# Multinomial Goodness-of-Fit Statistics When the Number of Variables is Large

M.M.K Dassanayake* M. Reiser†

**Abstract**
The Pearson and likelihood ratio statistics are commonly used to test goodness-of-fit for models applied to count data from a multinomial distribution. When data are from a table formed by the cross-classification of a large number of manifest variables, the common statistics may have low power and inaccurate Type I error level due to sparseness. Several statistics defined on marginal distribution have been proposed to remedy this issue. Some of these statistics, fit to binary cross classified variables, have good performance for Type I error rate and power when the data table is formed from a moderate number of manifest variables. However, when the number of manifest variables becomes larger than 20, these statistics have limitations in terms of computer resources. This paper compares the performance of several Goodness-of-fit statistics for multinomial data when number manifest variables is larger than or equal to 25. The study will also investigate performance of a bootstrap method to obtain p-values for Pearson-Fisher statistic, fit to confirmatory dichotomous variable factor analysis model, when the number of manifest variables is larger than or equal to 25.

**Key Words:** Item response model, Sparseness, Bootstrap, Multinomial distribution

## 1. Introduction

In multinomial models we often consider the null hypothesis $H_o$: $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$, where $\boldsymbol{\pi}$ is a T-dimensional vector of multinomial probabilities, and $\boldsymbol{\pi}(\boldsymbol{\beta})$ is a vector of the multinomial probabilities as a function of parameters in the vector $\boldsymbol{\beta}$. When the model parameters $\boldsymbol{\beta}$ are unknown and estimated, the null hypothesis $H_o$: $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$ is often tested with the Pearson-Fisher statistic:

$$\chi^2_{PF} = \sum_s z_s^2, \tag{1.1}$$

where

$$z_s = \sqrt{n}(\pi_s(\hat{\boldsymbol{\beta}}))^{-\frac{1}{2}}\left(\hat{\mathrm{p}}_s - \pi_s(\hat{\boldsymbol{\beta}})\right)$$

and where, $\hat{\mathrm{p}}_s$ is element $s$ of $\hat{\mathbf{p}}$, vector of multinomial proportions, n is total sample size, $\hat{\boldsymbol{\beta}}$ parameter estimator vector, $\pi_s(\boldsymbol{\beta})$ is the expected proportion for cell $s$ and $\pi_s(\hat{\boldsymbol{\beta}})$ is the estimated expected proportion for cell $s$.

*School of Mathematical and Statistical Sciences, Arizona State University, P.O. Box 871804, Tempe, AZ 85287, USA

†School of Mathematical and Statistical Sciences, Arizona State University, P.O. Box 871804, Tempe, AZ 85287, USA

Under the large sample theory conditions, the Pearson-Fisher statistic has an asymptotic chi-square distribution with T-g-1 degrees of freedom where, $T$ is the number of cells and $g$ is the number of estimated model parameters (Koehler and Larantz, 1980). However, this assumption may not reasonable for analyzing a sparse table. A sparse table is one where there are many cells with small counts and/or zeros. When the data table is sparse, Pearson's chi-square statistic may have lower power and inaccurate Type I error.

Over the past years several statistics has been proposed to remedy this issue. Some of these statistics formed on lower-order marginals, fit to binary cross classified variables, have good performance for Type I error rate and power when the data table is formed from a moderate number of manifest variables. However, when the number of manifest variables becomes larger than 20, these statistics have limitations in terms of computer resources. This paper investigates the performance of the Tollenaar and Mooijaart (2003) statistic for multinomial data when number manifest variables is larger than or equal to 25. Mathematical details related to Tollenaar and Mooijaart (2003) statistic and the reasons for choosing this statistics over other statistics formed on lower-order marginals are given in the Section 3 and 4.

This study will also investigate performance of a bootstrap method to obtain p-values for Pearson-Fisher statistic, fit to confirmatory dichotomous variable factor analysis model, when the number of manifest variables is large.

## 2. Marginal Proportions

Traditional statistic such as Pearson's chi-square uses the joint frequencies to calculate goodness of fit for a model that has been fit to a cross-classified table. This section presents a transformation from joint proportions or frequencies to marginal proportions.

### 2.1 First- and Second-Order Marginals

The relationship between joint proportions and marginals can be shown by using zeros and 1's to code the levels of dichotomous response random variables, $Y_i, i = 1, 2, \ldots, q$, where $Y_i$ follow the Bernoulli distribution with parameter $P_i$ . Then, a $q$-dimensional vector of zeros and 1's, sometimes called a response pattern, will indicate a specific cell from the contingency table formed by the cross-classification of $q$ response variables. For dichotomous response variables, a response pattern is a sequence of zeros and 1's with length $q$. The $T = 2^q$-dimensional set of response patterns can be generated by varying the levels of the $q^{th}$ variable most rapidly, the $q^{th} - 1$ variable next, etc. Define $\boldsymbol{V}$ as the $T$ by $q$ matrix with response patterns as rows.

