LIMITED INFORMATION GOODNESS-OF-FIT TESTING IN MULTIDIMENSIONAL CONTINGENCY TABLES*

IE Working Paper MK8-107-I 15-02-2005

Alberto Maydeu

Instituto de Empresa
Serrano 105
28006 Madrid, Spain
alberto.maydeu@ie.edu

Harry Joe

Dept. of Statistics
University of British Columbia
333-6356 Agricultural Rd.
Vancouver, BC, V6T 122-Canada
harry@stat.ubc.ca

Abstract

We introduce a family of goodness-of-fit statistics for testing composite null hypotheses in multidimensional contingency tables of arbitrary dimensions. These statistics are quadratic forms in marginal residuals up to order r. They are asymptotically chi-square under the null hypothesis when parameters are estimated using any consistent and asymptotically normal estimator. We show that when r is small (r = 2) the proposed statistics have more accurate empirical Type I errors and are more powerful than Pearson's X^2 for a widely used item response model. Also, we show that the proposed statistics (but not X^2 even for the maximum likelihood estimate) are asymptotically chi-squared under the null hypothesis when applied to subtables.

Keywords

multivariate discrete data, categorical data analysis

Depósito Legal: M-20073 I.S.S.N.: 1579-4873

* This research has been supported by grant BSO2003-08507 of the Spanish Ministry of Science and Technology, a Distinguished Young Investigator Award of the Catalan Government, and an NSERC Canada grant.

1 Introduction

Consider the problem of modeling N independent and identically distributed observations on n discrete random variables consisting respectively of $K_1, \ldots, K_n$ categories. This type of data arises, for instance, in surveys, educational tests, or social science questionnaires when the number of choices is not constant over items. The observed data can be gathered in a n-dimensional contingency table with $C = \prod_i^n K_i$ cells. For a parametric model, $\boldsymbol{\pi}(\boldsymbol{\theta})$ is the C-dimensional vector of cell probabilities $\boldsymbol{\pi}$, where the q-dimensional parameter vector $\boldsymbol{\theta}$ is to be estimated from the data. For assessing the fit of the model, consider a composite null hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$ versus $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$. Researchers confronted with testing such a composite hypothesis face two problems. First, how to assess the overall goodness-of-fit of the hypothesized model, and second, how to determine the source of the misfit in poorly fitting models.

The two most commonly used goodness-of-fit statistics for testing the overall goodness of fit of a parametric model in multivariate categorical data analysis are Pearson's $X^2 = 2N \sum_{c=1}^{C} (p_c - \pi_c)^2/\pi_c$, and the likelihood ratio statistic $G^2 = 2N \sum_{c=1}^{C} p_c \ln(p_c/\pi_c)$. When the model holds, the two statistics are asymptotically equivalent. Under H_0, they are asymptotically distributed as chi-square with $C - q - 1$ degrees of freedom. However, it is well known that in sparse tables the empirical Type I error rates of the X^2 and G^2 test statistics do not match their expected rates under their asymptotic distribution. Of the two statistics, X^2 is less adversely affected by the sparseness of the contingency table that G^2 (Koehler & Larntz, 1980). One reason for the poor empirical performance of X^2 is that the empirical variance of X^2 and its variance under its reference asymptotic distribution differ by a term that depends on the inverse of the cell probabilities (Cochran, 1952). When the cell probabilities become small the discrepancy between the empirical and asymptotic variances of X^2 can be large and the type I error for X^2 will be larger than the α level based on its asymptotic critical value. Thus, the accuracy of the type I errors will depend on the model being fitted to the table (as it determines the cell probabilities), but also on the size of the contingency table. This is because when the size of the contingency table is large, the cell probabilities must be small (Bartholomew & Tzamourani, 1999). However, for C and $\boldsymbol{\pi}(\boldsymbol{\theta})$ fixed the accuracy of the the asymptotic p-values for X^2 depends also on sample size, N. As N becomes smaller some of the cell proportions increasingly become more poorly estimated (their estimates will be zero) and the empirical Type I errors of X^2 will become inaccurate. The degree of sparseness N/C summarizes the relationship between sample size and model size. Thus, the accuracy of the asymptotic p-values for X^2 depend on the model and the degree of sparseness of the contingency table.

Three alternative strategies have been proposed for obtaining Type I errors when the accuracy of the asymptotic p-values of X^2 is suspect: (a) pooling cells, (b) resampling methods, and (c) limited information methods. Our new statistical procedures are in category (c); we point out the advantages of (c) over (a) and (b) below.

Regarding (a), pooling cells before the model is fitted is a useful approach as it reduces the size of the contingency table, and thus the degree of sparseness. However, there is a limit in the amount of pooling that can be performed without distorting the purpose of the analysis. Also, pooling cells ad-hoc after the model has been fitted may result in a test statistic with an unknown asymptotic null sampling distribution. Regarding (b), generating the empirical sampling distribution of the goodness-of-fit statistic using a resampling method such as the parametric bootstrap method (e.g., Collins et al, 1993; Bartholomew & Tzamourani, 1999) may result in trustworthy p-values (but see Tollenaar & Moijaart, 2003). However, resampling methods may be very time consuming if the researcher is interested in comparing the fit of several models. On the other hand, limited information methods use only the information contained in the low order marginals of the contingency table to assess the model, and amounts to pooling cells a priori. The cells are pooled in a systematic way, so that the resulting statistics have a known asymptotic null distribution. These procedures are computationally much more efficient than resampling methods.

There have been several proposals in Psychometrics to use low order marginals in goodness-of-fit assessment of *binary* contingency tables, most notably Christoffersson (1975), Reiser (1996), Bartholomew and Leung (2002), Maydeu-Olivares (2001a, 2001b), and Maydeu-Olivares and Joe (2005). Limited information statistics appear as a viable framework to assess the overall goodness-of-fit of models for multidimensional contingency tables as they have more accurate empirical Type-I errors and can be asymptotically more powerful than full information statistics such as X^2 (Maydeu-Olivares & Joe, 2005; see also Reiser & VanderBergh, 1994). However, the only limited information test statistic proposed to date for multidimensional contingency tables of arbitrary dimensions (Maydeu-Olivares, in press, a) is only valid when parameters are estimated using the sequential estimator described in Jöreskog (1994) and implemented in Lisrel (Jöreskog & Sörbom, 2001).

A second challenge a researcher must confront when modeling multivariate categorical data is to identify the source of the misfit when the overall test suggests significant misfit. The inspection of cell residuals is often not very useful to this aim. It is difficult to find trends in inspecting these residuals, and even for moderate n the number of residuals to be inspected is too large. Perhaps most importantly, Bartholomew and Tzamourani (1999) point out that because the cell frequencies are integers and the expected frequencies in large tables must be very small, the resulting residuals will be either very small or very large. To overcome this challenge, numerous authors have advocated examining residuals from the two-way and three-way margins to assess the goodness-of-fit in binary contingency tables. Some key references in this literature are Reiser (1996), Reiser and Lin (1999), Reiser and VanderBergh (1994), Bartholomew and Tzamourani (1999), Bartholomew and Leung (2002), and Maydeu-Olivares and Joe (2005). However, when the observed variables are not binary, the number of marginal residuals grows very rapidly as the number of categories and variables increases, and it may be difficult to draw useful information by inspecting individual marginal residuals. To overcome this problem, it has been suggested (Drasgow, Levine, Tsien, Williams & Mead, 1995) to compute X^2 for single variables, pairs and triplets. However, X^2 applied to subtables is not asymptotically chi-square under the null hypothesis even for the maximum likelihood estimator.

In this paper, the main ideas and results of Maydeu-Olivares and Joe (2005) for the binary case ($K_i = 2$ for all i) are extended in two directions. First, we provide goodness-of-fit test statistics for multidimensional contingency tables of arbitrary dimensions. The statistics are quadratic forms in the residuals of marginal tables up to order r, for small r. These test statistics are asymptotically chi-square for any $\sqrt{N}$-consistent and asymptotically normal estimator. The extension is straightforward but the computational implementation is more cumbersome. Second, we provide statistics for assessing the goodness-of-fit in r-dimensional subtables. These statistics are also asymptotically chi-square under the same conditions than the statistics to assess the overall goodness-of-fit and they can be useful to identify the source of the misfit in poorly fitting models.

The remaining of the paper is organized as follows. In Section 2 we provide a convenient representation of multivariate categorical data which are a random sample from a multivariate multinomial (MVM) distribution, and we also provide the asymptotic distribution of multivariate marginal residuals for different estimators. In Section 3, we consider extensions of the family of limited information statistics M_r proposed by Maydeu-Olivares and Joe (2005). These statistics can be used with nominal categorical variables as they are invariant to arbitrary relabeling of the categories. Section 3 also includes a small simulation study to illustrate the small sample distributions of M_r (for small r) and X^2. In Section 4, we consider the use of marginal residuals and M_r statistics on r-dimensional subtables to identify the source of the misfit. Section 5 contains two examples to illustrate our results. Finally, Section 6 has conclusions and a discussion of further research.

For completeness, we also discuss in an Appendix goodness-of-fit testing of simple null hypotheses under MVM assumptions as a straightforward extension of the results of Maydeu-Olivares and Joe (2005) for multivariate Bernoulli assumptions. Computational details for estimation, evaluation of M_r and simulations are also given in the Appendix.

2 Multivariate multinomial distributions and asymptotic distribution of marginal residuals

In this section, we define the notation used in the remainder of this paper and we give two representations of the MVM distribution. One of them uses the cell probabilities, while the other uses a set of multivariate marginal probabilities. There is a one-to-one linear map between the two representations. We also provide the asymptotic distribution of cell residuals and of marginal residuals for MVM models where the parameters have been estimated using (a) ML or another best asymptotically normal estimator, and (b) a $\sqrt{N}$-consistent and asymptotically normal estimator (including limited information estimators).

2.1 Representation of the MVM distribution

By a MVM distribution, we mean a multivariate distribution with univariate margins that are multinomial. If the ith ($1 \leq i \leq n$) variable consists of $K_i \geq 2$ categories labeled as $0, 1, \ldots, K_i - 1$, with respective probabilities $p_{i0}, \ldots, p_{i,K_i-1}$, then one observation of the ith variable Y_i has a Multinomial$(1; p_{i0}, \ldots, p_{i,K_i-1})$ distribution. Using indicator functions, we give a representation of the MVM distribution. In the case where each $K_i = 2$, the representation is the same as that of Teugels (1990).

