LIMITED AND FULL INFORMATION ESTIMATION AND GOODNESS-OF-FIT TESTING IN 2^n CONTINGENCY TABLES

IE Working Paper **MK8-105-I** **05 / 05 / 2003**

Alberto Maydeu Olivares

Instituto de Empresa
Marketing Dept.
C / María de Molina 11-15,
28006 – Madrid
España
Alberto.Maydeu@ie.edu

Harry Joe

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

Abstract

High-dimensional contingency tables tend to be sparse and standard goodness-of-fit statistics such as X^2 cannot be used without pooling categories. As an improvement on arbitrary pooling, for goodness-of-fit of large 2^n contingency tables, we propose a class of quadratic form statistics based on the residuals of margins or multivariate moments up to order r. Further, the marginal residuals are useful for diagnosing lack of fit of parametric models. These classes of test statistics are asymptotically chi-square and have better small sample properties than X^2. We also show that these classes of test statistics have better power than X^2 for some useful multivariate binary models. Related to this class of test statistics is a class of limited information estimators based on low-dimensional margins. We show that these estimators have high efficiency for one commonly used latent trait model for binary data.

Keywords

item response modeling, quadratic form statistics, low-dimensional margins, limited information, multivariate Bernoulli distribution

1 Introduction

It is common in the Social Sciences to encounter 2^n contingency tables, where n can be as large as several hundreds. These tables arise for instance by collecting the responses of a sample of individuals to a survey, a personality inventory, or an educational test consisting of n items, each with two possible responses. A researcher confronted to the problem of modeling such a 2^n contingency table faces several challenges. Perhaps the most important challenge is how to assess the overall goodness-of-fit of the hypothesized model. For large n, most often binary contingency tables become sparse and the empirical Type I error rates of X^2 and G^2 test statistics do not match their expected rates under their asymptotic distribution. This problem can be overcome by generating the empirical sampling distribution of the statistic using the parametric bootstrap method (e.g., Collins et al, 1993; Bartholomew & Tzamourani, 1999). However, this approach may be very time consuming if the researcher is interested in comparing the fit of several models.

If, as it is often the case, the overall tests suggests significant misfit, a second challenge that a researcher must confront is to identify the source of the 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 & 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 these two challenges, numerous authors, particularly in Psychometrics, have advocated using residuals for pairs and triplets of variables to assess the goodness-of-fit in 2^n contingency tables. Some key references in these literature are Reiser (1996), Reiser & Lin (1999), Reiser & VanderBergh (1994), Bartholomew & Tzamourani (1999), and Bartholomew & Leung (2002).

A third challenge a researcher may face when dealing with large binary tables is a parameter estimation problem. Take for instance latent trait models (for an overview see Bartholomew & Knott, 1999) which are extremely popular in the Social Sciences. If the distribution of the latent traits is assumed to be multivariate normal, as it is most often the case, computing the binary pattern probabilities is very difficult as the number of latent traits increases. However, estimation for these models using only univariate and bivariate information is relatively straightforward. There is a long tradition in Psychometrics of employing estimation methods that only use information from the low order marginals of the table (e.g., Christoffersson, 1975; Muthén, 1978, 1984, 1993). Here, we refer to testing and estimation methods that only use the information contained in the low order margins of the contingency table as *limited information* methods. There have also been some proposals in Statistics in using limited information methods (Joe, 1996: Chapter 10). Limited information

methods naturally yield limited information testing procedures, whose asymptotic properties are well known (see Christoffersson, 1975; Muthén, 1978, 1993; Maydeu-Olivares, 2001). However, the asymptotic distribution of full information test statistics when the parameters have been estimated using limited information procedures has never been studied.