For instance when $q = 3$,

$$\boldsymbol{V} = \begin{pmatrix} 0\,0\,0 \\ 0\,0\,1 \\ 0\,1\,0 \\ 0\,1\,1 \\ 1\,0\,0 \\ 1\,0\,1 \\ 1\,1\,0 \\ 1\,1\,1 \end{pmatrix} .$$

Let $v_{is}$ represent element $i$ of response pattern $s$, $s = 1, 2, \ldots, T$. Then, under the model $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$, the first-order marginal proportion for variable $Y_i$ can be defined as

$$P_i(\boldsymbol{\beta}) = \text{Prob}(Y_i = 1|\boldsymbol{\beta}) = \sum_s v_{is}\pi_s(\boldsymbol{\beta}),$$

and the true first-order marginal proportion is given by

$$P_i = \text{Prob}(Y_i = 1) = \sum_s v_{is}\pi_s \, .$$

Under the model, the second-order marginal proportion for variables $Y_i$ and $Y_j$ can be defined as

$$P_{ij}(\boldsymbol{\beta}) = \text{Prob}(Y_i = 1, Y_j = 1|\boldsymbol{\beta}) = \sum_s v_{is}v_{js}\pi_s(\boldsymbol{\beta}),$$

where $j = 1, 2, \ldots, q-1$; $i = j+1, \ldots q$, and the true second-order marginal proportion is given by

$$P_{ij} = \text{Prob}(Y_i = 1, Y_j = 1) = \sum_s v_{is}v_{js}\pi_s \, .$$

### 2.2 Higher-Order Marginals

A general matrix $\mathbf{H}_{[t:u]}$ to obtain marginals of any order can be defined using Hadamard products among the columns of $\boldsymbol{V}$. The symbol $\mathbf{H}_{[t:u]}$ , $t \leq u \leq q$, denotes the transformation matrix that would produce marginals from order $t$ up to and including order $u$. Furthermore, $\mathbf{H}_{[t]} \equiv \mathbf{H}_{[t:t]}$ and $\mathbf{H} \equiv \mathbf{H}_{[t:u]}$ . $\mathbf{H}_{[1:q]}$ gives a mapping from joint proportions to the set of $(2^q - 1)$ marginal proportions:

$$\boldsymbol{P} = \mathbf{H}_{[1:q]}\boldsymbol{\pi} \, ,$$

where

$$\boldsymbol{P} = (P_1,\ P_2,\ P_3, \ldots P_q,\ P_{12},\ P_{13}, \ldots P_{q-1,q},\ P_{1,1,2} \ldots P_{q-2,q-1,q} \ldots P_{1,2,3 \ldots q})'$$

is the vector of marginal proportions. For example, when q=3,

$$\mathbf{H}_{[1:3]} = \begin{pmatrix} 0\,1\,0\,1\,0\,1\,0\,1 \\ 0\,0\,1\,1\,0\,0\,1\,1 \\ 0\,0\,0\,0\,1\,1\,1\,1 \\ \cdots \\ 0\,0\,0\,1\,0\,0\,0\,1 \\ 0\,0\,0\,0\,0\,1\,0\,1 \\ 0\,0\,0\,0\,0\,0\,1\,1 \\ \cdots \\ 0\,0\,0\,0\,0\,0\,0\,1 \end{pmatrix}.$$

Based on the above definition, second-order marginal proportions for variables $Y_i$ and $Y_j$ can also be obtained by,

$$\boldsymbol{P}_{[2]} = \mathbf{H}_{[2]}\boldsymbol{\pi} \tag{2.1}$$

where,

$$\mathbf{H}_{[2]} = \begin{pmatrix} (\boldsymbol{v}_1 \circ \boldsymbol{v}_2)' \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_3)' \\ \vdots \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_q)' \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_3)' \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_4)' \\ \vdots \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_q)' \\ \vdots \\ (\boldsymbol{v}_{q-1} \circ \boldsymbol{v}_q)' \end{pmatrix},$$

where $\boldsymbol{v}_f$ represents column $f$ of matrix $\boldsymbol{V}$, and $\boldsymbol{v}_f \circ \boldsymbol{v}_g$ represents the Hadamard product of columns $f$ and $g$.

## 3. Test statistics based on lower-order marginals

As indicated before, one way of remedying the problem of sparseness is to consider focused test statistics that are based on only the low-order marginals, which are sums of joint frequencies. Any statistic formed from a sum of the components, not necessarily ones based on marginal frequencies, can be considered a focused statistic. Summing a subset of components to create a focused test statistic could increase the power against certain alternatives. Focused tests using lower-order

marginals can be used in a wide variety of applications including log-linear models, categorical variable factor analysis and repeated measures on categorical variables.

Christoffersson (1975) first introduced the idea of using first- and second-order marginals for a test of fit in dichotomous variable factor analysis. Transforming to the notation in this paper, this statistic can be written as,

$$\chi^2_{Ch} = \bar{\mathbf{r}}'\mathbf{H}'_{[1:2]}(D(\mathbf{p}) - \mathbf{p}\mathbf{p}')^{-1}\mathbf{H}_{[1:2]}\bar{\mathbf{r}} \tag{3.1}$$

where $\bar{\mathbf{r}} = \hat{\mathbf{p}} - \boldsymbol{\pi}(\bar{\boldsymbol{\beta}})$, $\bar{\boldsymbol{\beta}}$ is the generalized least squares estimator of $\boldsymbol{\beta}$. $\chi^2_{Ch}$ has an asymptotic chi-square distribution with $2^q - g$ degrees of freedom, where $g$ = number of model parameters to be estimated. The statistic could be generalized to include higher-order marginals, but even if marginals from first- to order q were included, this statistic would not be equivalent to the Pearson-Fisher statistic. Muthén (1978) improved $\chi^2_{Ch}$ statistic, but both used observed proportions for the calculation of covariance matrix and neither presented their test as having higher power or as a remedy for sparse data.

Reiser(1996, 2008) and Reiser and Lin (1999) proposed statistics for $H_0 : \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$ that can be obtained from orthogonal components defined on marginal proportions. These statistics have higher power under some circumstances, and they usually perform well when applied to sparse frequency tables. Define the unstandardized residual $\mathrm{r}_s = \hat{\mathrm{p}}_s - \pi_s(\hat{\boldsymbol{\beta}})$, and denote the vector of unstandardized residuals as $\mathbf{r}$ with element $\mathrm{r}_s$.