With the notation $Y_i = j$ meaning that Y_i has category j, we define the following indicator variables for $Y_1, \ldots, Y_n$:

$$I_{ij} = I(Y_i = j), \quad j = 1, \ldots, K_i - 1, \quad i = 1, \ldots, n \tag{2.1}$$

The univariate moments are $\mathrm{E}\,(I_{ij})$, $j = 1, \ldots, K_i - 1$, $i = 1, \ldots, n$; the bivariate moments are $\mathrm{E}\,[I_{i_1 j_1} I_{i_2 j_2}] = \Pr(Y_{i_1} = j_1, Y_{i_2} = j_2)$, $j_1 = 1, \ldots, K_{i_1} - 1$, $j_2 = 1, \ldots, K_{i_2} - 1$, $1 \leq i_1 < i_2 \leq n$. The trivariate up to n-dimensional moments can be defined in a similar way. Note that these moments consist of all joint and marginal probabilities of $Y_1, \ldots, Y_n$ that do not involve category 0 for any variables.

The distribution is characterized by all of the moments involving the I_{ij} up to the nth moments, in that all joint probabilities, including those involving the 0 categories, can be deduced from these moments. This follows by letting $I_{i0} = 1 - I_{i1} - \cdots - I_{iK_i}$; then

$$\Pr(Y_1 = j_1, \ldots, Y_n = j_n) = \mathrm{E}\,[I_{1j_1} \cdots I_{nj_p}],$$

and after expanding out any term with $j_i = 0$, this is a linear combination of the moments not involving any category of 0.

Consider the set $\mathcal{A}_r$ of expectations or moments that come from products of 1 to r indicators in (2.1). Then all probabilities up to the rth dimensional margins can be obtained from the set $\mathcal{A}_r$. There are no redundant moments in $\mathcal{A}_r$ in that no moment can be obtained as a linear combination of other moments in $\mathcal{A}_r$. The cardinality of $\mathcal{A}_r$ is equal to

$$s(r) \stackrel{\text{def}}{=} \sum_{j=1}^{r} \sum_{1\le i_1<\cdots<i_j\le n} \prod_{\ell=1}^{j} (K_{i_\ell} - 1) \tag{2.2}$$

which is smaller than the number $\sum_{1\le i_1<\cdots<i_r\le n} \prod_{\ell=1}^{r} K_{i_\ell}$ of rth order marginal probabilities. For example, if $K_i = K$ for $i = 1, \ldots, n$, then the number of rth order marginal probabilities is $\binom{n}{r}K^r$ which is larger than $s(r) = \sum_{j=1}^{r} \binom{n}{j}(K-1)^j$, the cardinality of $\mathcal{A}_r$. Note that $s(n) = C - 1$, with $C = \prod_{i=1}^{n} K_i$.

In the next section, we will be constructing quadratic form statistics based on residuals corresponding to the moments in $\mathcal{A}_r$. Because of the relationships mentioned above, the quadratic form statistics can also be expressed in terms of the residuals associated with all rth order marginal probabilities. It is an advantage computationally to work with the set $\mathcal{A}_r$ so that we can deal with smaller matrices in the quadratic form statistics. Note that even the cardinality of $\mathcal{A}_r$ increases rapidly as K_i and n increase. Also, for any goodness-of-fit statistic defined based on the moments up to order r, it is necessary to check/prove that the statistic is invariant to the labeling of the categories, since it is generally arbitrary which category is labeled as category 0.

Further insight into the relationship between the multivariate moment and the cell representation is obtained by using a notation analogous to that employed in Maydeu-Olivares and Joe (2005). In what follows we assume for notational ease that $K_i = K$ for all i. Consider a n-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ of K-category random variables, with $\pi_i(j) = \Pr(Y_i = j)$, $i = 1, \ldots, n$, and joint distribution:

$$\pi_{\mathbf{y}} = \Pr(Y_i = y_i, i = 1, \ldots, n), \quad \mathbf{y} = (y_1, \ldots, y_n), \quad y_i \in \{0, \ldots, K-1\}.$$

When we consider a parametric model with parameter vector $\boldsymbol{\theta}$, we write $\pi_{\mathbf{y}}(\boldsymbol{\theta})$ for an individual probability and $\boldsymbol{\pi}(\boldsymbol{\theta})$ for the vector of K^n joint probabilities. Also, we write $\dot{\boldsymbol{\pi}}_1$ for the $n(K-1)$ vector of univariate marginal probabilities. Similarly, we write $\dot{\boldsymbol{\pi}}_2$ for the $\binom{n}{2}(K-1)^2$ vector of bivariate marginal probabilities, and so forth up to $\dot{\boldsymbol{\pi}}_n$, the $\binom{n}{n}(K-1)^n$ vector of nth way marginal probabilities. Finally, let $\dot{\boldsymbol{\pi}}' = (\dot{\boldsymbol{\pi}}_1', \dot{\boldsymbol{\pi}}_2', \ldots, \dot{\boldsymbol{\pi}}_n')'$. Then, we can write $\dot{\boldsymbol{\pi}} = \mathbf{T}\boldsymbol{\pi}$, where $\mathbf{T}$ is a $(K^n - 1) \times K^n$ matrix of 1s and 0s, of full row rank (if K_i is not constant, then $\mathbf{T}$ is $(C-1) \times C$).

$\mathbf{T}$ can be partitioned according to the partitioning of $\dot{\boldsymbol{\pi}}$,

$$\begin{pmatrix} \dot{\boldsymbol{\pi}}_1 \\ \dot{\boldsymbol{\pi}}_2 \\ \vdots \\ \dot{\boldsymbol{\pi}}_n \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{n1} \\ \mathbf{T}_{n2} \\ \vdots \\ \mathbf{T}_{nn} \end{pmatrix} \boldsymbol{\pi}.$$

The vector of multivariate moments up to order r ($r \leq n$), denoted by $\boldsymbol{\pi}_r = (\dot{\boldsymbol{\pi}}_1', \ldots, \dot{\boldsymbol{\pi}}_r')'$, can be written as

$$\boldsymbol{\pi}_r = \mathbf{T}_r \boldsymbol{\pi},$$

where $\mathbf{T}_r = (\mathbf{T}_{n1}', \ldots, \mathbf{T}_{nr}')'$. Note that by definition $\boldsymbol{\pi}_n = \dot{\boldsymbol{\pi}}$. That is, $\mathbf{T}_r$ is the mapping of the C-dimensional vector of cell probabilities to the moments in $\mathcal{A}_r$.

Because the pattern $\pi_{0\cdots 0}$ is not used to obtain the marginal moments $\dot{\boldsymbol{\pi}}$, the first column of $\mathbf{T}$ is a column of zeros, so we can partition $\mathbf{T} = (\, \mathbf{0} \quad \dot{\mathbf{T}} \,)$, and write $\dot{\boldsymbol{\pi}} = \dot{\mathbf{T}}\breve{\boldsymbol{\pi}}$ with $\boldsymbol{\pi} = \begin{pmatrix} \pi_{0\cdots 0} \\ \breve{\boldsymbol{\pi}} \end{pmatrix}$. Since $\pi_{0\cdots 0} = 1 - \mathbf{1}'\breve{\boldsymbol{\pi}}$, the inverse relationship between $\dot{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ is

$$\boldsymbol{\pi} = \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} -\mathbf{1}'\dot{\mathbf{T}}^{-1} \\ \dot{\mathbf{T}}^{-1} \end{pmatrix} \dot{\boldsymbol{\pi}}.$$

2.2 Asymptotic distribution of marginal residuals for a fixed a priori parameter vector

For a random sample of size N from a MVM model $\boldsymbol{\pi}(\boldsymbol{\theta})$, let $\mathbf{p}$ and $\dot{\mathbf{p}}$ denote the C-dimensional vector of cell proportions, and the $(C-1)$-dimensional vector of sample joint moments, respectively. Then

$$\sqrt{N}\left(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}(\boldsymbol{\theta})\right) = \mathbf{T}\sqrt{N}\left(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})\right). \tag{2.3}$$

Since (Agresti, 2002)

$$\sqrt{N}\left(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})\right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma}(\boldsymbol{\theta})), \quad \text{where} \quad \boldsymbol{\Gamma}(\boldsymbol{\theta}) = \mathbf{D}(\boldsymbol{\theta}) - \boldsymbol{\pi}(\boldsymbol{\theta})\boldsymbol{\pi}'(\boldsymbol{\theta}), \quad \mathbf{D}(\boldsymbol{\theta}) = \text{ diag}\left(\boldsymbol{\pi}(\boldsymbol{\theta})\right),$$

it follows from (2.3) that

$$\sqrt{N}\left(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}(\boldsymbol{\theta})\right) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{\Xi}(\boldsymbol{\theta})\big), \quad \boldsymbol{\Xi}(\boldsymbol{\theta}) = \mathbf{T}\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{T}'.$$

Also, let $\mathbf{p}_r$ be the vector of sample moments up to order r; it has dimension $s(r)$ as given in (2.2). Then,

$$\sqrt{N}\left(\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta})\right) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{\Xi}_r(\boldsymbol{\theta})\big), \quad \boldsymbol{\Xi}_r(\boldsymbol{\theta}) = \mathbf{T}_r\boldsymbol{\Gamma}(\boldsymbol{\theta})\mathbf{T}_r'. \tag{2.4}$$

Here we have provided the asymptotic distribution of residual marginals for MVM parametric models $\boldsymbol{\pi}(\boldsymbol{\theta})$ for a fixed a priori vector $\boldsymbol{\theta}$ of dimension q. In practice, in most applications for multivariate categorical data, one is interested in comparing one or more MVM models where $\boldsymbol{\theta}$ is estimated from the data. We next provide the asymptotic distribution of residual marginals when parameters are estimated, via maximum likelihood (ML) or another estimation method.

2.3 Asymptotic distribution of marginal residuals for best asymptotic normal (BAN) estimators

Let $\boldsymbol{\pi}(\boldsymbol{\theta})$ be a parametric MVM model with parameters $\boldsymbol{\theta}$ to be estimated from the data. In this subsection we consider the case where the q-dimensional vector $\boldsymbol{\theta}$ is estimated using a consistent and asymptotically normal minimum variance (or BAN) estimator such as the maximum likelihood estimator (MLE) or the minimum chi-square estimator. We assume that the usual regularity conditions on the model are satisfied so as to fulfill the consistency and asymptotic normality of the $\boldsymbol{\theta}$ estimates. The matrices below mostly depend on $\boldsymbol{\theta}$ but we omit this for notational ease.

