_{y_i} \sum_{y_j} \hat{N}^{(ij)}_{y_i y_j} \log \pi^{(ij)}_{y_i y_j}(\boldsymbol{\theta}), \tag{7}$$

  where $\hat{N}^{(ij)}_{y_i y_j} = \sum_h w_h [\boldsymbol{y}_i^{(h)} = y_i, \boldsymbol{y}_j^{(h)} = y_j]$.

Pairwise likelihood estimation

- For a pair of variables Y_i and Y_j, $i, j = 1, \ldots, p$ and $i < j$, define

$$\pi^{(ij)}_{y_i y_j}(\boldsymbol{\theta}) = \mathrm{P}_{\boldsymbol{\theta}}(Y_i = y_i, Y_j = y_j), \qquad y_i, y_j \in \{0, 1\}. \tag{6}$$

  There are $\tilde{R} = 4 \times \binom{p}{2}$ such probabilities, with $\sum_{y_i, y_j} \pi^{(ij)}_{y_i y_j}(\boldsymbol{\theta}) = 1$.
- The pairwise log-likelihood takes the form (Katsikatsou et al., 2012)

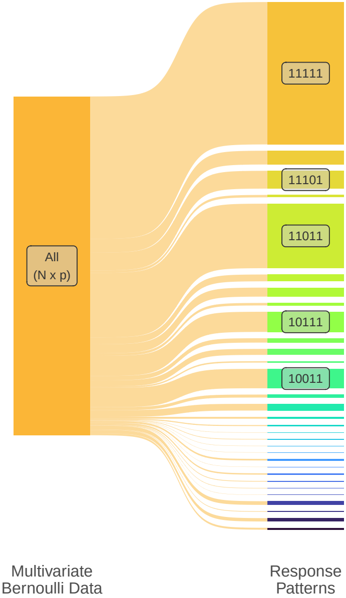
$$\log \mathcal{L}_{\mathrm{P}}(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \sum_{i<j} \sum_{y_i} \sum_{y_j} \hat{N}^{(ij)}_{y_i y_j} \log \pi^{(ij)}_{y_i y_j}(\boldsymbol{\theta}), \tag{7}$$

  where $\hat{N}^{(ij)}_{y_i y_j} = \sum_h w_h [\boldsymbol{y}^{(h)}_i = y_i, \boldsymbol{y}^{(h)}_j = y_j]$.
- Let $\hat{\boldsymbol{\theta}}_{\mathrm{PL}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{P}}(\boldsymbol{\theta} \mid \boldsymbol{Y})$. Under certain regularity conditions,

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{PL}} - \boldsymbol{\theta}) \xrightarrow{D} \mathrm{N}_m \left(\mathbf{0}, \left\{ \mathcal{H}(\boldsymbol{\theta}) \mathcal{J}(\boldsymbol{\theta})^{-1} \mathcal{H}(\boldsymbol{\theta}) \right\}^{-1} \right), \tag{8}$$

  where (Varin et al., 2011)
  - $\mathcal{H}(\boldsymbol{\theta}) = -\operatorname{E} \nabla^2 \log \mathcal{L}_{\mathrm{P}}(\boldsymbol{\theta} \mid \boldsymbol{Y})$ is the *sensitivity matrix*; and
  - $\mathcal{J}(\boldsymbol{\theta}) = \operatorname{Var}\left(\nabla \log \mathcal{L}_{\mathrm{P}}(\boldsymbol{\theta} \mid \boldsymbol{Y}) \right)$ is the *variability matrix*.

Goodness-of-fit (GOF)



- GOF tests are usually constructed by inspecting the fit of the joint probabilities $\hat{\pi}_r := \pi_r(\hat{\boldsymbol{\theta}})$.
- E.g.
  - LR: $X^2 = 2N \sum_r \hat{p}_r \log(\hat{p}_r / \hat{\pi}_r)$;
  - Pearson: $X^2 = N \sum_r (\hat{p}_r - \hat{\pi}_r)^2 / \hat{\pi}_r$,