$\sqrt{n}\,\mathbf{r}$ has asymptotic covariance matrix $\boldsymbol{\Omega}_{\mathbf{r}}$, where

$$\boldsymbol{\Omega}_{\mathbf{r}} = (D(\boldsymbol{\pi}(\boldsymbol{\beta})) - \boldsymbol{\pi}(\boldsymbol{\beta})\boldsymbol{\pi}(\boldsymbol{\beta})' - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{G}'),$$

and where

$$\begin{aligned} D(\boldsymbol{\pi}(\boldsymbol{\beta})) &= \text{diagonal matrix with } (s,s) \text{ element equal to } \pi_s(\boldsymbol{\beta}), \\ \mathbf{A} &= D(\boldsymbol{\pi}(\boldsymbol{\beta}))^{-1/2}\frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \\ \text{and } \mathbf{G} &= \frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \ . \end{aligned}$$

Then consider the linear combination $\boldsymbol{e} = \mathbf{H}\mathbf{r}$. If $\mathbf{H}$ contains $2^q - g - 1$ linearly independent rows corresponding to marginals from order 1 to $q$, then define the statistic

$$\chi^2_{[1:q]} = n\mathbf{r}'\mathbf{H}'\boldsymbol{\Omega}_{\boldsymbol{e}}^{-1}\mathbf{H}\mathbf{r}.$$

Here the statistic is evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is now consistent and efficient for $\boldsymbol{\beta}$, such as the maximum likelihood estimator, and where $\boldsymbol{\Omega}_{\boldsymbol{e}} = \mathbf{H}\boldsymbol{\Omega}_{\mathbf{r}}\mathbf{H}'$. With the added condition that the rows of $\mathbf{H}$ are linearly independent of the columns of $\mathbf{G}$, i.e., $\text{rank}(\mathbf{H}'\vdots\mathbf{G}) = T + g$, $\chi^2_{[1:q]}$ can be shown to be equivalent to $\chi^2_{PF}$. See Reiser (2008). To obtain orthogonal components, define the upper

triangular matrix $\boldsymbol{F}$ such that $\boldsymbol{F}'\boldsymbol{\Omega_e}\boldsymbol{F} = \boldsymbol{I}$. $\boldsymbol{F} = (\boldsymbol{C}')^{-1}$, where $\boldsymbol{C}$ is the Cholesky factor of $\boldsymbol{\Omega_e}$. Then writing $\boldsymbol{\Omega_e}$ as $\boldsymbol{C}\boldsymbol{C}'$,

$$\begin{aligned}\chi^2_{PF} &= n\mathbf{r}'\mathbf{H}'(\hat{\boldsymbol{C}}')^{-1}\hat{\boldsymbol{C}}'(\hat{\boldsymbol{C}}\hat{\boldsymbol{C}}')^{-1}\hat{\boldsymbol{C}}(\hat{\boldsymbol{C}})^{-1}\mathbf{H}\mathbf{r} \\ &= n\mathbf{r}'\mathbf{H}'\widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}'\mathbf{H}\mathbf{r}\end{aligned}$$

where $\widehat{\boldsymbol{F}}$ and $\hat{\boldsymbol{C}}$ are the matrices $\boldsymbol{F}$ and $\boldsymbol{C}$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

Define

$$\hat{\boldsymbol{\gamma}} = n^{\frac{1}{2}}\widehat{\boldsymbol{F}}'\mathbf{H}\mathbf{r} = n^{\frac{1}{2}}\widehat{\mathbf{H}}^*\mathbf{r}$$

Then

$$\chi^2_{PF} = \hat{\boldsymbol{\gamma}}'\hat{\boldsymbol{\gamma}} = \sum_{j=1}^{j=T-g-1}\hat{\gamma}_j^2,$$

and the elements $\hat{\gamma}_j^2$ are orthogonal components of $\chi^2_{PF}$. Since $\widehat{\mathbf{H}}^*\mathbf{r}$ has asymptotic covariance matrix $\boldsymbol{F}'\boldsymbol{\Omega_e}\boldsymbol{F} = \boldsymbol{I}_{T-g-1}$, the elements $\hat{\gamma}_j^2$ are asymptotically independent $\chi^2_1$ random variables.

By summing subset of these components one can obtain limited-information statistics. The statistic on first- and second-order marginals from Reiser (1996) is

$$\chi^2_{[1:2]} = \sum_{j=1}^{j=q(q+1)/2}\hat{\gamma}_j^2,$$

and the statistic on second-order marginals from Reiser and Lin (1999) is

$$\chi^2_{[2]} = \sum_{j=q+1}^{j=q(q+1)/2}\hat{\gamma}_j^2.$$

Joe (1993) and Maydeu-Olivares and Joe (2001, 2005, 2006)proposed a class of chi-square tests for sparse dichotomous and multidimensional data with applications to the item response model, a form of categorical variable factor analysis. Their approach is closely related to that of Reiser (1996) but their focused statistic $M_2$ does not correspond to the same decomposition of the $\chi^2_{PF}$. For $\mathbf{e} = \mathbf{H}_{[1:r]}\mathbf{r}$ and $\mathbf{r} = \hat{\mathbf{p}} - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$,

$$\mathbf{M}_r = \mathbf{e}'\hat{\mathbf{C}}_r\mathbf{e} \tag{3.2}$$

where $\hat{\mathbf{C}}_r = (\mathbf{H}\hat{\mathbf{T}}\mathbf{H}')^{-1} - (\mathbf{H}\hat{\mathbf{T}}\mathbf{H}')^{-1}\mathbf{H}\hat{\mathbf{G}}(\hat{\mathbf{G}}'\mathbf{H}'(\mathbf{H}\hat{\mathbf{T}}\mathbf{H}')^{-1}\mathbf{H}\hat{\mathbf{G}})^{-1}\hat{\mathbf{G}}'\mathbf{H}'(\mathbf{H}\hat{\mathbf{T}}\mathbf{H}')^{-1}$ and $\hat{\mathbf{T}} = D(\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})) - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})'$. $\mathbf{H}$ is always equal to $\mathbf{H}_{[1:r]}$ when applied to the definition of $\mathbf{M}_r$.