Suppose we have a sample of size N from a MVM distribution. Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimator (ML) or another consistent minimum variance estimator. Then (Bishop, Fienberg & Holland, 1975),

$$\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{B}\sqrt{N}\,(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1), \qquad \mathbf{B} = \boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1}, \tag{2.5}$$

and $\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1})$, where $\boldsymbol{\mathcal{I}} = \boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta}$ is the Fisher information matrix. Letting $\hat{\mathbf{e}} = \mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}) = \mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta}) - \boldsymbol{\Delta}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(N^{-1/2})$ denote the vector of cell residuals, we have $\sqrt{N}\,\hat{\mathbf{e}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$ with asymptotic covariance matrix

$$\boldsymbol{\Sigma} = (\mathbf{I} - \boldsymbol{\Delta}\mathbf{B})\boldsymbol{\Gamma}(\mathbf{I} - \boldsymbol{\Delta}\mathbf{B})' = \boldsymbol{\Gamma} - \boldsymbol{\Delta}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}'. \tag{2.6}$$

For the marginal residuals, $\hat{\mathbf{e}}_r = \mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}) = \mathbf{T}_r\hat{\mathbf{e}}$, $\sqrt{N}\,\hat{\mathbf{e}}_r \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_r)$, where

$$\boldsymbol{\Sigma}_r = \mathbf{T}_r\boldsymbol{\Sigma}\mathbf{T}_r' = \boldsymbol{\Xi}_r - \boldsymbol{\Delta}_r\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}_r'. \quad \text{with } \boldsymbol{\Xi}_r = \mathbf{T}_r\boldsymbol{\Gamma}\mathbf{T}_r', \quad \boldsymbol{\Delta}_r = \frac{\partial\boldsymbol{\pi}_r(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}. \tag{2.7}$$

Note that the dimension of $\boldsymbol{\Delta}_r$ is $s(r) \times q$.

2.4 Asymptotic distribution of marginal residuals for other consistent and asymptotically normal estimators

When the n-dimensional probabilities may be too difficult to compute (for instance when they involve high-dimensional numerical integration) other simpler estimators may be computationally more convenient than BAN estimators. In Psychometrics, limited information estimators for discretized multivariate normal structural models that estimate the parameters using only univariate and bivariate information are rather popular and they have been implemented in commercially available computer programs such as Lisrel (Jöreskog & Sörbom, 2001), Mplus (Muthén & Muthén, 2001), and EQS (Bentler, 1995).

In this subsection we consider the asymptotic distribution of marginal residuals when the model parameters are estimated using some alternative $\sqrt{N}$-consistent and asymptotically normal estimator $\tilde{\boldsymbol{\theta}}$. The results of the previous subsection become a special case. As before, we assume that the usual regularity conditions on the model are satisfied so as to fulfill the consistency and asymptotic normality of the estimates. In particular, we assume that $\tilde{\boldsymbol{\theta}}$ satisfies

$$\sqrt{N}\,(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{H}\sqrt{N}\,(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1), \tag{2.8}$$

for some $q \times C$ matrix $\mathbf{H}$. Many estimators, among them the limited information estimators $\tilde{\boldsymbol{\theta}}$ considered by Christoffersson (1975), Jöreskog (1994; see also Maydeu-Olivares, in press, a), Jöreskog and Moustaki (2001), Lee, Poon and Bentler (1995), Maydeu-Olivares (2001b), and Muthén (1978, 1984, 1993) are special cases of this framework.

The asymptotic distribution of the vector of cell residuals $\tilde{\mathbf{e}} = \mathbf{p} - \boldsymbol{\pi}(\tilde{\boldsymbol{\theta}})$ for (2.8) can be obtained as follows. Note that $\boldsymbol{\pi}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\pi}(\boldsymbol{\theta}) = \boldsymbol{\Delta}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(N^{-1/2}) = \boldsymbol{\Delta}\mathbf{H}(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(N^{-1/2})$. Since $\mathbf{p} - \boldsymbol{\pi}(\tilde{\boldsymbol{\theta}}) = [\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})] - [\boldsymbol{\pi}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\pi}(\boldsymbol{\theta})]$, then $\sqrt{N}\,\tilde{\mathbf{e}} = (\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1)$. Thus, the asymptotic distribution of the cell residuals is $\sqrt{N}\,\tilde{\mathbf{e}} \xrightarrow{d} N(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}})$ with asymptotic covariance matrix

$$\widetilde{\boldsymbol{\Sigma}} = (\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})\boldsymbol{\Gamma}(\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})'. \tag{2.9}$$

Next we consider residuals up to order r only. Let the vector of residuals of the moments be $\tilde{\mathbf{e}}_r = \mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})$. Since $\tilde{\mathbf{e}}_r = \mathbf{T}_r\tilde{\mathbf{e}}$, the asymptotic distribution of these marginal residuals is (using (2.7)) $\sqrt{N}\,\tilde{\mathbf{e}}_r \xrightarrow{d} N(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_r)$, with

$$\widetilde{\boldsymbol{\Sigma}}_r = \mathbf{T}_r\widetilde{\boldsymbol{\Sigma}}\mathbf{T}_r' = (\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{H})\boldsymbol{\Gamma}(\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{H})' = \boldsymbol{\Xi}_r - \boldsymbol{\Delta}_r\mathbf{H}\boldsymbol{\Gamma}\mathbf{T}_r' - \mathbf{T}_r\boldsymbol{\Gamma}\mathbf{H}'\boldsymbol{\Delta}_r' + \boldsymbol{\Delta}_r[\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}']\boldsymbol{\Delta}_r', \tag{2.10}$$

where $\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}'$ is the asymptotic covariance matrix of $\sqrt{N}\,\tilde{\boldsymbol{\theta}}$.

3 Overall goodness-of-fit testing using marginal residuals

In this section we consider testing a composite null hypothesis using quadratic forms in the marginal residuals. That is, we consider the hypothesis H_0: $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$ versus H_1: $\boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$, when parameters are estimated using a method that yields $\sqrt{N}$-consistent and asymptotically normal estimates. Let r_0 be the smallest integer r such that the model is (locally) identified from the marginal residuals up to order r. Then, for $r \geq r_0$, the matrix $\boldsymbol{\Delta}_r$ is of full column rank q. Also, we assume that $s(r) > q$ so as to exclude the case $s(r) = q$.

3.1 The family of test statistics M_r

In the special case $K_i = 2$, Maydeu-Olivares and Joe (2005) introduced the family of statistics M_r for testing composite null hypotheses for multivariate binary models. Their results readily extend to MVM models for contingency tables of arbitrary dimensions. The notation is basically the same but the dimension of the matrices is larger, and numerical computations are harder.

Consider a $s(r) \times (s(r)-q)$ orthogonal complement to $\mathbf{\Delta}_r$ (given in (2.7)), say $\mathbf{\Delta}_r^{(c)}$, such that ${\mathbf{\Delta}_r^{(c)}}'\mathbf{\Delta}_r = \mathbf{0}$. Then, from (2.10) and (2.7), $\sqrt{N}\,[\mathbf{\Delta}_r^{(c)}]'\tilde{\mathbf{e}}_r = {\mathbf{\Delta}_r^{(c)}}'\sqrt{N}\left(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})\right)$ has asymptotic covariance matrix

$$
{\mathbf{\Delta}_r^{(c)}}'\widetilde{\mathbf{\Sigma}}_r\mathbf{\Delta}_r^{(c)} = {\mathbf{\Delta}_r^{(c)}}'\mathbf{\Xi}_r\mathbf{\Delta}_r^{(c)}. \tag{3.1}
$$

Next, let

$$
\mathbf{C}_r = \mathbf{C}_r(\boldsymbol{\theta}) = \mathbf{\Delta}_r^{(c)}\big([\mathbf{\Delta}_r^{(c)}]'\mathbf{\Xi}_r\mathbf{\Delta}_r^{(c)}\big)^{-1}[\mathbf{\Delta}_r^{(c)}]' = \mathbf{\Xi}_r^{-1} - \mathbf{\Xi}_r^{-1}\mathbf{\Delta}_r\big(\mathbf{\Delta}_r'\mathbf{\Xi}_r^{-1}\mathbf{\Delta}_r\big)^{-1}\mathbf{\Delta}_r'\mathbf{\Xi}_r^{-1}. \tag{3.2}
$$

Note that $\mathbf{C}_r$ is invariant to the choice of orthogonal complement (if $\mathbf{\Delta}_r^{(c)}$ is a full rank orthogonal complement, then so $\mathbf{\Delta}_r^{(c)}\mathbf{A}$ for a nonsingular matrix $\mathbf{A}$), and the last equality in (3.2) follows from a result in Rao (1973: p. 77). Then,

$$
\sqrt{N}\,[\mathbf{\Delta}_r^{(c)}]'\tilde{\mathbf{e}}_r \xrightarrow{d} N\big(\mathbf{0}, {\mathbf{\Delta}_r^{(c)}}'\mathbf{\Xi}_r\mathbf{\Delta}_r^{(c)}\big). \tag{3.3}
$$

The limited information statistic M_r of order r is given by

$$
M_r = M_r(\tilde{\boldsymbol{\theta}}) = N\tilde{\mathbf{e}}_r'\widetilde{\mathbf{\Delta}}_r^{(c)}\Big([\widetilde{\mathbf{\Delta}}_r^{(c)}]'\widetilde{\mathbf{\Xi}}_r\widetilde{\mathbf{\Delta}}_r^{(c)}\Big)^{-1}[\widetilde{\mathbf{\Delta}}_r^{(c)}]'\tilde{\mathbf{e}}_r = N\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})\big)'\widetilde{\mathbf{C}}_r\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})\big). \tag{3.4}
$$

In (3.4), $\widetilde{\mathbf{C}}_r$ denotes $\mathbf{C}_r(\tilde{\boldsymbol{\theta}})$ and other matrices are also evaluated at $\tilde{\boldsymbol{\theta}}$. This holds for any $\sqrt{N}$-consistent and asymptotically normal estimator $\tilde{\boldsymbol{\theta}}$, including the BAN estimators, denoted as $\hat{\boldsymbol{\theta}}$ in Section 2.3. It is straightforward to verify that $\mathbf{C}_r = \mathbf{C}_r\widetilde{\mathbf{\Sigma}}_r\mathbf{C}_r$, that is, $\widetilde{\mathbf{\Sigma}}_r$ is a generalized inverse of $\mathbf{C}_r$. By using Slutsky's theorem,

$$
M_r \xrightarrow{d} \chi^2_{s(r)-q},
$$

where the degrees of freedom are obtained from a result in Rao (1973: p. 30) using the fact that $\mathbf{\Delta}_r^{(c)}$ is of full column rank $s(r)-q$ and hence $\mathbf{C}_r$ is also of rank $s(r)-q$.