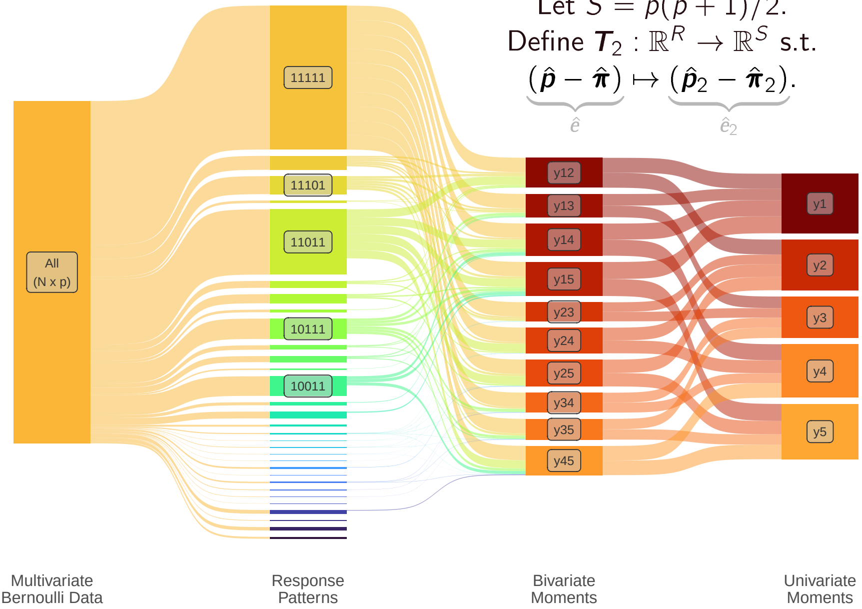
  These tests are asymptotically distributed as chi square.
- Likely to face sparsity issues (small or zero cell counts) which distort the approximation to the chi square.

Lower-order residuals

Let $S = p(p+1)/2$.

Define $\boldsymbol{T}_2 : \mathbb{R}^R \to \mathbb{R}^S$ s.t.

$$\underbrace{(\hat{\boldsymbol{p}} - \hat{\boldsymbol{\pi}})}_{\hat{e}} \mapsto \underbrace{(\hat{\boldsymbol{p}}_2 - \hat{\boldsymbol{\pi}}_2)}_{\hat{e}_2}.$$



Limited information GOF tests

- We show, via usual linearisation arguments, that as $N \to \infty$,

$$\sqrt{N}\hat{\boldsymbol{e}}_2 = \sqrt{N}\boldsymbol{T}_2\hat{\boldsymbol{e}} \xrightarrow{\text{D}} \text{N}_S(\boldsymbol{0}, \boldsymbol{\Omega}_2), \tag{9}$$

  where $\boldsymbol{\Omega}_2 = \big(\boldsymbol{I} - \boldsymbol{\Delta}_2\mathcal{H}(\boldsymbol{\theta})^{-1}\boldsymbol{B}(\boldsymbol{\theta})\big)\boldsymbol{\Sigma}_2\big(\boldsymbol{I} - \boldsymbol{\Delta}_2\mathcal{H}(\boldsymbol{\theta})^{-1}\boldsymbol{B}(\boldsymbol{\theta})\big)^\top$, and
  - $\boldsymbol{\Sigma}_2 = \boldsymbol{T}_2\boldsymbol{\Sigma}\boldsymbol{T}_2^\top$ (uni & bivariate multinomial matrix);
  - $\boldsymbol{\Delta}_2 = \boldsymbol{T}_2\big(\partial\pi_r(\boldsymbol{\theta})/\partial\theta_k\big)_{r,k}$ (uni & bivariate derivatives);
  - $\mathcal{H}(\boldsymbol{\theta})$ is the sensitivity matrix; and
  - $\boldsymbol{B}(\boldsymbol{\theta})$ is some transformation matrix dependent on $\boldsymbol{\theta}$.

Limited information GOF tests

- We show, via usual linearisation arguments, that as $N \to \infty$,

$$\sqrt{N}\hat{e}_2 = \sqrt{N}\boldsymbol{T}_2\hat{e} \xrightarrow{\text{D}} \text{N}_S(\mathbf{0}, \boldsymbol{\Omega}_2), \tag{9}$$

where $\boldsymbol{\Omega}_2 = \big(\boldsymbol{I} - \boldsymbol{\Delta}_2\mathcal{H}(\boldsymbol{\theta})^{-1}\boldsymbol{B}(\boldsymbol{\theta})\big)\boldsymbol{\Sigma}_2\big(\boldsymbol{I} - \boldsymbol{\Delta}_2\mathcal{H}(\boldsymbol{\theta})^{-1}\boldsymbol{B}(\boldsymbol{\theta})\big)^\top$, and

  - $\boldsymbol{\Sigma}_2 = \boldsymbol{T}_2\boldsymbol{\Sigma}\boldsymbol{T}_2^\top$ (uni & bivariate multinomial matrix);
  - $\boldsymbol{\Delta}_2 = \boldsymbol{T}_2\big(\partial\pi_r(\boldsymbol{\theta})/\partial\theta_k\big)_{r,k}$ (uni & bivariate derivatives);
  - $\mathcal{H}(\boldsymbol{\theta})$ is the sensitivity matrix; and
  - $\boldsymbol{B}(\boldsymbol{\theta})$ is some transformation matrix dependent on $\boldsymbol{\theta}$.