Tollenaar and Mooijaart (2003) proposed a statistic,

$$\chi^2_{red} = n\mathbf{e}'(\mathbf{H}_{[1:2]}\hat{\mathbf{T}}\mathbf{H}'_{[1:2]})^{-1}\mathbf{e} \tag{3.3}$$

where,

$$\hat{\mathbf{T}} = D(\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})) - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})'.$$

The Tollenaar and Mooijaart (2003) statistic is a reduced version of $\chi^2_{[1:2]}$ statistic (Reiser, 2008). The difference lies in the covariance matrix $\mathbf{T}$ not including the term $\mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{G}'$, where $\mathbf{G} = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ and $\mathbf{A} = \mathbf{D}(\boldsymbol{\pi})^{-1/2}\mathbf{G}$. As indicated by Tollenaar and Mooijaart (2003), omitting this term may substantially reduce computations. Since $\chi^2_{[1:2]}$ and $\chi^2_{red}$ have different covariance matrices, the degrees of freedom are different.

### 3.1 Application to Factor Analysis

When categorical manifest variables are hypothesized to be associated with a continuous latent variable, the model is known as categorical variable factor analysis and sometimes as the item response theory model. In order to investigate the challenges of a large number of variables and intense computations, a comparison of the statistics reviewed in the previous section will be presented using this model with one factor.

According to the categorical factor model, the probability of the response to a manifest variable, sometimes also referred to as an item, can be given by a logistic item response function:

$$P(Y_i = 1 \mid \boldsymbol{\beta}_i', X = x) = (1 + \exp(-\beta_{i0} - \beta_{i1}x))^{-1} \tag{3.4}$$

where $Y_i$ represents the response to item $i$,
$\beta_{i0}$ = intercept parameter for item $i$
$\beta_{i1}$ = slope parameter for item $i$
$\boldsymbol{\beta}_i' = (\beta_{0i}, \beta_{1i})$
$x$ = value taken on by latent random variable $X$

Since

$$P(Y_i = 0 \mid \boldsymbol{\beta}_i', X = x) = 1.0 - \pi(Y_i = 1 \mid \boldsymbol{\beta}_i', X = x),$$

it follows that

$$P(Y_i = y_i \mid \boldsymbol{\beta}_i', x) = P(Y_i = 1 \mid \boldsymbol{\beta}_i', x)^{y_i}[1.0 - P(Y_i = 1 \mid \boldsymbol{\beta}_i', x)]^{1-y_i}$$

It is assumed that, *conditional* upon the latent variable, responses to the manifest variables are independent. Let $\boldsymbol{Y}$ represent a random vector of responses to the items, with element $Y_i$, and let $\mathbf{y}$ represent a realized value of $\boldsymbol{Y}$. Then

$$P(\boldsymbol{Y} = \mathbf{y} \mid \boldsymbol{\beta}, x) = \prod_{i=1}^{k} \pi(Y_i = 1 \mid \boldsymbol{\beta}, x)^{y_i}[1 - \pi(Y_i = 1 \mid \boldsymbol{\beta}, x)]^{1-y_i} \tag{3.5}$$

$$\text{where } \boldsymbol{\beta} = \begin{pmatrix} \beta_{01} & \beta_{i1} \\ \beta_{02} & \beta_{12} \\ \beta_{03} & \beta_{13} \\ \vdots & \vdots \\ \beta_{0q} & \beta_{1q} \end{pmatrix}.$$

Finally, the probability of response pattern $s$, say, is obtained by taking the expected value of the conditional probability over the distribution of $X$ in the population, and is sometimes called the marginal probability:

$$\pi_s(\boldsymbol{\beta}) = \pi(\boldsymbol{Y} = \mathbf{y}_s \mid \boldsymbol{\beta}) = \int_{-\infty}^{\infty} \pi(\boldsymbol{Y} = \mathbf{y}_s \mid \boldsymbol{\beta}, x) f(x) dx \tag{3.6}$$

where $f(x)$ is the density function of X in the population of respondents.