Note that (3.4) does not use the generalized inverse of $\mathbf{\Sigma}_r$ because this may be numerically unstable with a small singular value. Also computation of $\mathbf{C}_r$, which depends on $\mathbf{\Delta}_r$ and $\mathbf{\Xi}_r$, is much easier than that of $\mathbf{\Sigma}_r$ (which depends also on $\boldsymbol{\mathcal{I}}$) or $\widetilde{\mathbf{\Sigma}}_r$ (which depends also on $\mathbf{H\Gamma T}_r'$ and $\mathbf{H\Gamma H}'$).

$\{M_r\}$ is a family of test statistics based on residuals up to r-variate margins whose members are $\{M_1, \cdots, M_n\}$. M_1 is defined only if $s(1) > q$, that is, for models that do not have many parameters; for example, it is not defined for the item response model that we use later in this paper. M_1 is a quadratic form in univariate residuals, whereas M_2 is a quadratic form in univariate and bivariate residuals, and so forth, up to M_n which is a full information test statistic. Following the technique used in the Appendix of Maydeu-Olivares and Joe (2005), M_n can be written as a quadratic form in the cell residuals as

$$
M_n = N\big(\mathbf{p} - \boldsymbol{\pi}(\tilde{\boldsymbol{\theta}})\big)'\widetilde{\mathbf{U}}\big(\mathbf{p} - \boldsymbol{\pi}(\tilde{\boldsymbol{\theta}})\big) \tag{3.5}
$$

with $\widetilde{\mathbf{U}} = \mathbf{U}(\tilde{\boldsymbol{\theta}})$, where $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{D}^{-1} - \mathbf{D}^{-1}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1}$. Also, they show that this statistic can be alternatively be written as

$$M_n = N\big(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}(\tilde{\boldsymbol{\theta}})\big)'\widetilde{\mathbf{C}}_n\big(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}(\tilde{\boldsymbol{\theta}})\big) \tag{3.6}$$

with $\widetilde{\mathbf{C}}_n = \mathbf{C}_n(\tilde{\boldsymbol{\theta}})$

Thus, we have shown that M_r is asymptotically $\chi^2_{s(n)-q}$ if $\tilde{\boldsymbol{\theta}}$ is any $\sqrt{N}$-consistent and asymptotically normal estimator of $\boldsymbol{\theta}$. Since $C = s(n)+1$ is the number of possible cells in the contingency table, we have shown that the full information test statistic $M_n = M_n(\tilde{\boldsymbol{\theta}})$ is asymptotically χ^2_{C-1-q} for this large class of consistent estimators.

Also with a proof very similar to that in the Appendix of Maydeu-Olivares and Joe (2005), M_r is invariant to the labeling of the categories, assuming that with permuted categories $\boldsymbol{\theta} \to \boldsymbol{\theta}_\Lambda$, a permuted vector, and $\tilde{\boldsymbol{\theta}}$ is an equivariant estimator.

Previously, there had not been any goodness-of-fit statistic that is asymptotically chi-square for any $\sqrt{N}$-consistent estimator of $\boldsymbol{\theta}$. In particular, the full information test statistic M_n can be used to assess the overall goodness-of-fit of categorical data models estimated using the limited information estimators implemented in Lisrel, Mplus, or EQS. Note that with $X^2(\tilde{\boldsymbol{\theta}})$ representing the X^2 statistic based on $\tilde{\boldsymbol{\theta}}$, the results in the Appendix of Maydeu-Olivares and Joe (2005) imply that $M_n(\tilde{\boldsymbol{\theta}}) \le X^2(\tilde{\boldsymbol{\theta}})$. That is, for a consistent estimator that is not the MLE, the asymptotic distribution of $X^2(\tilde{\boldsymbol{\theta}})$ is stochastically larger than χ^2_{C-1-q}. Also, $M_n = X^2$ when $\tilde{\boldsymbol{\theta}}$ is the MLE. But for other minimum variance asymptotically normal estimators, $M_n \le X^2$ and M_n and X^2 are equivalent only asymptotically.

Maydeu-Olivares and Joe (2005) pointed out that the asymptotic variance of M_r is influenced by the smallest marginal probability of dimension $\min\{2r, n\}$. Therefore, the asymptotic null distribution of M_r can be acceptable if the rth order margins are not sparse, and larger sample sizes are needed as r increases for the null asymptotics to be valid. This was illustrated using a simulation study where a two-parameter logistic model (Lord & Novick, 1968) was estimated by MLE. For the less sparse situations, the small sample behavior of $M_n = X^2$ was close to its asymptotic reference distribution. But as sparseness increased the empirical Type I errors of X^2 first — and with increased sparseness M_3 as well — departed from its expected rates. Only the empirical Type I errors of M_2 remained accurate throughout the different sparseness conditions considered in their study. In the next subsection we extend their simulation results by (a) considering an item response (IRT) model for variables where $K_i > 2$, (b) considering much larger contingency tables, and by (c) investigating the behavior of the test statistics for a limited information estimator.

3.2 Small sample performance of M_r

For an illustration of the small sample performance of M_r consider a unidimensional *item response model* (e.g., van der Linden & Hambleton, 1997)

$$\Pr\left[\cap_{i=1}^{n}\{Y_i = y_i\}\right] = \int_{-\infty}^{\infty} \prod_{i=1}^{n} \Pr(Y_i = y_i \mid \eta)\, f(\eta) d\eta, \quad y_i \in \{0, \ldots, K-1\} \tag{3.7}$$

where $f(\eta)$ denotes the density of the a continuous unobserved variable (i.e, a latent trait). Note that under this family of models, the probabilities conditional on the latent trait are assumed to be independent. For ordered categorical variables, Samejima (1969) proposed letting $f(\eta)$ be a standard normal density function and

$$\Pr(Y_i = j \mid \eta) = \begin{cases} 1 - G(\alpha_{i,1} + \beta_i \eta) & \text{if } j = 0 \\ G(\alpha_{i,j} + \beta_i \eta) - G(\alpha_{i,j+1} + \beta_i \eta) & \text{if } 0 < j < K-1 \\ G(\alpha_{i,K-1} + \beta_i \eta) & \text{if } j = K-1 \end{cases} \tag{3.8}$$

where $G(z)$ equals either the standard logistic distribution function

$$\Psi(z) = [1 + \exp\{-z\}]^{-1} \tag{3.9}$$

or the standard normal distribution function

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}\, dx \tag{3.10}$$

Thus, in this model, for each item there is one slope parameter β_i and $K-1$ intercept parameters $\alpha_{i,j}$; $\alpha_{i,j}$ is decreasing in j for each i. Samejima (1969) referred to the model specified by equations (3.7–3.10) as the (logistic or normal) graded model. Bartholomew and Knott (1999) refer to these models as the logit-normit and normit-normit models, respectively. The family of models (3.7) are random effects members of the larger *generalized linear mixed model* (GLMM) family (see Agresti, 2002: Chapter 12).

Note that for model (3.7–3.9), the number of parameters is $q = nK$ so that M_r in (3.1) is defined only for $r \geq 2$ since $s(1) = n(K-1) < q$. M_3 would be useful to compute only if $s(3)/N$ is large enough. For most item response theory applications M_2 is the statistic of choice in the M_r family.

To illustrate the small sample behavior of M_2 for ML estimation, we generated data according to Samejima's logistic model for many different parameter vectors. We summarize some representative results in Table 1, which has three cases of (K, n): $(3, 5)$ with $C = 243$ cells; $(5, 5)$ with $C = 3125$ cells and $(5, 10)$ with $C = 9765625 \approx 10^7$ cells. The sample sizes are $N = 300$, 1000, and 3000. The procedure used to generate the data is explained in the Appendix subsection on computing notes.

For $K = 3$, $\boldsymbol{\alpha} = (-1, 1)$ for all items, and for $K = 5$, $\boldsymbol{\alpha} = (-1, -0.5, 0.5, 1)$ for all items. For $n = 5$, $\boldsymbol{\beta} = (1, 1.5, 2, 1.5, 1)$, whereas for $n = 10$, $\boldsymbol{\beta} = (1, 1.5, 2, 1.5, 1, 1, 1.5, 2, 1.5, 1)$.

A small model with $K = 3$ and $n = 5$ was chosen to show that the empirical rejection rates of M_2 are similar to those of X^2 when the latter are accurate. The other cases with larger C were chosen to show that the empirical rejection rates of M_2 remain accurate, unlike those of X^2, even for extremely sparse tables. As can be seen in Table 1, the empirical Type I errors for M_2 remain close to its nominal levels even at the highest degree of sparseness considered, whereas those of X^2 are only accurate in the small model with $K = 3$, $n = 5$.

INSERT TABLE 1 ABOUT HERE

Also, we use a bivariate composite likelihood (BCL) estimator (Zhao & Joe, 2005) under the same conditions as above to illustrate the behavior of M_2 for estimators that are not BAN. The results are shown in Table 2. The BCL estimator is the maximum of the sum of the $\binom{n}{2}$ bivariate marginal log-likelihoods, rather than the maximum of the joint n-dimensional log-likelihood. In one special setting, Jöreskog and Moustaki (2001) refer to this as the underlying bivariate normal (UBN) approach. If the trivariate margins are not sparse, one could consider the trivariate composite likelihood estimator. The asymptotic analysis of this method can be done using the theory of estimating equations (Godambe, 1991) and the asymptotic covariance matrix of the BCL estimator is an inverse Godambe information matrix, which can be compared with the inverse Fisher information matrix. We were able to compute both of these for different parameter vectors for (3.8), and look at ratios of the diagonals of these two matrices. For all cases that we computed for $n \leq 10$ and $2 \leq K \leq 5$, the asymptotic relative efficiency of any component of the BCL estimator is over 0.98; the average efficiency tends to slowly decrease as n increases.