- From this, LIGOF test statistics generally take the quadratic form

$$X^2 = N\hat{e}_2^\top\hat{\boldsymbol{\Xi}}\hat{e}_2, \tag{10}$$

where $\boldsymbol{\Xi}(\hat{\boldsymbol{\theta}}) =: \hat{\boldsymbol{\Xi}} \xrightarrow{\text{P}} \boldsymbol{\Xi}$ is some $S \times S$ weight matrix. Generally, this is a chi square variate whose d.f. is either known or has to be estimated using moment matching (Maydeu-Olivares & Joe, 2005) or Rao and Scott (1979, 1981, 1984) adjustments.

Weight matrices

$$X^2 = N\hat{e}_2^\top \hat{\boldsymbol{\Xi}} \hat{e}_2$$
$$\sqrt{N}\hat{e}_2 \approx \text{N}_S(\mathbf{0}, \boldsymbol{\Omega}_2)$$

	Name	$\boldsymbol{\Xi}$	D.f.	Notes
1	Wald	$\boldsymbol{\Omega}_2^+$	$S - m$	Possible rank issues
2	Wald (VCF)	$\boldsymbol{\Xi}\boldsymbol{\Omega}_2\boldsymbol{\Xi}$	$S - m$	Need not est. $\boldsymbol{\Omega}_2$
3	Wald (Diag.)	$\text{diag}(\boldsymbol{\Omega}_2)^{-1}$	est.	Moment match, order 3
4	Wald (Diag., RS)	$\text{diag}(\boldsymbol{\Omega}_2)^{-1}$	est.	Rao-Scott, order 2
5	Pearson	$\text{diag}(\boldsymbol{\pi}_2(\boldsymbol{\theta}))^{-1}$	est.	Moment match, order 3
6	Pearson (RS)	$\text{diag}(\boldsymbol{\pi}_2(\boldsymbol{\theta}))^{-1}$	est.	Rao-Scott, order 2

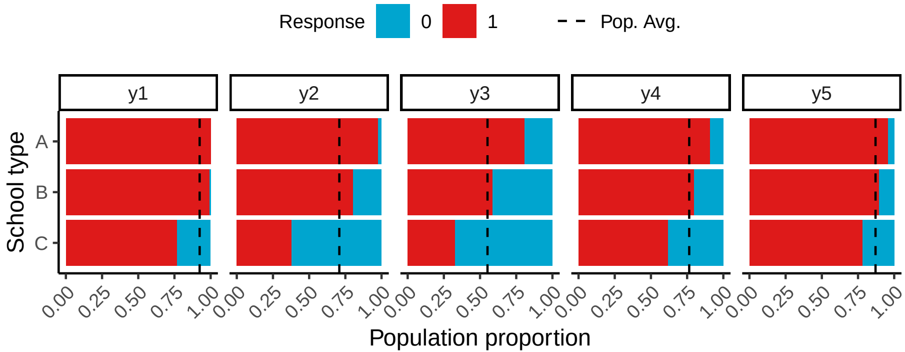
1–Reiser (1996); 2–Maydeu-Olivares and Joe (2005, 2006).

Setup

- $N \in \{500, 1000, 2000, 3000\}$ data were generated from a binary factor model with the following true parameter values:
  - Loadings: $\boldsymbol{\lambda} = (0.8, 0.7, 0.47, 0.38, 0.34, \dots)$
  - Factor correlations: $\rho = 0.3$ or $\boldsymbol{\rho} = (0.2, 0.3, 0.4)$
  - Thresholds: $\boldsymbol{\tau} = (-1.43, -0.55, -0.13, -0.82, -1.13, \dots)$
- Five scenarios considered
  1. 1 factor, 5 variables (1F 5V)
  2. 1 factor, 8 variables (1F 8V)
  3. 1 factor, 15 variables (1F 15V)
  4. 2 factor, 10 variables (2F 10V)
  5. 3 factor, 15 variables (3F 15V)
- For power analyses, models are intentionally misspecified by adding an extra, unaccounted for, latent variable in each scenario.
- Experiments repeated a total of $B = 1000$ times.