If $\boldsymbol{U}$ represents a $T$-dimensional multinomial random vector of frequencies associated with the response patterns, the distribution of $\boldsymbol{U}$ is given by

$$\pi(\boldsymbol{U} = \mathbf{n}) = n! \prod_{s=1}^{T} \frac{[\pi_s(\boldsymbol{\beta})]^{n_s}}{n_s!} \tag{3.7}$$

$$\begin{aligned} \text{where } \mathbf{n} &= \text{vector of observed frequencies} \\ n_s &= \text{element } s \text{ of } \mathbf{n} \\ n &= \text{total sample size} = \sum_{s=1}^{T} n_s \end{aligned}$$

## 4. Feasibility of $\chi^2_{red}$ statistic when the number of manifest variables is large

Some popular statistics based on lower-order marginals have been discussed in Section 3. However, when the manifest variables exceed 20, most of these statistics will become difficult or impossible to calculate due to computer resource limitations. From these statistics, calculation of $\chi^2_{Ch}$ is fairly straightforward since the covariance matrix, $\Sigma_{\bar{\mathbf{r}}} = D(\mathbf{p}) - \mathbf{p}\mathbf{p}'$ can be calculated from the observed counts or proportions. Simulations reported by Reiser and VandenBerg (1994) show that chi-square approximation for the distribution of $\chi^2_{Ch}$ is valid only up to 8 to 10 variables for typical sample sizes. For larger number of variables the data table becomes very sparse and then $\Sigma_{\bar{\mathbf{r}}}$ is not a consistent estimator. On the other hand, $\chi^2_{[t:u]}$ tends to perform well under commonly encountered sparse situations, and has been calculated for up to 20 variables. However, calculating $\chi^2_{[t:u]}$ requires calculation of $\mathbf{G} = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ which requires $2 * 2^{q+1}$ integrals, where q is the number of manifest variables to be evaluated by numerical quadrature for the factor analysis model. Using SAS PROC IML, these calculations can be accomplished in random access memory for 20 manifest variables if 6 to 8 GB of RAM are available, for $\mathbf{G}, \mathbf{H}, \mathbf{A}, \boldsymbol{\pi}(\boldsymbol{\beta})$ and $\hat{\mathbf{p}}$, in approximately 4

minutes of CPU time (Reiser, 2012). If the calculations are done using virtual memory, reading and writing to disk, then processing time for 20 variables is on the order of 30 hours. With 25 manifest variables, these calculations can take up to 64 GB of RAM. On the other hand, Tollenaar and Mooijaart (2003) statistic, stated in Section 3 does not require calculation of $\mathbf{G} = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. The Tollenaar and Mooijaart (2003) statistic

$$\chi^2_{red} = n\mathbf{e}'(\mathbf{H}_{[1:2]}\hat{\mathbf{T}}\mathbf{H}'_{[1:2]})^{-1}\mathbf{e} \tag{4.1}$$

where,

$$\hat{\mathbf{T}} = D(\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})) - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})'$$

is a reduced version of $\chi^2_{[1:2]}$ statistic. It is a statistic for simple null hypothesis but with adjusted degrees of freedom for estimated parameters. The difference lies in the covariance matrix $\hat{\mathbf{T}}$, which does not include the term $\mathbf{G}(\hat{\mathbf{A}}'\hat{\mathbf{A}})^{-1}\mathbf{G}'$ in the $\chi^2_{red}$ statistic. This term represents variance due by estimating model parameters $\boldsymbol{\beta}$. As indicated by Tollenaar and Mooijaart (2003), omitting this term may substantially reduce computations. For instance, if the two parameter IRT model is fitted to 20 manifest variables, it requires $8*2^{20}*40$ bytes or 0.335 GB to store just the $\mathbf{G}$ matrix in SAS. With 25 variables, this amount will increase to $8 * 2^{25} * 50$ bytes or approximately 13.4 GB. Note that the two parameter IRT model contains both intercept and slope parameters, thus, it requires to take derivatives with respect to both intercept and slope. Hence, the $\mathbf{G}$ matrix will have $2q$ rows. The memory requirement when both the $\mathbf{A}$ matrix and $\mathbf{G}$ matrix are in memory is approximately 2*13.4 = 26.8 GB for 25 manifest variables. After calculation of the term $(\mathbf{A}'\mathbf{A})^{-1}$, the $\mathbf{A}$ matrix can be be discarded from the memory, which will save around 13.4 GB.

While $\chi^2_{red}$ does not require the term $\mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{G}'$, it still requires the $\mathbf{H}_{[1:2]}$ matrix, which becomes very large with a large number of manifest variables. For instance, with 20 manifest variables, it requires $8 * 2^{20} * 210$ bytes or approximately 1.76 GB to store $\mathbf{H}_{[1:2]}$ matrix. With 25 manifest variables this amount will increase up to 87.24 GB. This is a huge memory requirement for just one matrix, even with modern computer standards. One way to remedy this problem is to replace matrix operations with loops over vectors that consists of the rows of $\mathbf{H}$. Another technique that maybe useful for calculating the entire $\mathbf{H}$ matrix is sparse matrix operations. There are two aspects to sparse matrix techniques, namely, sparse matrix storage and sparse matrix computations. Typically, computer programs represent an $M$ by $N$ matrix in a dense form as an array of size $M$ by $N$, making row-wise and column-wise arithmetic operations particularly efficient to compute. However, if many of these $M$ by $N$ numbers are zeros, then correspondingly many of these operations are unnecessary or trivial. Sparse matrix techniques exploit this fact by representing a matrix not as a complete array, but as a set of nonzero elements and their location (row and column) within the matrix. This will be ideal for our case since not only observed proportions are sparse but also the $\mathbf{H}$ matrix is sparse. By combining these techniques we have created a program to calculate the $\chi^2_{red}$ statistic that can be used for a larger number of manifest variables. This program will not store the $\mathbf{H}_{[1:2]}$ matrix but rather generate the rows of $\mathbf{H}_{[1:2]}$ matrix at each element of $(\mathbf{H}_{[1:2]}\hat{\mathbf{T}}\mathbf{H}'_{[1:2]})$.