As can be seen in Table 2, the finite sample null distribution of the M_2 statistic with $\tilde{\boldsymbol{\theta}}$ behaves very similarly to M_2 with the MLE. Although we have only studied the null hypothesis small sample performance for one (commonly used) model for item response categorical data, we expect the behavior to be similar for other models, for the maximum likelihood estimator and other $\sqrt{N}$-consistent estimators.

INSERT TABLE 2 ABOUT HERE

3.3 Power comparison of X^2 and M_r when data are not sparse

For the binary case, Maydeu-Olivares and Joe (2005) have an asymptotic power comparison under a sequence of local alternatives for model (3.7–3.9). They report that M_2 and M_3 typically had more power asymptotically than X^2 for the null hypothesis of a common slope parameter.

For $K_i = K > 2$, we have done some simulations that show a similar behavior for M_2 for finite sample sizes. A finite sample power comparison of X^2 and M_r is meaningful only in the non-sparse cases where X^2 can be used. Consequently, Table 3 has some summaries for representative cases for small sample power comparison for $n = 5, K = 3$, using model (3.7–3.9) with constant $\beta_i = \beta$ as the null nested model.

INSERT TABLE 3 ABOUT HERE

4 Using marginal residuals to assess the source of misfit

When the M_r statistic suggests a model misfit, the vector of standardized marginal residuals can be inspected. This is $N[\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})]$, the differences of observed and model expected counts or moments for margins, divided by the square root of $\mathrm{diag}\,(\widetilde{\boldsymbol{\Sigma}}(\tilde{\boldsymbol{\theta}}))$. Note that this vector includes only those categories for which no category index is 0. The remaining residuals can be obtained based on zero sum constraints or by computing the residuals from inverse coded categories (the M_r statistic is invariant to the inverse coding).

In large models, particularly when the number of categories for some variables is large, there will be a large number of marginal residuals involved and it may be difficult to draw useful information. Furthermore, the standardized residuals may be difficult to compute in large models.

A more fruitful avenue to assess the source of misfit might be to examine the rth dimensional marginal tables. Note that this is like multiple testing after a jointly significance result. An analogy is Fisher's least significant difference following a significant F-ratio in ANOVA. In other words, we recommend assessing the source of misfit by computing $M_r^{(b)}(\tilde{\boldsymbol{\theta}})$ for each subset b of $\{1, \dots, n\}$ with cardinality r. For a submodel for r-dimensional margins, with $C_r(b) = \prod_{i \in b} K_i$ cells depending on $q_r(b)$ parameters, $M_r^{(b)}(\tilde{\boldsymbol{\theta}})$ has an asymptotic null chi-square distribution with $C_r(b) - q_r(b) - 1$ degrees of freedom, provided the submodel is identified, the estimator is consistent and asymptotically normal, and $C_r(b) - 1 > q_r(b)$. When $r = 2$, we write $M_2^{(ij)}$ for $1 \le i < j \le n$. Also if $K_i = K$ for all i, then $C_2(b) = K^2$.

To see this, consider M_r applied to the r-variate subset b. Let the vector of sample and model moments for this subset be denoted as $\dot{\mathbf{p}}_{rb}$, and $\dot{\boldsymbol{\pi}}_{rb}(\tilde{\boldsymbol{\theta}}_b)$, respectively, both of dimension $C_r(b)-1$. Typically $\boldsymbol{\theta}_b$ is a subset of the vector $\boldsymbol{\theta}$. Let $q_r(b)$ be the dimension of $\boldsymbol{\theta}_b$. Using (3.6), we can write M_r in this case as

$$M_r^{(b)}(\tilde{\boldsymbol{\theta}}) = M_r^{(b)}(\tilde{\boldsymbol{\theta}}_b) = N(\dot{\mathbf{p}}_{rb} - \dot{\boldsymbol{\pi}}_{rb}(\tilde{\boldsymbol{\theta}}_b))'\widetilde{\mathbf{C}}_{rb}(\dot{\mathbf{p}}_{rb} - \dot{\boldsymbol{\pi}}_{rb}(\tilde{\boldsymbol{\theta}}_b))$$

for some $\sqrt{N}$-consistent and asymptotically normal estimator $\tilde{\boldsymbol{\theta}}$. We assume that $\boldsymbol{\Delta}_{rb} = \partial\dot{\boldsymbol{\pi}}_{rb}(\boldsymbol{\theta}_b)/\partial\boldsymbol{\theta}_b'$ is of full rank $q_r(b)$, so that the submodel is (locally) identified. Also, we assume that $C_r(b)-1-q_r(b)>0$. The matrix of the above quadratic form is

$$\widetilde{\mathbf{C}}_{rb} = \mathbf{C}_{rb}(\tilde{\boldsymbol{\theta}}_b) = \boldsymbol{\Delta}_{rb}^{(c)}\left([\boldsymbol{\Delta}_{rb}^{(c)}]'\boldsymbol{\Xi}_b\boldsymbol{\Delta}_{rb}^{(c)}\right)^{-1}[\boldsymbol{\Delta}_{rb}^{(c)}]'$$

evaluated at $\tilde{\boldsymbol{\theta}}_b$, where $\boldsymbol{\Delta}_{rb}^{(c)}$ is an orthogonal complement to $\boldsymbol{\Delta}_{rb}$, and $\boldsymbol{\Xi}_{rb}$ is N times the asymptotic covariance matrix of $\dot{\mathbf{p}}_{rb} - \dot{\boldsymbol{\pi}}_{rb}(\boldsymbol{\theta}_b)$. Now, $(\dot{\mathbf{p}}_{rb} - \dot{\boldsymbol{\pi}}_{rb}(\tilde{\boldsymbol{\theta}}_b)) = \mathbf{T}_{rb}(\mathbf{p}-\boldsymbol{\pi}(\tilde{\boldsymbol{\theta}}))$ for some $(C_r(b)-1)\times C$ matrix $\mathbf{T}_{rb}$. Thus, using (2.9), the asymptotic covariance matrix of $\sqrt{N}\,(\dot{\mathbf{p}}_{rb} - \dot{\boldsymbol{\pi}}_{rb}(\tilde{\boldsymbol{\theta}}_b))$ is $\boldsymbol{\Sigma}_{rb} = \mathbf{T}_{rb}(\mathbf{I}-\boldsymbol{\Delta}\mathbf{H})\boldsymbol{\Gamma}(\mathbf{I}-\boldsymbol{\Delta}\mathbf{H})'\mathbf{T}_{rb}' = (\mathbf{T}_{rb}-\boldsymbol{\Delta}_{rb}\mathbf{H})\boldsymbol{\Gamma}(\mathbf{T}_{rb}-\boldsymbol{\Delta}_{rb}\mathbf{H})'$.

A necessary and sufficient condition for $M_r^{(b)}$ to be asymptotically distributed (under H_0) as a chi-square with ν degrees of freedom in this setup is (Schott, 1997: Theorem 9.10)

$$\boldsymbol{\Sigma}_{rb}\mathbf{C}_{rb}\boldsymbol{\Sigma}_{rb}\mathbf{C}_{rb}\boldsymbol{\Sigma}_{rb} = \boldsymbol{\Sigma}_{rb}\mathbf{C}_{rb}\boldsymbol{\Sigma}_{rb} \quad \text{for any } \boldsymbol{\theta}, \tag{3.9}$$

where $\nu = \operatorname{tr}(\mathbf{C}_{rb}\boldsymbol{\Sigma}_{rb})$. Since $\boldsymbol{\Xi}_{rb} = \mathbf{T}_{rb}\boldsymbol{\Gamma}\mathbf{T}_{rb}'$, it can be readily verified that $\mathbf{C}_{rb} = \mathbf{C}_{rb}\boldsymbol{\Sigma}_{rb}\mathbf{C}_{rb}$. That is, $\boldsymbol{\Sigma}_{rb}$ is a generalized inverse for $\mathbf{C}_{rb}$. So, (3.9) is satisfied. Also, the degrees of freedom are obtained using the fact that $\boldsymbol{\Delta}_{rb}^{(c)}$ is of full column rank $C_r(b)-1-q_r(b)$ and hence $\mathbf{C}_{rb}$ is also of rank $C_r(b)-1-q_r(b)$. Thus, the null distribution of $M_r^{(b)}(\tilde{\boldsymbol{\theta}}_b)$ is asymptotically chi-square with degrees of freedom $C_r(b)-1-q_r(b)$.

On the other hand, Pearson's X^2 is not asymptotically chi-square under H_0 when applied to subsets of variables even for BAN estimators. To see this, from the Appendix of Maydeu-Olivares and Joe (2005), X^2 applied to the r-variate subset b can be written as $X_b^2 = N(\dot{\mathbf{p}}_b - \dot{\boldsymbol{\pi}}_b(\tilde{\boldsymbol{\theta}}_b))'\boldsymbol{\Xi}_{rb}^{-1}(\dot{\mathbf{p}}_b - \dot{\boldsymbol{\pi}}_b(\tilde{\boldsymbol{\theta}}_b))$. Now, using (2.6), the asymptotic covariance matrix of $\sqrt{N}\,(\dot{\mathbf{p}}_b - \dot{\boldsymbol{\pi}}_b(\tilde{\boldsymbol{\theta}}_b))$ for BAN estimators such as the MLE is $\boldsymbol{\Sigma}_{rb} = \mathbf{T}_{rb}(\boldsymbol{\Gamma} - \boldsymbol{\Delta}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}')\mathbf{T}_{rb}' = \boldsymbol{\Xi}_{rb} - \boldsymbol{\Delta}_{rb}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}_{rb}' = \boldsymbol{\Xi}_{rb} - \mathbf{A}$, where $\mathbf{A} = \boldsymbol{\Delta}_{rb}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}_{rb}'$ is symmetric. For this $\boldsymbol{\Sigma}_{rb}$, it can be readily verified that $\boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb} = \boldsymbol{\Xi}_{rb} - 2\mathbf{A} + \mathbf{A}\boldsymbol{\Xi}_{rb}^{-1}\mathbf{A}$ and $\boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb} = \boldsymbol{\Xi}_{rb} - 3\mathbf{A} + 3\mathbf{A}\boldsymbol{\Xi}_{rb}^{-1}\mathbf{A} - \mathbf{A}\boldsymbol{\Xi}_{rb}^{-1}\mathbf{A}\boldsymbol{\Xi}_{rb}^{-1}\mathbf{A}$, so that $\boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb} \neq \boldsymbol{\Sigma}_{rb}\boldsymbol{\Xi}_{rb}^{-1}\boldsymbol{\Sigma}_{rb}$ in general. To get a null asymptotic distribution that is chi-square, a BAN estimator based on the variables in the subset b must be used.