What is needed is a unified framework for limited information estimation and testing in 2^n contingency tables. We provide such a framework in this paper under multivariate Bernoulli sampling. In Section 2, we provide a convenient representation of the multivariate Bernoulli (MVB) distribution using its joint moments. From the asymptotic distribution of sample joint moments (marginal proportions), we obtain the asymptotic distribution of marginal residuals. In Section 3, a family of limited information quadratic form statistics, based on these marginal residuals, to assess the goodness-of-fit of simple null hypotheses is proposed. These statistics are asymptotically chi-square distributed, and Pearson's full information X^2 statistic is a special case of this family. In Section 4, we extend the results of Section 3 to composite null hypotheses, the common situation for applications. Two classes of estimators are considered: (a) minimum variance full information estimators such as maximum likelihood, and (b) consistent and asymptotically normal estimators. The latter includes limited information estimators. A family of limited information goodness-of-fit test statistics is proposed whose members are asymptotically chi-square for both classes of estimators. In order to study asymptotic power of our new statistics, we derive results for the asymptotic distribution under a sequence of local alternatives for testing one form of a nested null model. In Section 5, a family of limited information estimators, closely linked to our proposed family of limited information goodness-of-fit tests, is proposed. These estimators are computationally advantageous when the multivariate binary probabilities are difficult to compute. We show that these estimators are highly efficient for one common latent trait model. Section 6 has an example of binary item response data from Bartholomew & Knott (1999) to illustrate our results. Finally, Section 7 has conclusions and a discussion of further research.

2 Multivariate Bernoulli (MVB) distributions and asymptotic distribution of sample moments

In this section, we give a characterization of the MVB distribution in terms of multivariate moments, and define the notation used in the remainder of this paper.

Consider an n-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ of Bernoulli random variables, with $\pi_i = \Pr(Y_i = 1)$, $i = 1, \ldots, n$, and joint distribution:

$$\pi_{\mathbf{y}} = \Pr(Y_i = y_i, i = 1, \ldots, n) \quad \mathbf{y} = (y_1, \ldots, y_n), \quad y_i \in \{0, 1\}. \tag{2.1}$$

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 2^n joint probabilities. One convenient ordering of the elements of $\boldsymbol{\pi}(\boldsymbol{\theta})$ is by order of the values of $\mathbf{y}'\mathbf{1} = 0, 1, \ldots, n$, and by lexicographical ordering within a constant sum. An example with $n = 3$ is given below.

The n-variate Bernoulli distribution may be alternatively characterized by the (2^n-1)-dimensional vector $\dot{\boldsymbol{\pi}}$ of its joint moments (Teugels, 1990); $\dot{\boldsymbol{\pi}}' = (\dot{\boldsymbol{\pi}}_1', \dot{\boldsymbol{\pi}}_2', \ldots, \dot{\boldsymbol{\pi}}_n')'$ where $\dot{\boldsymbol{\pi}}_1' = (\pi_1, \ldots, \pi_n)'$, $\dot{\boldsymbol{\pi}}_2$ is the $\binom{n}{2}$-dimensional vector of bivariate moments with elements $\mathrm{E}(Y_iY_j) = \Pr(Y_i = 1, Y_j = 1) = \pi_{ij}$, $j < i$, and so on up to $\dot{\boldsymbol{\pi}}_n = \mathrm{E}(Y_1 \cdots Y_n) = \Pr(Y_1 = \cdots = Y_n = 1)$.

There is a $(2^n - 1) \times 2^n$ matrix $\mathbf{T}$ of 1s and 0s, of full row rank, such that $\dot{\boldsymbol{\pi}} = \mathbf{T}\boldsymbol{\pi}$. $\mathbf{T}$ is an upper triangular matrix if $\boldsymbol{\pi}$ is ordered as described above. For example for $n = 3$, one has

$$\begin{pmatrix} \dot{\pi}_1 \\ \dot{\pi}_2 \\ \dot{\pi}_3 \\ \cdots \\ \dot{\pi}_{12} \\ \dot{\pi}_{13} \\ \dot{\pi}_{23} \\ \cdots \\ \dot{\pi}_{123} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ & & & & \cdots & & & \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ & & & & \cdots & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_{000} \\ \pi_{100} \\ \pi_{010} \\ \pi_{001} \\ \pi_{110} \\ \pi_{101} \\ \pi_{011} \\ \pi_{111} \end{pmatrix}.$$