Therefore, to calculate the term $(\mathbf{H}_{[1:2]}\hat{\mathbf{T}}\mathbf{H}'_{[1:2]})$ of the $\chi^2_{red}$ statistic, this program only need to store two columns of $\mathbf{V}$ matrix to generate the second-order marginal $\mathbf{H}_{(2,i)}$ and another two columns of $\mathbf{V}$ matrix to generate the second-order marginal $\mathbf{H}_{(2,j)}$, where $j \geq i$ and i,j=1,...,q*(q-1)/2. Note, it also need to store the fitted proportions $\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$ and the vectors $\mathbf{H}_{(2,i)}$ and $\mathbf{H}_{(2,j)}$. Hence, by using this method for 25 manifest variables, it will only require $7*2^{25}$ bytes or approximately 0.2348 GB to generate the elements of $(\mathbf{H}_{[1:2]}\hat{\mathbf{T}}\mathbf{H}'_{[1:2]})$. This is huge memory saving compared to the 87.24 GB that is required to store just the $\mathbf{H}_{[1:2]}$ matrix for 25 manifest variables, but there will be a very large increase in number of loops. A brief description of the steps of this program are given as follows:

1. For each $l$ and $m$ create two corresponding columns of the $\mathbf{V}$ matrix, $l = 1, ..., q$ and $m = l + 1, ..., q$.
2. Do a element-wise multiplication of those two columns to obtain the second-order $\mathbf{H}_{(2,i)}$.
3. Use an embedded loop and create column $l$ and $n$ of the $\mathbf{V}$ matrix, where $n \geq m$.
4. Do a element-wise multiplication of the two columns in Step 3 to obtain second-order marginal $\mathbf{H}_{(2,j)}$, where $j \geq i$ and i,j=1,...,q*(q-1)/2.
5. Then, use the equation $\Sigma_{vec_p} = \sum_{i,j}\left((\mathbf{H}_{(2,i)} \circ \boldsymbol{\pi}(\hat{\boldsymbol{\beta}}) \circ \mathbf{H}_{(2,j)}) - \mathbf{H}'_{(2,i)} * \boldsymbol{\pi}(\hat{\boldsymbol{\beta}}) * \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})' * \mathbf{H}_{(2,j)}\right)$ to generate the $p^{th}$ element of the $\Sigma_{vec}$ where, $\Sigma_{vec}$ is the covariance matrix $(\mathbf{H}_{[1:2]}\hat{\mathbf{T}}\mathbf{H}'_{[1:2]})$ in vector form.
6. Use another loop over rows of $\mathbf{H}$ to obtain the vector $\mathbf{e}$ using the equation $\boldsymbol{e} = \mathbf{H}_{[1:2]}(\hat{\mathbf{p}} - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}}))$, where $\hat{\mathbf{p}}$ is the observed proportions. As in the Step 1 and 2, the loop is used to reduce the memory requirement of the $\mathbf{H}_{[1:2]}$ matrix. Calculation of the rows of the $\mathbf{H}_{[1:2]}$ matrix is similar to Step 1 and 2. For each element, rows of $\mathbf{H}_{[1:2]}$ will be multiplied by the vector $(\hat{\mathbf{p}} - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}}))$ to create the $r^{th}$ element of the vector $\mathbf{e}$, $r = 1, ..., q * (q - 1)/2$.
7. Use SQRVECH function in SAS to transform $\Sigma_{vec}$ into a symmetric square matrix, say $\hat{\Sigma}_{\chi^2_{red}}$.
8. Finally, use the equation $\chi^2_{red} = n\mathbf{e}'(\mathbf{H}_{[1:2]}\hat{\mathbf{T}}\mathbf{H}'_{[1:2]})^{-1}\mathbf{e} = n\mathbf{e}'(\hat{\Sigma}_{\chi^2_{red}})^{-1}\mathbf{e}$ to calculate the $\chi^2_{red}$ statistic.

The table below shows results for given observed and fitted probabilities for calculating $\chi^2_{red}$ in SAS using this method. Note, these results are for only one pseudo data set.

**Table 1:** Time and memory requirements for $\chi^2_{red}$

| No. of variables | Real time | User CPU time | System CPU time | Memory |
|---|---|---|---|---|
| 15 variables | 8.32 sec. | 6.81 sec. | 1.51 sec. | 0.0037 GB |
| 20 variables | 13 min 3 sec | 10 min 28 sec | 2 min 35 sec | 0.0996 GB |
| 25 variables | 19 min 21 sec | 14 min 5 sec | 5 min 16 sec | 3.15 GB |

Next, a Monte-Carlo simulation study was performed to test the performance of $\chi^2_{red}$ for 25 manifest variables. Due to the time limitations only Type I error study was performed. Empirical power study is recommended as a future work.

The design of Type I error study is as follows:

| | |
|---|---|
| Model (data generation) | categorical variable factor analysis model with one latent factor |
| Model (fitted) | categorical variable factor analysis model with one latent factor |
| Number of observed variables | q=25 |
| Number of simulation samples | 500 |
| Sample size | n=500 |

For the Monte-Carlo simulation study, data was generated from one factor IRT model. For the slope parameters of the model, pattern $(.1, .1, .1, 2.4, 2.4, 2.4, .2, .2)$ was repeated. Intercepts of the model were kept at zero. Result related to the simulation is given in the table below.

**Table 2:** Type I error results

| No. varaibles | Type I error rates |
|---|---|
| 25 var | 0.066 |

According to the results in the Table 2, the empirical Type I error rates are within $0.05 \pm 1.96 * \sqrt{0.05 * 0.95/n}$. Note, with 25 manifest variables and sample size, n=500 there can be spareness even in the $2 * 2$ sub-table. Yet, the above Type I error results indicates $\chi^2_{red}$ has good performance for Type I error rate even when the number of manifest variables are large as 25.