5 Data examples

In this section we provide two numerical data examples to illustrate our results. In these examples we used Samejima's (1969) graded logistic model to fit questionnaire data using MLE. In the first example a small model is considered, $C = 3^5 = 243$, where the contingency table is not very sparse. In the second example we fit a larger model, $C = 5^{10} \approx 10^7$ to illustrate a highly sparse situation.

5.1 The Satisfaction with Life Scale data

The Satisfaction with Life Scale (SWLS: Diener, Emmons, Larsen & Griffin, 1985) is a widely used questionnaire consisting of $n = 5$ statements intended to obtain a global cognitive judgment of one's life. The SWLS is usually responded using a 7-point rating scale. However, Kramp (2005) investigated experimentally the effects of varying the number of response options in several rating scales, among them the SWLS. Here we shall fit Samejima's graded logistic model to an experimental version of the SWLS where respondents were asked to employ the following three point scale 0 = disagree, 1 = neither agree nor disagree, and 2 = agree. The sample size is $N = 429$, so the contingency table is not very sparse ($N/C = 1.77$). However, even in this situation 141 cells have zero counts. As a consequence of these zero observed counts the full information test statistics X^2 and G^2 yield very different conclusions: $X^2 = 310$, $p = 0.0002$ and $G^2 = 199$, $p = 0.91$, both on 227 degrees of freedom. The M_2 statistic, on the other hand, suggests that the model does not fit well, but not as poorly as X^2: $M_2 = 57.05$ on 35 degrees of freedom, $p = 0.01$. Notice that in this case, since we are using maximum likelihood estimation, $X^2 = M_5$.

As the model does not fit well, we proceed to investigate the source of the misfit. Large standardized cell residuals were obtained for the patterns (01222), (20122), (00210), (11211), (02000), (22120), (02200), (00122), (21000), (22102), (10202), (00021), (22010). We can not meaningfully extract any trend in these patterns. As an alternative, we computed goodness-of-fit statistics for bivariate subtables.

Each bivariate table depends on $2(K-1)$ intercepts and 2 slopes. Thus, there are $(K^2-1)-2(K-1)-2 = 2$ degrees of freedom when M_2 is applied to bivariate subtables. We can not assess how well this model fits each item separately using M_1 as the univariate submodels are not identified. There are $K - 1$ mathematically independent probabilities in each univariate table. But each univariate table depends on $K - 1$ intercepts α and one slope β.

We provide in Table 4 the bivariate statistics computed for every pair of variables. As can be seen in this Table, the pairwise $M_2^{(ij)}$ statistics suggest that the model does not fit well for item 2. To verify this conjecture we re-estimated the model to each subset of $n - 1 = 4$ items. The results are shown in Table 5.

INSERT TABLES 4 AND 5 ABOUT HERE

The results of this table strongly suggest that the fit of Samejima's graded logistic model to these data can be improved by removing item 2, as suggested by the bivariate $M_2^{(ij)}$ statistics. Also notice that in this table there are some discrepancies between $X^2 = M_4^{(-i)}$ and $M_2^{(-i)}$ with the ith item deleted, which suggests, given our simulation results in Section 3.2, that sparseness can still have some adverse effects on the small sample behavior of X^2 even in such small tables.

5.2 The Negative Problem Orientation data

Following Drasgow et al. (1995), Maydeu-Olivares (in press, b) used X^2 statistics for single items, item pairs and item triplets to compare in a descriptive fashion the fit of several unidimensional IRT models to each of the five scales of the Social Problem Solving Inventory-Revised (SPSI-R: D'Zurilla, Nezu & Maydeu-Olivares, 2002). The models considered were Samejima's graded logistic model, Masters's (1982) partial credit model, Thissen and Steinberg's (1986) extension of the latter, and Bock's (1972) nominal model. In all scales Samejima's graded logistic model yielded the best fit. However, since the statistics employed to compare the models had an unknown sampling distribution, nothing could be concluded about the absolute fit of the models. In this example, we shall re-analyze Maydeu-Olivares' (in press, b) data from one of the SPSI-R scales, the Negative Problem Orientation (NPO) scale, to investigate whether the best fitting model, Samejima's graded logistic model, indeed provides an adequate fit to the data.

The NPO scale consists of 10 items intended to measure individual differences in (a) viewing a problem as a significant threat to well-being, (b) doubting one's personal ability to solve problems successfully, and (c) easily becoming frustrated and upset when confronted with problems in everyday living. Individuals are asked to respond to each item using one of five categories: "0 = Not at all true of me", "1 = Slightly true of me", "2 = Moderately true of me", "3 = Very true of me", "4 = Extremely true of me". The sample size is $N = 1053$.

Samejima's graded logistic model was estimated by maximum likelihood. The parameter estimates and standard errors are reported in Table 6. The number of degrees of freedom available for testing using X^2 and G^2 is very large, $df = 9765574$, and each statistic offers a very different picture: $X^2 \approx 6 \times 10^7 \gg df$, $G^2 \approx 13000 \ll df$. Given the extremely large degree of sparseness of the data, neither statistic can be trusted and we resort to M_2 to assess the overall fit of the model. With 710 degrees of freedom we obtained $M_2 \approx 1500$, $p \ll .001$. Thus, the model fits very poorly.

To assess the source of misfit we used as in the previous example pairwise $M_2^{(ij)}$ statistics, each with 2 degrees of freedom. These statistics are shown in Table 7. In this table, we used a Bonferroni adjustment for the $M_2^{(ij)}$ statistics. Thus, those statistics that exceed 35.82, the upper $0.05/45 = .0011$ quantile of the χ^2_{14} distribution, are indicated with an asterisk.

INSERT TABLES 6 AND 7 ABOUT HERE

Even with this correction, Table 7 reveals that the misfit of the model can not be attributed to any particular item. Rather, it is widespread. Thus, we conclude that although Maydeu-Olivares (in press, b) results suggest that Samejima's logistic graded model was the best fitting model for these data among a set of parametric IRT models, this model does not provide a satisfactory fit to this questionnaire. An alternative model is needed.

6 Discussion and conclusions

Applied researchers confronted with the problem of modeling sparse multidimensional contingency tables are faced with the problem of how to assess the overall goodness of fit of the model, and should the overall fit be poor how to identify the source of the misfit. In this paper we have extended previous work by Maydeu-Olivares and Joe (2005) on limited information goodness-of-fit testing of composite hypotheses in multidimensional binary contingency tables to multidimensional contingency tables of arbitrary dimensions. We have shown that their M_r family of overall goodness-of-fit statistics extends readily to the general case. Provided that the model is identified from the margins up to order r, M_r is asymptotically chi-squared distributed for any $\sqrt{N}$-consistent and asymptotically normal estimator. The simulations presented in this paper suggest that in large and/or sparse contingency tables M_r for small r ($r = 2, 3$) should be employed instead of X^2 as the former have more precise empirical Type I errors and may be more powerful than the latter.

Also, to assess the source of misfit, we have suggested employing the M_r statistic for r-dimensional subtables. Provided the subtable's model is identified, the $M_r^{(b)}$ statistics are asymptotically chi-square with degrees of freedom equal to the number of cells in the subtable minus the number of parameters involved in the subtable minus one. This result holds for any $\sqrt{N}$-consistent and asymptotically normal estimator. Furthermore, we have shown that X^2 applied to subtables is not asymptotically chi-square even for the MLE. These $M_r^{(b)}$ statistics applied to subtables may be very useful to identify the source of the misfit in multidimensional contingency tables where the number of categories is large.

The family of statistics M_r compares favorably to the use of resampling methods for goodness-of-fit assessment of composite null hypothesis in multidimensional contingency tables. On the one hand, one can obtain a p-value for the overall fit of the model with considerably less computing effort than by resampling methods. On the other hand, they provide a way to detect the source of misfit of the model. However, there are two obvious limitations to the use of the approach advocated here. First, the model must be identified from the margins. In practice, most models of interest — such as the IRT model considered here — can be identified from the bivariate or trivariate margins. The second limitation is computational. When some of the observed variables have a large number of categories K_i, even computing M_2 for $n > 15$ can be computationally infeasible as the dimension $s(2)$ of the matrices $\mathbf{\Xi}_r$ and $\mathbf{\Delta}_r$ gets too large.

When the categorical data are ordinal, then there exists an alternative set of limited information test statistics that are invariant to the set of permissible transformations of the ordinal data and that can be used with much larger models than those feasible using the M_r family of test statistics. This alternative approach suitable only for ordinal variables will be discussed in a separate report.

In closing, while we have focused on testing composite hypotheses, the common situation in applications, the general framework discussed here can also be applied to testing simple null hypotheses. This is discussed in the Appendix. Also, although the applications and simulations in this paper are focused on item response models, the theory introduced in this paper is completely general for multivariate discrete data. For example, for multivariate continuous variables with a copula model (e.g., Joe 1997) there is no general approach for assessing goodness-of-fit other than discretizing the variables. Applying the family of M_r statistics to discretized continuous variables to assess goodness-of-fit is another topic for future investigation.

Appendix

Goodness-of-fit testing of simple null hypotheses

For testing the overall goodness-of-fit of a simple null hypotheses $H_0 : \boldsymbol{\pi}(\boldsymbol{\theta})$ for a fixed a priori vector $\boldsymbol{\theta}$ of dimension q, and as an alternative to X^2 in sparse tables, Maydeu-Olivares and Joe (2005) proposed using the family of limited information test statistics

$$L_r = N(\mathbf{p}_r - \boldsymbol{\pi}_r)' \boldsymbol{\Xi}_r^{-1} (\mathbf{p}_r - \boldsymbol{\pi}_r), \quad r = 1, \ldots, n.$$

The choice of r depends on the sparseness of the contingency table. From (2.4), under H_0, the L_r statistics converge in distribution to a $\chi^2_{s(r)}$ distribution as $N \to \infty$. For $r = n$, $L_n = X^2$.