The first column of $\mathbf{T}$ is a column of zeros, so we can partition $\mathbf{T} = (\,\mathbf{0} \quad \dot{\mathbf{T}}\,)$ and $\dot{\boldsymbol{\pi}} = \dot{\mathbf{T}}\check{\boldsymbol{\pi}}$, with $\boldsymbol{\pi} = \begin{pmatrix} \pi_{0\cdots0} \\ \check{\boldsymbol{\pi}} \end{pmatrix}$. Since $\pi_{0\cdots0} = 1 - \mathbf{1}'\check{\boldsymbol{\pi}}$, the inverse relationship between $\dot{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ is

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

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

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

Furthermore, the vector of joint moments of the multivariate Bernoulli distribution up to order $r \leq n$, 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 $\boldsymbol{\pi}_n = \dot{\boldsymbol{\pi}}$.

For a random sample of size N from (2.1), let $\mathbf{p}$ and $\dot{\mathbf{p}}$ denote the 2^n-dimensional vector of cell proportions, and the $(2^n - 1)$-dimensional vector of sample joint moments, respectively. Then

$$\sqrt{N}\,(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}) = \mathbf{T}\sqrt{N}\,(\mathbf{p} - \boldsymbol{\pi}). \tag{2.2}$$

Since (Agresti, 1990)

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

it follows from (2.2) that

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

Let $\dot{p}_a$ and $\dot{p}_b$ be any two elements of $\dot{\mathbf{p}}$ (not necessarily univariate proportions). Then, the elements of $\boldsymbol{\Xi}$ are of the form $N\mathrm{Var}\,(\dot{p}_a)=\dot{\pi}_a(1-\dot{\pi}_a)$, $N\mathrm{Cov}\,(\dot{p}_a,\dot{p}_b)=\dot{\pi}_{a\cup b}-\dot{\pi}_a\dot{\pi}_b$, so that for example when $n\geq 3$, for $i\neq j$, $j=k$, $N\mathrm{Var}\,(\dot{p}_{ij})=\dot{\pi}_{ij}(1-\dot{\pi}_{ij})$, and $N\mathrm{Cov}\,(\dot{p}_{ij},\dot{p}_k)=\dot{\pi}_{ij}-\dot{\pi}_{ij}\dot{\pi}_k=\dot{\pi}_{ij}(1-\dot{\pi}_k)$; whereas for i,j,k distinct, $N\mathrm{Cov}\,(\dot{p}_{ij},\dot{p}_k)=\dot{\pi}_{ijk}-\dot{\pi}_{ij}\dot{\pi}_k$.

Also, let $\mathbf{p}_r$ be the vector of sample moments up to order r; it has dimension $s=s(r)=\sum_{i=1}^{r}\binom{n}{i}$. Then,

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

Since $\mathbf{T}_r$ is of full row rank s, $\boldsymbol{\Xi}_r$ is also of full rank s (see Rao 1973: p. 30).

3 Limited information tests of simple null hypotheses

Consider a simple null hypotheses $H_0:\boldsymbol{\pi}=\boldsymbol{\pi}_0$. The two statistics most widely used in this situation are the likelihood ratio test statistic, $G^2=2N\sum_c p_c\ln[p_c/\pi_c]$, and Pearson's test statistic, $X^2=N\sum_c(p_c-\pi_c)^2/(\pi_c)$. Under the null hypothesis (e.g. Agresti, 1990), $G^2=X^2+o_p(1)\xrightarrow{d}\chi^2_{2^n-1}$. However, in sparse tables, when $N/2^n$ is small, the empirical distribution of these statistics is not well approximated by their limiting chi-square distribution (e.g., Koehler & Larntz, 1980).