## 5. Bootstrap method

The section will introduce a bootstrap method to obtain p-values for Pearson-Fisher statistic, fit to confirmatory dichotomous variable factor analysis model when the number of manifest variables is large.

When there are 25 manifest variables, the cross-classified table has $2^{25}$, or 33,554,432 cells. If the sample size for testing the fit of a model is a few hundred observations, then the data table will be sparse and many cells will have counts of zero or 1. As discussed in the previous sections, when the data are sparse, the asymptotic chi-square approximation for the distribution of the Pearson and likelihood ratio statistics may not be valid. Extensive simulations have also shown that p-values obtained from the chi-square distribution for a test of the categorical factor analysis model on a sample of size 1000 start to become unreliable at about 6 to 8 manifest variables, depending on the skew of distribution of the frequencies (Reiser and VandenBerg, 1994).

Not only sparseness, but also computer resources become an issue when the number of manifest variables exceeds 20. There are limits on individual objects statistical software can store. For

example, having 30 manifest variables would require approximately $8 * 30 * 2^{30}$ bytes or 257.6 GB to store the $\mathbf{H}$ matrix in R or SAS assuming double precision storage. If the interest is to store only observed probabilities and fitted probabilities, with 30 manifest variables it will only require approximately 16 GB. Due to these reasons most of the simulations found in the literature are limited to 20 manifest variables. But, in an application such as educational testing, the number of manifest variables could be 50 or more, and with 50 manifest variables, it will require $8 * 2^{50}$ bytes or 9,007,199.25 GB to store the fitted probabilities.

We will introduce the following method using the omnibus $\chi^2_{PF}$ statistic to overcome these issues. Calculation of the Pearson statistic itself does not necessarily encounter memory limits for large number of manifest variables because the contribution of each cell can be calculated individually and cumulated. Processing requirements of $\chi^2_{PF}$ are not a concern for 30 or more variables because calculation of $\boldsymbol{\pi}_s(\hat{\boldsymbol{\beta}})$ is required only for the cells where $n_s > 0$, and even with a large number of manifest variables, the number of cells where $n_s > 0$ can be no more than the sample size. The contribution for the cells with $n_s = 0$ is equal to $n \sum_s I(n_s = 0)\pi_s(\hat{\boldsymbol{\beta}})$ and can be obtain by subtraction since,

$$\sum_s I(n_s > 0)\pi_s(\hat{\boldsymbol{\beta}}) + \sum_s I(n_s = 0)\pi_s(\hat{\boldsymbol{\beta}}) = 1 \tag{5.1}$$

where, $I$ is the indicator function. Calculation of $\chi^2_{[2]}$, for example, requires much more storage. Since computational requirements may not present a barrier, obtaining p-values for $\chi^2_{PF}$ by using the parametric bootstrap may be feasible even for a very large number of variables. The theory of the parametric bootstrap is quite similar to that of the nonparametric bootstrap, the only difference is that instead of simulating bootstrap samples that are independent and identically distributed (iid) from the empirical distribution (the nonparametric estimate of the distribution of the data) the parametric bootstrap procedure simulates bootstrap samples that are iid from the estimated parametric model.

The method that is introduce here will require only the observed patterns and hence less memory requirement. A brief description of the steps of this method are given as follows:

1. Assume $\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$ is true. The model $\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$ could be any categorical variable model.

2. Treat the fitted proportions $\boldsymbol{\pi}_s(\hat{\boldsymbol{\beta}})$ under the model as population proportions.

3. Draw random samples from the multinomial distribution with these fitted proportions as parameters of the distribution.

4. For each sample, estimate the categorical variable model used in Step 1. For a instance, if the IRT model was used in Step 1 to get $\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$ then, IRT model will be estimated for each sample from Step 3.

5. If $n_s > 0$, use multivariate Gaussian quadrature to obtain the expected proportions and calculate $\chi^2_{PF_{n_s>0}}$.

6. If $n_s = 0$, use the equation 5.1 to obtain $\chi^2_{PF_{n_s=0}}$.
7. Sum $\chi^2_{PF_{n_s=0}}$ and $\chi^2_{PF_{n_s>0}}$ to obtain $\chi^2_{PF}$.
8. Repeat step 5,6 and 7 for each sample.
9. Obtain p-value by calculating the proportion of $\chi^2_{PF}$ values from bootstrap samples that are greater than the $\chi^2_{PF}$ value from the original sample.

In order to evaluate the performance of this method, Type I error study was performed. Note, the $\chi^2_{PF}$ is an omnibus test that gives little guidance of the source of poor fit and can be outperformed by focused or directional tests of lower-order.

The design of Type I error study is as follows:

| | |
|---|---|
| Model (data generation) | categorical variable factor analysis model with one latent factor |
| Model (fitted) | categorical variable factor analysis model with one latent factor |
| Number of observed variables | q=8, q=15, q=18, q=20, q=25 |
| Number of simulation samples | 1000 |
| Sample size | n=500 |
| Number of bootstrap samples | 500 |

A Monte-Carlo simulation studies were performed with the information described in the Table above. One thousand data sets were generated from the one factor model. For the slope parameters of the one factor model, the pattern $(.1, .1, .1, 2.4, 2.4, 2.4, .2, .2)$ was repeated. Intercepts of the model were kept at zero. After generating the data, a two-parameter IRT model was estimated for each of these data sets and Type I error rate related to the $\chi^2_{PF}$ was calculated. Results related to 8, 15, and 20 variables are given in the Table below.