Complex design

Simulate a population of $1e6$ students clustered within classrooms and stratified by school type (correlating with abilities).

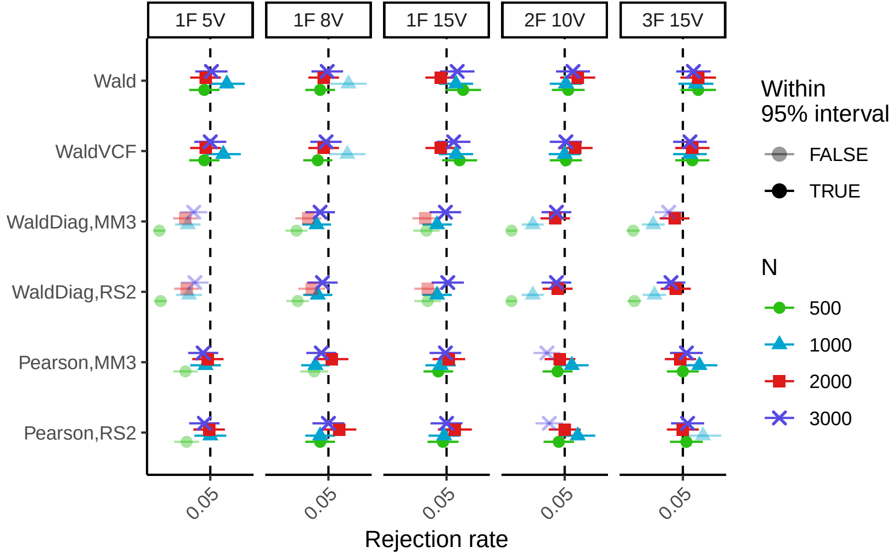


Multi-stage sampling: Sample n_S schools per strata via SRS, then sample 1 classroom via SRS, then select all students in classroom.

Other designs can be considered, e.g. cluster sampling or single-stage samples

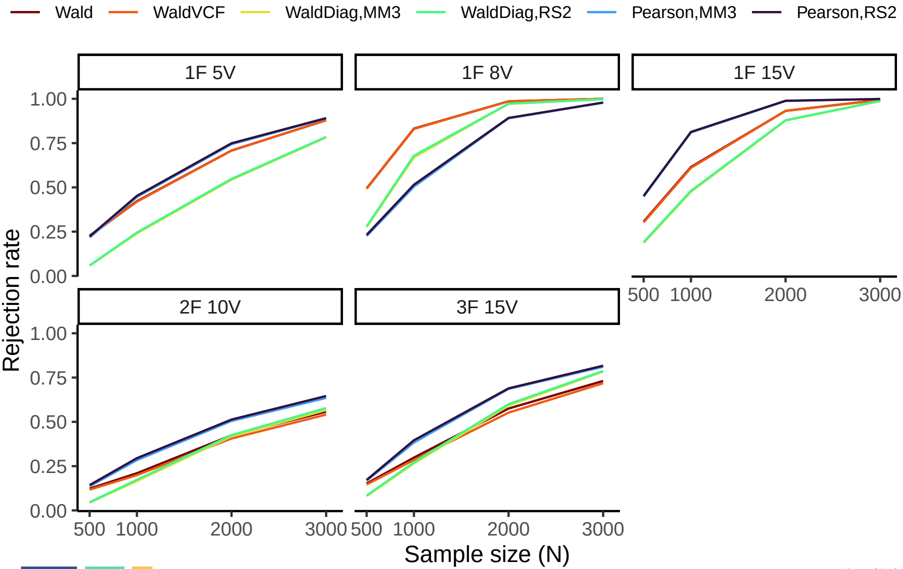
Results

SRS type I error rates ($\alpha = 5\%$)