The poor approximation of X^2 to its reference asymptotic distribution in sparse 2^n tables can be attributed to fact that the mean and variance of its reference asymptotic distribution are 2^n-1 and $2(2^n-1)$, respectively, but $\mathrm{E}\,(X^2)=2^n-1$ and $\mathrm{Var}\,(X^2)=2(2^n-1)+N^{-1}[2-2\cdot 2^n-2^{2n}+\sum_c\pi_c^{-1}]$ (Read & Cressie 1988: pp. 176–179). Thus, the discrepancy between the empirical variance of X^2 and its variance under its reference asymptotic distribution can be large when some probabilities π_c are small, and for sparse tables, the type I error X^2 will be larger than the α level based on its asymptotic critical value.

On the other hand, we show in the Appendix that X^2 can be written as

$$X^2=N\big(\dot{\mathbf{p}}-\dot{\boldsymbol{\pi}}\big)'\boldsymbol{\Xi}^{-1}\big(\dot{\mathbf{p}}-\dot{\boldsymbol{\pi}}\big).$$

That is, X^2 can be written as a weighted discrepancy between the sample and expected joint moments of the MVB distribution. But large samples are needed to accurately estimate high order joint sample moments. As an alternative to X^2 in sparse tables we propose testing whether the

sample joint moments match the population moments up to order r, where r depends on the size n of the model relative to sample size N, using the family of limited information test statistics

$$L_r = N(\mathbf{p}_r - \boldsymbol{\pi}_r)' \boldsymbol{\Xi}_r^{-1} (\mathbf{p}_r - \boldsymbol{\pi}_r), \quad r = 1, \ldots, n. \tag{3.1}$$

For $r = n$, $L_n = X^2$ (see proof in Appendix). L_r converges in distribution to a $\chi^2_{s(r)}$ distribution as $N \to \infty$. We also show in the Appendix that L_r is invariant to the relabeling of the categories indexed by 0 and 1.

Only probabilities up to $\min\{2r, n\}$ enter in the computation of L_r and the $O(N^{-1})$ term of $\mathrm{Var}\,(L_r)$ is most influenced by the smallest marginal probability of dimension $\min\{2r, n\}$. Hence we would expect L_r for small r to have a distribution closer to chi-square for small N even when there are some small probabilities π_c.

If the L_r test suggests significant misfit marginal residuals can be inspected to identify the source of the misfit. Again, letting $\dot{p}_a$ be an arbitrary marginal proportion, the standardized residual is $\sqrt{N}\,(\dot{p}_a - \dot{\boldsymbol{\pi}}_a)/\sqrt{\xi_{aa}}$, where is the ξ_{aa} is the ath diagonal element of $\boldsymbol{\Xi}$. The asymptotic distribution of this residual is standard normal.

To illustrate the small sample behavior of L_r, $r = 1, 2, 3$, against X^2, Table 1 has summaries of simulated type I errors using the asymptotic $\alpha = 0.05$ level critical values. For null MVB distributions, we used examples from the exchangeable beta-binomial MVB model with Bernoulli parameter η and dependence parameter γ (see Joe 1997, Section 7.1; and (3.3) below). Table 1 has two different null MVB distributions, the one based on $(\eta, \gamma) = (0.8, 0.5)$ has much smaller π_c values than that based on $(\eta, \gamma) = (0.5, 0.5)$. Table 1 clearly demonstrates the theory referred to above. Note that the asymptotic critical values for L_1, L_2 are quite good even for small $N/2^n$ ratios.

Bartholomew & Leung (2002) proposed a statistic for testing both simple and composite hypotheses that is closely related to L_r. Their statistic can be written as

$$N\big(\dot{\mathbf{p}}_2 - \dot{\boldsymbol{\pi}}_2\big)' \Big(\mathrm{diag}\,\big(\dot{\boldsymbol{\Xi}}_2\big)\Big)^{-1} \big(\dot{\mathbf{p}}_2 - \dot{\boldsymbol{\pi}}_2\big),$$

where $\dot{\boldsymbol{\Xi}}_2$ denotes the asymptotic covariance matrix of $\sqrt{N}\,(\dot{\mathbf{p}}_2 - \dot{\boldsymbol{\pi}}_2)$. This statistic is not asymptotically chi-square distributed even in the case of simple null hypotheses. Bartholomew & Leung (2002) used the first three moments of this statistic to approximate its sampling distribution using a chi-square distribution.