**Table 3:** Type I error rates comparison

| No. varaibles | Bootstrap Method | Mplus(MonteCarlo) |
|---|---|---|
| 8 var | 0.046 | 0.042 |
| 15 var | 0.044 | 0.161 |
| 20 var | 0.342 | 0.380 |

**Table 4:** Time requirements for the Bootstrap method

| No. varaibles | Time (in sec) |
|---|---|
| 8 var | 29 |
| 15 var | 68 |
| 20 var | 360 |

According to the results in the Table 3, for moderately large number of manifest variables, the bootstrap method performed well in terms of Type I error rates. When the number of manifest variables exceeds 20, the Type I error rates started to inflate. However, we believe the Type I error rates can be improved by increasing the number of bootstrap samples. Due to the time limitations we had to restrict our simulation to 500 bootstrap samples.

## 6. Discussion, Drawbacks and Future work

In this study we have investigated performance of the Tollenaar and Mooijaart (2003) $\chi^2_{red}$ statistics when the number manifest variables is large. Results indicate $\chi^2_{red}$ has good performance for Type I error rate even when the number of manifest variables as large as 25. One of the other goals of this research was to create memory and time efficient program to calculate goodness-of-fit statistics for large number of variables. The program that we have created improved the memory consumption. The largest amount of RAM the program consumed during the calculation of the Tollenaar and Mooijaart (2003) statistics was 3.15 GB. However, the number of loops this program require thus the computer time increased rapidly with $q$. For instance, 15 manifest variables would require $105 * (106/2) = 5,565$ loops to calculate components of the matrix $(\mathbf{H}_{[1:2]}\hat{\mathbf{T}}\mathbf{H}'_{[1:2]})$ and $15 * (14/2) = 105$ loops to calculate the $\mathbf{e}$ vector. Similarly, 20 manifest variables would require $20*(19/2)+190*(191/2) = 18,335$ loops and 25 manifest variables would require $25*(24/2)+300 * (301/2) = 45,450$ loops. Therefore, the drawback of the this method is the large number of loops and cpu time.

Performance of a bootstrap based method to obtain p-values for Pearson-Fisher statistic was also investigated when the number of manifest variables is large. For moderately large number of manifest variables, the bootstrap method performed well in terms of Type I error rates. When the number of manifest variables exceeds 20, the Type I error rates started to inflate. This might be due to the small number of bootstrap samples used in the simulations study. Therefore, as a future work, we suggest to increase the number of bootstrap samples to 2000 or more. The main issue that we encountered with the bootstrap method is it requires $2^q$ expected probabilities to generate the bootstrap samples. When the number of manifest variables increases this may cause computer resource limitations. For a instance, 30 manifest variables would require calculation of $2^{30} = 1,073,741,824$ expected probabilities.

## REFERENCES


Agresti, A., & Yang, M. (1987). An empirical investigation of some effects of sparsness in contingency tables. *Computational Statistics and Data Analysis, 5* , 9-21.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse $2^P$ contingency tables. *British Journal of Mathematical and Statistical Psychology, 55*, 1-15.

Bartholomew, D.J., Knott, M. & Moustaki (2011). *Latent variable models and factor analysis: A unified approach, 3rd Edition.* New York: Wiley.

Birch, M. W. (1964). A new proof of the Pearson-Fisher Theorem. *Annals of Mathematical Statistics, 35*, 818-824.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179-197.

Cagnone, S., & Mignani, S. (2007). Assessing the goodness of fit of a latent variable model for binary data. *Metron*, LXV, 337-361.

Cai, L., Maydeu-Olivares, A., Coffman, D.L., & Thissen, D. (2006). Limited information goodness of fit testing of item response theory models for sparse 2p tables. *British Journal of Mathematical and Statistical Psychology*, 59, 173-194.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40* , 5-32.

Dassanayake, M.K., and M. Reiser. 2015. Lack-of-Fit Diagnostics Based on Standardized Residuals and Orthogonal Components of Pearson's Chi-Square. In JSM Proceedings, Social Statistics Section. Alexandria, VA: American Statistical Association. 614-628.

Goodnight, J. H. (1978). The sweep Operator: Its importance in Statistical Computing. SAS Technical Report R-106, SAS Institute, Cary, NC.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics, 29*, 205-220.

Joreskog & Moustaki (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, 36, 347-387.

Koehler, K., & Larantz, K. (1980). An emperical investigation of goodness-of fit statis- tics for sparse multinomials. *Journal of the American Statistical Association, 75* , 336-344.

Maydeu-Olivares, A. (2001). Limited information estimation and testing of thurstonian models for paired comparison data under multiple judgement sampling. *Psychometrika, 66*, 209-228.

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*, 1009-1020.

Maydeu-Olivares, A. & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71, 713-732.

Maydue-Olivares, A., Garcia-forero, c., Gallardo-Pujol, D., & Renom, J. (2009). Testing categorized bivariate normality with two-stage polychoric correlation estimates. *Methodology: European Journal of Research Methods for the behavioral and social sciences,* 5, 131-136.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43, 551-560.

Muthén, L.K & Muthén, B.O. (1998-2010). Mplus User's Guide. Sixth edition. Los Angeles, CA.

Rayner, J. C. W., & Best, D. J. (1989). *Smooth Tests of Goodness of Fit.* Oxford: New York.

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

Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology*, 61(2), 331-360.

Reiser, M., & Lin, G. (1999). A goodness-of-fit test for the latent class model when expected frequencies are small. In M. Sobel & M. Becker (Eds), *Sociological Methodology 1999*, 81-111. Boston: Blackwell.

Takane, Y., and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 292-408.

Tollenaar, N., & Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology, 56*, 271-288.