Limited information goodness-of-fit tests for ordinal factor models

Haziq Jamil

Limited information approaches overcome sparsity issues and computational challenges in traditional goodness-of-fit tests. This paper describes the implementation of LIGOF tests for ordinal factor models that have been fitted using the `{lavaan}` package in R. The tests are computationally efficient and reliable, and adapted to suit whichever parameter estimation procedure was used to fit the model. The implementation is available as an R package called `{lavaan.ligof}`.

1 Introduction

2 Methods

2.1 Ordinal data

Consider the case of analysing multivariate data $\mathbf{y} = (y_1, \ldots, y_p)^\top$, where each item y_i is an ordinal random variable with m_i categories, $i = 1, \ldots, p$. Let $\mathcal{R} = \{\mathbf{c} = (c_1, \ldots, c_p)^\top \mid c_i \in \{1, \ldots, m_i\}\}$ be the set of all possible response patterns, and let $R = \prod_i m_i$ be the cardinality of this set. The joint probability of observing a response pattern $\mathbf{c}_r \in \mathcal{R}$ is given by

$$\pi_r = \Pr(\mathbf{y} = \mathbf{c}_r) = \Pr(y_1 = \mathbf{c}_{r1}, \ldots, y_p = \mathbf{c}_{rp}), \qquad r = 1, \ldots, R,$$

with $\sum_r \pi_R = 1$. Collect all response probabilities into the vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_R)^\top \in [0, 1]^R$. An example with $p = 3$, $m_1 = 2$, and $m_2 = m_3 = 3$ is given below. In total, there are $R = 2 \times 3 \times 3 = 18$ response patterns as shown in Table 1.

Later on we wish to use lower-order residuals to assess the fit of a model to the data, which first requires a description of lower-order moments and its connection to the joint response probabilities. Marginally, each y_i can be viewed as a multinoulli random variable with event probabilities $\pi_k^{(i)} = \Pr(y_i = k)$, $k = 1, \ldots m_i$, that sum to one. Therefore, this univariate distribution is characterised by its $(m_i - 1)$ *moments* $\pi_2^{(i)}, \ldots, \pi_{m_i}^{(i)}$, with the first moment

Table 1: Response patterns for $p = 3$ with $m_1 = 2$, and $m_2 = m_3 = 3$.

r	y_1	y_2	y_3	Pattern
1	1	1	1	111
2	1	1	2	112
3	1	1	3	113
4	1	2	1	121
5	1	2	2	122
6	1	2	3	123
7	1	3	1	131
8	1	3	2	132
9	1	3	3	133

r	y_1	y_2	y_3	Pattern
10	2	1	1	211
11	2	1	2	212
12	2	1	3	213
13	2	2	1	221
14	2	2	2	222
15	2	2	3	223
16	2	3	1	231
17	2	3	2	232
18	2	3	3	233

being redundant due to the sum to unity constraint. All univariate moments can be collected into the vector $\dot{\boldsymbol{\pi}}_1 = (\pi_k^{(i)})^\top$ whose dimension is $S_1 = \sum_i (m_i - 1)$. In a similar light, the bivariate distribution of (y_i, y_j) is characterised by its $(m_i - 1)(m_j - 1)$ *joint moments* $\pi_{k,l}^{(ij)} = \Pr(y_i = k, y_j = l)$, $k = 2, \ldots, m_i$, $l = 2, \ldots, m_j$. Also collect all bivariate moments into the vector $\dot{\boldsymbol{\pi}}_2 = (\pi_{k,l}^{(ij)})^\top$ whose dimension is $S_2 = \sum_{i<j} (m_i - 1)(m_j - 1)$. Finally, denote by $\boldsymbol{\pi}_2 = (\dot{\boldsymbol{\pi}}_1^\top, \dot{\boldsymbol{\pi}}_2^\top)^\top$ the vector of multivariate moments up to order 2, which is a vector of length $S = S_1 + S_2$.

Because the lower order moments are contained in the higher order moments, the vector $\boldsymbol{\pi}_2$ can be extracted from the joint probabilities $\boldsymbol{\pi}$ via a linear operation $\boldsymbol{\pi}_2 = \mathbf{T}_2 \boldsymbol{\pi}$ (Jamil et al., 2025). As an example, continuing from the $p = 3$ instance above, the moments for the first variable y_1, $\Pr(y_1 = 2)$ can be obtained by *summing* over all joint probabilities whose patterns contain $y_1 = 2$. The positions of these joint probabilities in the vector $\boldsymbol{\pi}$ are picked up by the first row of the matrix $\mathbf{T}_2$. Similarly, the two bivariate moments of (y_1, y_2), i.e. $\pi_{22}^{(12)}$ and $\pi_{23}^{(12)}$ are obtained by summing over the joint probabilities whose patterns contain $y_1 = 2$ and $y_2 = 2$, and $y_1 = 2$ and $y_2 = 3$, respectively.

Note that this construction of lower-order moments generalises to any order $q \leq p$, but the total number of moments up to order q grows combinatorially in both p and the category counts m_i, yielding design matrices $\mathbf{T}_q$ that can become computationally burdensome. Moreover, although we arbitrarily dropped the first moment in the foregoing construction, the choice of which category to omit is immaterial. This is because category probabilities sum to one, so excluding any one category produces a similar-dimensional parameterisation algebraically equivalent to excluding any other. For further details, consult Reiser (1996) and Maydeu-Olivares & Joe (2006).