We now consider the power of L_r for different r. To do so, we derive the asymptotic distribution of L_r under a sequence of local alternatives for a parametric MVB model. Let $\boldsymbol{\pi}(\boldsymbol{\theta})$ be a parametric MVB model with parameters $\boldsymbol{\theta}$. Let $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and let the family of local alternatives be

$$H_{1N} : \boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{\epsilon}/\sqrt{N}\,. \tag{3.2}$$

Let $\boldsymbol{\delta} = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\boldsymbol{\epsilon}$. Under (3.2), from Bishop et al (1975: p. 471)

$$\sqrt{N}\,(\mathbf{p} - \boldsymbol{\pi}_0)] \xrightarrow{d} N(\boldsymbol{\delta}, \mathbf{D}_0 - \boldsymbol{\pi}_0\boldsymbol{\pi}_0') \quad \text{and} \quad \sqrt{N}\,(\mathbf{p}_r - \boldsymbol{\pi}_{0r})] \xrightarrow{d} N(\mathbf{T}_r\boldsymbol{\delta}, \boldsymbol{\Xi}_{r0}),$$

where $\boldsymbol{\Xi}_{r0} = \mathbf{T}_r(\mathbf{D}_0 - \boldsymbol{\pi}_0\boldsymbol{\pi}_0')\mathbf{T}_r'$. Therefore under (3.2), the limiting distributions of X^2 and L_r are noncentral χ^2 as $N \to \infty$. The noncentrality parameter for X^2 is $\boldsymbol{\delta}'\mathbf{D}_0^{-1}\boldsymbol{\delta}$, and the noncentrality parameter for L_r is $\lambda_r = (\mathbf{T}_r\boldsymbol{\delta})'\boldsymbol{\Xi}_{r0}^{-1}(\mathbf{T}_r\boldsymbol{\delta})$. Hence the power of L_r under the sequence of local alternatives at level α is the probability that a $\chi'^2_s(\lambda_r)$ random variable exceeds the upper 100αth percentile from the χ^2 distribution with $s = \sum_{i=1}^{r}\binom{n}{i}$ degrees of freedom.

To illustrate the power of the L_r statistics, we compute the asymptotic power of X^2 and L_r ($r = 1, 2, 3$) under the local alternatives for families of parametric MVB models. There are a number of parametric MVB models, for which $\boldsymbol{\theta}$ consists of univariate and bivariate parameters. A simple one is the multivariate binary beta-binomial model [see (7.4) of Joe (1997)], which is a two-parameter exchangeable MVB model. For this model, with η being the marginal Bernoulli parameter and γ being the dependence parameter (correlation is $\gamma/(1+\gamma)$), the joint distribution in dimension n is

$$\pi_{\mathbf{y}} = \pi_{\mathbf{y}}(\eta, \gamma) = \frac{\prod_{i=0}^{k-1}(\eta + i\gamma)\prod_{i=0}^{n-k-1}[1 - \eta + i\gamma]}{\prod_{i=0}^{n-1}(1 + i\gamma)}, \quad k = 0, \ldots, n;\ y_1 + \ldots + y_n = k. \tag{3.3}$$

A representative summary of the asymptotic power results is given in Table 2. For (3.3), $\boldsymbol{\theta} = (\eta, \gamma)'$, hence L_1 has no power when $\epsilon_1 = 0$ (or univariate margins for alternative same as the null), but for $\epsilon_1 \neq 0$, L_1 has more power than X^2. For $n = 3$, L_3 is the same as X^2 so that they have same power, and for $n > 3$, L_3 has more power than X^2. For $n > 2$, L_2 always has more power than X^2. When $\epsilon_1 \neq 0$ and $\gamma > 0$, L_1 is most powerful, and when $\epsilon_1 = 0$, L_2 is most powerful. These results may be a little surprising because one might have expected more asymptotic power when more information is employed (higher r), but note that all of the information in the beta-binomial MVB distribution can be summarized in the bivariate margins ($r = 2$).