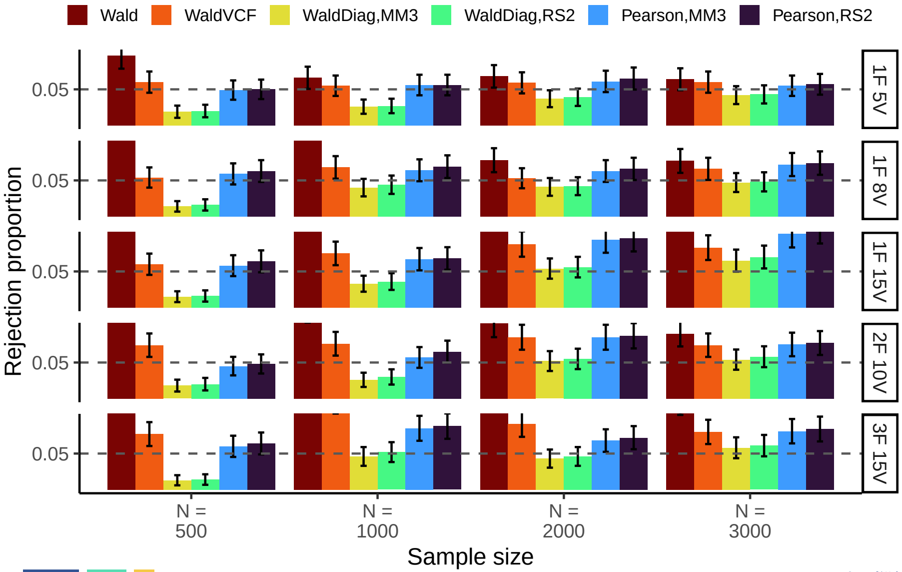
Results

SRS power analysis ($\alpha = 5\%$)



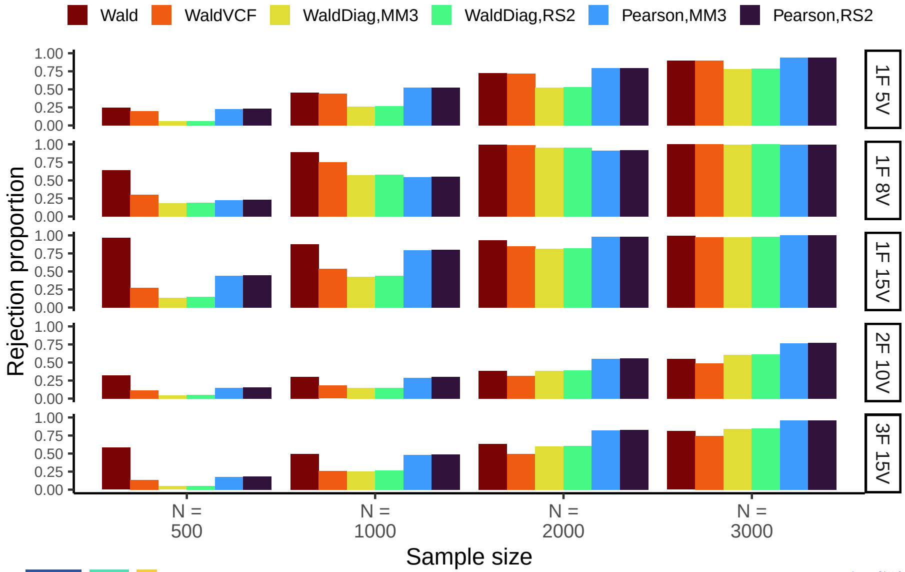
Results

Complex type I error rates ($\alpha = 5\%$)



Results

Complex power analysis ($\alpha = 5\%$)



Conclusions

- Pairwise likelihood estimation alleviates some issues associated with the UV approach in binary factor models.
- Sparsity impairs the dependability of GOF tests but are circumvented by considering lower order statistics.
- Wald-type and Pearson-type tests are investigated under simple random and complex sampling.
    - SRS: Wald and Pearson type tests generally perform as expected, but not the Diagonal Wald test.
    - Complex: Traditional Wald tests tend to give poor results, but our proposed Diagonal Wald test is more dependable.

Thanks!

Visit `haziqj.ml/lavaan.bingof` for further details and to try out our R package to implement these LIGOF tests.

References

Agresti, A. (2012). *Categorical data analysis* (Vol. 792). John Wiley & Sons.

Fuller, W. A. (2011). *Sampling statistics*. John Wiley & Sons.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012).Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243–4258.

Maydeu-Olivares, A., & Joe, H. (2005).Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006).Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713.

References

Rao, J. N. K., & Scott, A. J. (1979).Chi-squared tests for analysis of categorical data from complex surveys. *Proceedings of the American Statistical Association, section on Survey Research Methods*, 58–66.

Rao, J. N. K., & Scott, A. J. (1981).The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, *76*(374), 221–230.

Rao, J. N. K., & Scott, A. J. (1984).On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 46–60.

Reiser, M. (1996).Analysis of residuals for the multinomial item response model. *Psychometrika*, *61*, 509–528.

Varin, C., Reid, N., & Firth, D. (2011).An overview of composite likelihood methods. *Statistica Sinica*, 5–42.