If the L_r test suggests significant misfit marginal then $L_r^{(b)} = X_b^2$ for the r-dimensional subtables can be obtained to identify the source of the misfit. Under the null hypothesis, these statistics applied to subtables are asymptotically chi-square; the degrees of freedom for margin b is $[\prod_{i \in b} K_i] - 1$.

Some computing details

Consider a model, such as that given in (3.7–3.9), that has a form that is closed under margins. Then any probability in the rth order margin and in $\boldsymbol{\Xi}_r$ can be computed directly without marginalizing the n-dimensional joint distribution. That is, for computations, one can avoid the large matrix $\mathbf{T}_r$ in Section 2.2, where it was presented for notational convenience. M_r depends on $\mathbf{C}_r$ evaluated at $\tilde{\boldsymbol{\theta}}$, which depends on the matrices $\boldsymbol{\Delta}_r$ and $\boldsymbol{\Xi}_r$ evaluated at $\tilde{\boldsymbol{\theta}}$. The matrix of partial derivatives $\boldsymbol{\Delta}_r$ can be computed at the same time as $\boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})$. The computation of $\boldsymbol{\Xi}_r$ is a bit more involved. $\boldsymbol{\Xi}_r$ is the covariance matrix of the vector of sample proportions of margins of order r or less. A term in $\boldsymbol{\Xi}_r$ has the form $m_c(\mathbf{y}_c) - m_a(\mathbf{y}_a) m_b(\mathbf{y}_b)$, where a, b are subsets of $\{1, \ldots, n\}$ of dimension between 1 and r, $c = a \cup b$, and m_a, m_b, m_c are marginal probabilities. Efficient computation of $\boldsymbol{\Xi}_r$ relies on a systematic way of enumerating the marginal probabilities corresponding to terms in $\mathcal{A}_r$.

Next we discuss computation of the MLE for model (3.7–3.9). There are similar considerations for other item response models. With Gauss-Hermite quadrature for evaluation of marginal probabilities of (3.7) and its derivatives, we have coded the computation of the MLE with the Newton-Raphson method. This is an alternative to the EM algorithm (e.g., Bock & Aitkin, 1981). It has the advantage that the inverse observed Fisher information matrix, used as the estimated covariance matrix, is computed at the same time. For the covariance matrix of $\hat{\mathbf{e}}_r$ in (2.7), the expected Fisher information matrix is needed. Computing the information matrix is much harder than computing $\boldsymbol{\Xi}_r$ because the former requires summing through the probabilities in (3.7) for all $C = K^n$ n-dimensional probabilities, and this is essentially only feasible if $K^n < 10^9$. With most efficient use of computer memory, Fisher information $\boldsymbol{\mathcal{I}}$ can be computed as $\sum_{\mathbf{y}} \frac{\partial \pi_{\mathbf{y}}}{\partial \boldsymbol{\theta}} \left(\frac{\partial \pi_{\mathbf{y}}}{\partial \boldsymbol{\theta}} \right)' / \pi_{\mathbf{y}}$.

The BCL estimator for model (3.7–3.9) can also be obtained with Gauss-Hermite quadrature and the Newton-Raphson method. The computations require the evaluation of marginal probabilities of (3.7) and its derivatives for dimensions 2, 3, and 4. We next indicate how to evaluate $\widetilde{\boldsymbol{\Sigma}}_2$ in (2.10), without any matrices of order C. This technique applies to any $\sqrt{N}$-consistent estimator that can be considered as a solution to a set of estimating equations. Let $\pi^{(ij)}_{k_1k_2}(\boldsymbol{\theta}) = \Pr(Y_i = k_1, Y_j = k_2)$ and $p^{(ij)}_{k_1k_2}$ be the sample counterpart. Then the BCL estimator $\tilde{\boldsymbol{\theta}}$ maximizes

$$L_2(\boldsymbol{\theta}) = N \sum_{i<j} \sum_{k_1} \sum_{k_2} p^{(ij)}_{k_1k_2} \log \pi^{(ij)}_{k_1k_2}(\boldsymbol{\theta}).$$

From second order Taylor approximation,

$$\mathbf{0} = N^{-1}\frac{\partial L_2(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = N^{-1}\frac{\partial L_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + N^{-1}\frac{\partial^2 L_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) + O_p(N^{-1})$$

and using the theory of estimating equations (see for example Zhao & Joe 2005),

$$\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} \approx \mathbf{W}^{-1}N^{-1}\frac{\partial L_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \qquad \text{where} \qquad \mathbf{W} = \mathbf{W}(\boldsymbol{\theta}) = -N^{-1}\mathrm{E}\left[\frac{\partial^2 L_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right].$$

Note that

$$\begin{aligned} N^{-1}\frac{\partial L_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{i<j} \sum_{k_1} \sum_{k_2} p^{(ij)}_{k_1k_2} \frac{\partial \pi^{(ij)}_{k_1k_2}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big/ \pi^{(ij)}_{k_1k_2}(\boldsymbol{\theta}) \\ &= \sum_{i<j} \sum_{k_1} \sum_{k_2} \Big[p^{(ij)}_{k_1k_2} - \pi^{(ij)}_{k_1k_2}(\boldsymbol{\theta})\Big] \frac{\partial \pi^{(ij)}_{k_1k_2}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big/ \pi^{(ij)}_{k_1k_2}(\boldsymbol{\theta}) \qquad (A.1) \end{aligned}$$

Let $\boldsymbol{\pi}^*_2$ be a vector containing all model-based bivariate marginal probabilities (including those with 0 indices). Also, let $\mathbf{p}^*_2$ be its sample counterpart. From (A.1), there is a matrix $\mathbf{K}$ such that $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} \approx \mathbf{W}^{-1}\mathbf{K}(\mathbf{p}^*_2 - \boldsymbol{\pi}^*_2)$. From Section 2, each element of $\boldsymbol{\pi}^*_2$ is either an element of $\boldsymbol{\pi}_2$ or a linear function of elements of $\boldsymbol{\pi}_2$. Hence $\mathbf{p}^*_2 - \boldsymbol{\pi}^*_2 = \mathbf{S}(\mathbf{p}_2 - \boldsymbol{\pi}_2)$ for a matrix $\mathbf{S}$. Putting everything together, $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta} \approx \mathbf{H}_2(\mathbf{p}_2 - \boldsymbol{\pi}_2)$ where $\mathbf{H}_2 = \mathbf{W}^{-1}\mathbf{K}\mathbf{S}$. Since only probabilities in $\boldsymbol{\pi}_2$ are involved, $\widetilde{\boldsymbol{\Sigma}}_2$ in (2.10) can be written as

$$\widetilde{\boldsymbol{\Sigma}}_2 = \boldsymbol{\Xi}_2 - \boldsymbol{\Delta}_2\mathbf{H}_2\boldsymbol{\Xi}_2 - \boldsymbol{\Xi}_2\mathbf{H}_2'\boldsymbol{\Delta}_2' + \boldsymbol{\Delta}_2[\mathbf{H}_2\boldsymbol{\Xi}_2\mathbf{H}_2']\boldsymbol{\Delta}_2'$$

For computer implementation in all of the above, a systematic way is needed to convert a multi-indexed margin to a row or column index in a matrix.

When using Gauss-Hermite quadrature for ML estimation, one must be careful in the simulation of (3.7)–(3.9) for the assessment of the null distribution of M_r. For a fixed number of quadrature points n_q, the accuracy decreases as the slope parameters increase in absolute value. This is checked by comparing Romberg integration with Gauss-Hermite quadrature. Hence, the number of quadrature points needs to increase as the slope parameter increases in order to achieve a desired accurary; $n_q = 48$ is acceptable provided β values do not exceed 3 in absolute value.

The null distribution of X^2 and M_2 depends on the simulation method if the MLE (or another estimator) is obtained based on Gauss-Hermite quadrature (GH) of the model probabilities. Rather than a standard normal latent random variable Z, GH calculations with n_q quadrature points implicitly assume that the latent rv Z' is discrete with mass w_i at point x_i for $i = 1, \ldots, n_q$ (note that $\sum_i w_i = 1$). Hence if simulating with Z and estimating and calculating M_2 and X^2 with Z', the resulting 'null distribution' of M_2 and X^2 will be stochastically larger than the (asymptotic) χ^2 distribution if the sample size N is large (relative to number of vector categories C). This is because Z is different from Z' and the goodness-of-fit statistics can discriminate these two for large N. If estimation is based on GH with Z', then simulation with Z means that a non-null model that is close to null is used, and the M_2 and X^2 statistics will tend to be a bit larger than simulation with Z'. A rough calculation shows that the distribution of M_r in this case is approximately non-central chi-square with noncentrality parameter $N\boldsymbol{\delta}_r'\mathbf{C}_r\boldsymbol{\delta}_r$ where $\boldsymbol{\delta}_r$ is the vector is differences in marginal moments up to order r for probabilities based on latent variables Z and Z'. This behavior was readily seen in simulation results of Z versus Z'.

References

Agresti, A. (2002). *Categorical Data Analysis* (second edition). New York: Wiley.

Bartholomew, D.J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis.* (second edition). London: Arnold.

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. & Tzamourani, P. (1999). The goodness-of-fit of latent trait models in attitude measurement. *Sociolological Methods and Research*, **27**, 525–546.

Bentler, P.M. (1995). *EQS.* Encino, CA: Multivariate Software Inc.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, **37**, 29–51.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press.

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

Cochran, W.G. (1952). The X^2 test of goodness of fit. *Annals of Mathematical Statistics*, **23**, 315–345.

Collins, L.M., Fidler, P.L., Wugalter, S.E. & Long, J. (1993). Goodness of fit testing for latent class models. *Multivariate Behavioral Research*, **28**, 375–389.

Diener, E., Emmons, R.A., Larsen, R.J. & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, **49**, 71–75.

Drasgow, F., Levine, M.V., Tsien, S., Williams, B., & Mead, A. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, **19**, 143–165.

D'Zurilla, T.J., Nezu, A.M. & Maydeu-Olivares, A. (2002). *Manual of the Social Problem-Solving Inventory-Revised*. North Tonawanda, NY: Multi-Health Systems, Inc.