	111	112	113	121	122	123	131	132	133	211	212	213	221	222	223	231	232	233
Y1=2	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Y2=2	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0
Y2=3	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1
Y3=2	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
Y3=3	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1
Y1=2,Y2=2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
Y1=2,Y2=3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
Y1=2,Y3=2	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0
Y1=2,Y3=3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1
Y2=2,Y3=2	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
Y2=2,Y3=3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Y2=3,Y3=2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
Y2=3,Y3=3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1

Figure 1: Matrix $\mathbf{T}_2$ for the case of $p = 3$ with $m_1 = 2$, and $m_2 = m_3 = 3$.

2.2 Confirmatory factor analysis

The confirmatory factor analysis (CFA) model imposes a structure on the joint response probabilities by assuming that the p observed variables are manifestations of a smaller set of q latent variables. In this way, the CFA may be viewed as a data-reduction technique since, effectively, the correlations among variables are modelled by a pre-specific factor structure using lower-dimensional data summaries.

CFA is typically used for continuous manifest variables, but it can also be applied to ordinal data. A common approach is the *underlying variable* (UV) approach, where the observed responses y_i are assumed to be discretised versions of continuous latent variables y_i^*. The connection is made through

$$y_i = \begin{cases} 1 & \tau_0^{(i)} < y_i^* < \tau_1^{(i)} \\ 2 & \tau_1^{(i)} < y_i^* < \tau_2^{(i)} \\ 3 & \tau_2^{(i)} < y_i^* < \tau_3^{(i)} \\ \vdots & \vdots \\ m_i & \tau_{m_i-1}^{(i)} < y_i^* < \tau_{m_i}^{(i)}, \end{cases}$$

with the *thresholds* $\tau_k^{(i)}$ for item i satisfying the ordering

$$-\infty \equiv \tau_0^{(i)} < \tau_1^{(i)} < \tau_2^{(i)} < \cdots < \tau_{m_i-1}^{(i)} < \tau_m^{(i)} \equiv +\infty.$$

Evidently, the model is invariant to a linear transformation, since scaling and shifting the underlying variables y_i^* do not affect the outcome of the ordinal variable y_i. For this reason

it is convenient to assume, for the purposes of model identifiability, a zero mean Gaussian distribution $\mathbf{y}^* \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{y}^*})$, where $\boldsymbol{\Sigma}_{\mathbf{y}^*}$ is a correlation matrix.

The underlying continuous variables $\mathbf{y}^*$, unlike their discrete counterparts $\mathbf{y}$, are now suitable to be modelled using a factor analysis model. Here, the goal is to find a set of latent factors $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^\top \in \mathbb{R}^q$, with $q \ll p$, that sufficiently explain the covariance structure of the p-dimensional variable space. This is achieved by the relationship

$$\mathbf{y}^* = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\Lambda}$ is a (often sparse) $p \times q$ matrix of factor loadings, and $\boldsymbol{\epsilon}$ is a vector of residuals. Certain distributional assumptions are made, namely that $\boldsymbol{\eta} \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi}$ a correlation matrix, $\boldsymbol{\epsilon} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Theta}_{\boldsymbol{\epsilon}})$ with $\boldsymbol{\Theta}_{\boldsymbol{\epsilon}} = \mathbf{I} - \mathrm{diag}(\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^\top)$, and that $\mathrm{Cov}(\boldsymbol{\eta}, \boldsymbol{\epsilon}) = \mathbf{0}$. Together, this implies that the polychoric correlation matrix of $\mathbf{y}$ is given by

$$\boldsymbol{\Sigma}_{\mathbf{y}^*} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^\top + \boldsymbol{\Theta}_{\boldsymbol{\epsilon}} \in \mathbb{R}^{p \times p}.$$

As a remark, the UV approach is commonly employed in the context of confirmatory factor analysis (CFA) models due to the ease of modelling, though other approaches such as item response theory (IRT) models are also available (Jöreskog & Moustaki, 2001).

For this factor analysis model, the parameters of interest are the non-zero entries $\boldsymbol{\lambda}$ of the loading matrix $\boldsymbol{\Lambda}$, the unique non-diagonal entries $\boldsymbol{\psi}$ in the factor correlation matrix $\boldsymbol{\Psi}$, and the thresholds $\boldsymbol{\tau}^{(i)} = (\tau_1^{(i)}, \ldots, \tau_{m_i-1}^{(i)})^\top$ for each ordinal item y_i. Collectively, these parameters are denoted by $\theta = (\boldsymbol{\lambda}^\top, \boldsymbol{\rho}^\top, \boldsymbol{\tau}^{(1)}, \ldots, \boldsymbol{\tau}^{(p)})^\top$ belonging to some parameter space Θ.

2.3 Parameter estimation

2.4 Distribution of residuals

2.5 Wald type tests

2.6 Pearson type tests

2.7 General GOF tests

2.8 Estimation of degrees of freedom

3 Usage

References

Jamil, H., Moustaki, I., & Skinner, C. (2025). Pairwise likelihood estimation and limited-information goodness-of-fit test statistics for binary factor analysis models under complex

survey sampling. *British Journal of Mathematical and Statistical Psychology*, *78*(1), 258–285. https://doi.org/10.1111/bmsp.12358

Jöreskog, K. G., & Moustaki, I. (2001). Factor Analysis of Ordinal Variables: A Comparison of Three Approaches. *Multivariate Behavioral Research*, *36*(3), 347–387. https://doi.org/10.1207/S15327906347-387

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

Reiser, M. (1996). Analysis of residuals for the multionmial item response model. *Psychometrika*, *61*(3), 509–528. https://doi.org/10.1007/BF02294552

4 Acknowledgements

I thank Rabi'ah Roslan for her diligent contributions as part of her undergraduate project and for the insightful discussions that helped shape this paper.