Godambe, V.P. (Ed.) (1991). *Estimating Functions*. Oxford: Oxford University Press.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman and Hall.

Jöreskog, K.G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, **59**, 381–389.

Jöreskog, K.G. & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, **36**, 347–387.

Jöreskog, K.G. & Sörbom, D. (2001). *LISREL 8*. Chicago, IL: Scientific Software.

Koehler, K. & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, **75**, 336–344.

Kramp, U. (2005). *Effects of the number of response options on personality rating scales*. Unpublished doctoral dissertation. University of Barcelona.

Lee, S.Y., Poon, W.Y., & Bentler, P.M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, **48**, 339–358.

Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149–174.

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

Maydeu-Olivares, A. (2001b). Multidimensional item response theory modeling of binary data: Large sample properties of NOHARM estimates. *Journal of Educational and Behavioral Statistics*, **26**, 49–69.

Maydeu-Olivares, A. (in press, a). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*.

Maydeu-Olivares, A. (in press, b). Further empirical results on parametric vs. non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research*.

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**, in press.

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

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, **49**, 115–132.

Muthén, B. (1993). Goodness of fit with categorical and other non normal variables. In K.A. Bollen & J.S. Long (Eds.) *Testing Structural Equation Models* (pp. 205–234). Newbury Park, CA: Sage.

Muthén, L. & Muthén, B. (2001). *MPLUS*. Los Angeles, CA: Muthén & Muthén.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. New York: Wiley.

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

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

Reiser, M. & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, **47**, 85–107.

Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Schott, J.R. (1997). *Matrix Analysis for Statistics*. New York: Wiley.

Teugels, J.L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *Journal of Multivariate Analysis*, **32**, 256–268.

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.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, **51**, 567–577.

van der Linden, W. J. & Hambleton, R.K. (Eds.) (1997). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

Zhao, Y. & Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*, to appear.

Table 1: Small sample distribution for X^2 and M_2 with ML estimation. Mean, variance and exceedances of asymptotic upper 0.2, 0.1, 0.05, 0.01 quantiles.

n	K	N	statistic	df	mean	var.	$\alpha = .2$	$\alpha = .1$	$\alpha = .05$	$\alpha = .01$
5	3	300	X^2	227	227.4	460.8	.21	.11	.058	.011
			M_2	35	35.1	67.1	.20	.10	.051	.009
5	3	1000	X^2	227	227.5	487.4	.21	.12	.066	.015
			M_2	35	34.9	71.8	.21	.11	.045	.006
5	3	3000	X^2	227	227.1	470.5	.19	.10	.056	.011
			M_2	35	35.1	70.4	.22	.11	.050	.005
5	5	300	X^2	3099	3094	75200	.35	.31	.27	.21
			M_2	155	155	300	.20	.10	.042	.005
5	5	1000	X^2	3099	3097	23675	.30	.24	.19	.11
			M_2	155	155	311	.20	.10	.053	.010
5	5	3000	X^2	3099	3097	12108	.26	.16	.10	.04
			M_2	155	155	301	.20	.09	.042	.012
10	5	300	X^2	9765574	9.68×10^6	6.07×10^{12}	.37	.37	.37	.36
			M_2	710	711	1482	.21	.11	.064	.011
10	5	1000	X^2	9765574	9.73×10^6	1.13×10^{12}	.42	.42	.42	.41
			M_2	710	710	1339	.18	.10	.055	.008
10	5	3000	X^2	9765574	9.77×10^6	3.60×10^{11}	.48	.48	.48	.48
			M_2	710	708	1315	.19	.08	.039	.008

NOTE: 1000 replications. MVM model given in (3.7–3.9).

Table 2: Small sample distribution for M_2 with bivariate composite likelihood estimator. Mean, variance and exceedances of asymptotic upper 0.2, 0.1, 0.05, 0.01 quantiles.

n	K	N	statistic	df	mean	var.	$\alpha = .2$	$\alpha = .1$	$\alpha = .05$	$\alpha = .01$
5	3	300	M_2	35	35.1	67.0	.20	.10	.050	.008
5	3	1000	M_2	35	34.9	71.7	.21	.10	.045	.006
5	3	3000	M_2	35	35.1	70.4	.22	.11	.049	.005
5	5	300	M_2	155	155	299	.20	.10	.041	.005
5	5	1000	M_2	155	155	309	.19	.10	.050	.009
5	5	3000	M_2	155	155	301	.20	.09	.041	.013
10	5	300	M_2	710	711	1435	.20	.11	.056	.009
10	5	1000	M_2	710	710	1327	.18	.10	.051	.007
10	5	3000	M_2	710	708	1310	.19	.09	.036	.009

Note: 1000 replications. MVM model given in (3.7–3.9).

Table 3: Small sample power for M_2 vs X^2 with MLE estimator; MVM model given in (3.7–3.9) with a common slope parameter for the null hypothesis. Exceedances of asymptotic upper 0.2, 0.1, 0.05, 0.01 quantiles based on 1000 replications.

$\boldsymbol{\alpha}$	β (alternative)	N	statistic	$\alpha = .2$	$\alpha = .1$	$\alpha = .05$	$\alpha = .01$
1 -1, 1 -1, 1 -1, 1 -1, 1 -1	1 1 1 .5 .5	500	X^2	0.37	0.21	0.12	0.04
			M_2	0.60	0.42	0.31	0.14
1 -1, 1 -1, 1 -1, 1 -1, 1 -1	1 1 1 1 .5	500	X^2	0.36	0.20	0.13	0.03
			M_2	0.56	0.40	0.29	0.11
1 -1, 1 -1, 1 -1, 1 -1, 1 -1	1 .9 .8 .9 .8	1200	X^2	0.24	0.11	0.05	0.03
			M_2	0.26	0.14	0.09	0.02
1 -1, 1 -1, 1 -1, .5 -.5, .5 -.5	1 1 1 1 .5	500	X^2	0.31	0.18	0.09	0.03
			M_2	0.54	0.37	0.26	0.11

Table 4: $M_2^{(ij)}$ statistics applied to bivariate subtables for the SWLS data

items	1	2	3	4	5
1	–	0.03	6.00	6.19	0.08
2	0.03	–	1.41	9.13*	8.71*
3	6.00	1.41	–	0.83	0.05
4	6.19	9.13*	0.83	–	1.52
5	0.08	8.71*	0.05	1.52	–

Note: statistics significant at the $\alpha = 0.05$ significance level are marked with a $*$.

Table 5: Overall goodness-of-fit results for subsets of $n-1=4$ items for the SWLS data

dropping	X^2	p	M_2	p
1	86.98	0.06	24.97	0.20
2	64.59	0.59	22.91	0.29
3	91.93	0.03	35.70	0.02
4	82.20	0.12	36.37	0.01
5	105.36	< 0.01	30.06	0.07

Notes: $X^2 = M_4^{(-i)}$; there are 68 degrees of freedom for X^2 and 20 for $M_2^{(-i)}$ (ith item deleted).

Table 6: Maximum likelihood estimates and SEs for the NPO data

parameters	estimates	SEs
$\alpha_{11},\dots,\alpha_{14}$	$1.97, 0.14, -1.52, -3.16$	$0.11, 0.09, 0.10, 0.15$
$\alpha_{21},\dots,\alpha_{24}$	$2.05, -0.06, -1.94, -3.99$	$0.12, 0.10, 0.12, 0.19$
$\alpha_{31},\dots,\alpha_{34}$	$2.29, 0.15, -1.39, -3.28$	$0.12, 0.09, 0.10, 0.15$
$\alpha_{41},\dots,\alpha_{44}$	$2.15, -0.02, -1.63, -3.52$	$0.12, 0.09, 0.11, 0.16$
$\alpha_{51},\dots,\alpha_{54}$	$0.89, -0.92, -2.46, -4.22$	$0.10, 0.10, 0.13, 0.19$
$\alpha_{61},\dots,\alpha_{64}$	$2.92, 0.73, -0.90, -3.05$	$0.14, 0.10, 0.10, 0.15$
$\alpha_{71},\dots,\alpha_{74}$	$1.65, -0.68, -2.52, -4.51$	$0.13, 0.11, 0.15, 0.22$
$\alpha_{81},\dots,\alpha_{84}$	$1.63, -0.18, -1.34, -2.76$	$0.10, 0.09, 0.10, 0.13$
$\alpha_{91},\dots,\alpha_{94}$	$1.12, -0.86, -2.52, -4.65$	$0.12, 0.12, 0.15, 0.22$
$\alpha_{10,1},\dots,\alpha_{10,4}$	$2.33, -0.37, -2.28, -4.63$	$0.14, 0.12, 0.15, 0.22$
$\beta_1,\dots,\beta_5$	$1.57, 2.06, 1.78, 1.71, 1.74$	$0.09, 0.11, 0.10, 0.10, 0.10$
$\beta_6,\dots,\beta_{10}$	$2.02, 2.43, 1.51, 2.47, 2.57$	$0.11, 0.13, 0.09, 0.14, 0.14$

Table 7: Bivariate subtable $M_2^{(ij)}$ for the NPO data

items	1	2	3	4	5	6	7	8	9	10
1	-	48.37*	23.35	31.37	12.84	39.73*	34.63*	26.67	51.11*	19.87
2	48.37*	-	55.31*	28.64	19.49	30.13	15.43	21.35	18.31	33.08
3	23.35	55.31*	-	28.40	24.46	36.47*	31.14	27.08	37.92*	35.73
4	31.37	28.64	28.40	-	31.67	57.42*	29.69	32.32	34.99	49.28*
5	12.84	19.49	24.46	31.67	-	30.51	32.70	46.82*	38.77*	27.45
6	39.73*	30.13	36.47*	57.42*	30.51	-	31.43	50.55*	50.97*	55.49*
7	34.63	15.43	31.14	29.69	32.70	31.43	-	45.23*	58.72*	92.67*
8	26.67	21.35	27.08	32.32	46.82*	50.55*	45.23*	-	25.75	29.34
9	51.11*	18.31	37.92*	34.99	38.77*	50.97*	58.72*	25.75	-	35.89*
10	19.87	33.08	35.73	49.28*	27.45	55.49*	92.67*	29.34	35.89*	-

Note: statistics significant at the $\alpha = 0.05/45 = .0011$ significance level are marked with a $*$.