For another comparison, we also considered a MVB distribution with higher order dependence parameters; one simple model for this is the Bahadur representation [see (7.21) of Joe (1997)] in the exchangeable case with up to third order terms. This model has one univariate, one bivariate and one trivariate parameter. In this case, L_2 and L_3 sometimes have more power than X^2 but not always. Also L_3 is sometimes more powerful than L_2 and definitely more powerful if the local alternative makes no change to the univariate and bivariate parameters.

The results of the power comparisons and small sample behavior show the usefulness of the class of L_r statistics for the case of a MVB parametric model and a simple null hypothesis. In small samples and sparse tables, the L_r statistics for small r are much more convenient than $L_n = X^2$ as the asymptotic chi-square approximation is valid for much smaller N.

4 Limited information tests of composite null hypotheses

In the preceding section we consider goodness-of-fit tests for MVB parametric models $\boldsymbol{\pi}(\boldsymbol{\theta})$ for a fixed a priori vector $\boldsymbol{\theta}$ of dimension q. In practice, in most applications for multivariate binary data, one is interested in comparing one or more MVB models where $\boldsymbol{\theta}$ is estimated from the data (i.e., composite null hypotheses). In this section, we study the analogs of the L_r statistics in (3.1) when parameters are estimated, via maximum likelihood or another estimation method. To do so, throughout this section we assume that that $\boldsymbol{\Delta} = \partial\boldsymbol{\pi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ is a $2^n \times q$ matrix with full column rank q, so that the model is identifiable. We also 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.

We shall first consider the case where the q-dimensional vector $\boldsymbol{\theta}$ is estimated using a consistent and asymptotically normal minimum variance estimator such as the maximum likelihood estimator or the minimum chi-square estimator.

4.1 Maximum likelihood and asymptotic minimum variance estimators

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

$$\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{B}\sqrt{N}\,(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1), \qquad \mathbf{B} = \boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1}, \tag{4.1}$$

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 $\sqrt{N}\,\hat{\mathbf{e}} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = (\mathbf{I} - \boldsymbol{\Delta}\mathbf{B})\boldsymbol{\Gamma}(\mathbf{I} - \boldsymbol{\Delta}\mathbf{B})' = \boldsymbol{\Gamma} - \boldsymbol{\Delta}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}'$.

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

$$\boldsymbol{\Sigma}_r = \mathbf{T}_r\boldsymbol{\Sigma}\mathbf{T}_r' = \boldsymbol{\Xi}_r - \boldsymbol{\Delta}_r\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}_r' \tag{4.2},$$

and

$$\boldsymbol{\Delta}_r = \frac{\partial\boldsymbol{\pi}_r(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} = \mathbf{T}_r\frac{\partial\boldsymbol{\pi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} = \mathbf{T}_r\boldsymbol{\Delta} \tag{4.3}$$

is a $s \times q$ matrix.

For an index a which is a subset of $\{1, \ldots, n\}$ of size less than or equal to r, the standardized marginal residual $\sqrt{N}\,\hat{e}_{r,a}/\sqrt{\boldsymbol{\Sigma}_{r,aa}(\hat{\boldsymbol{\theta}})}$ is asymptotically standard normal. The marginal residuals should be useful to assess the source of the misfit of a model.

We next consider testing composite null hypotheses of the model using limited information up to the r-dimensional joint moments. Let r_0 be the smallest integer r such that the model is (locally) identified from the joint moments up to order r. Then, for $r \geq r_0$, the matrix $\boldsymbol{\Delta}_r$ is of full column rank q. Note that this assumption ensures that $q < s$.

We could consider the statistic

$$N\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})\big)'\widehat{\boldsymbol{\Sigma}}_r^{+}\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})\big),$$

where $\widehat{\boldsymbol{\Sigma}}_r^{+}$ is the Moore-Penrose inverse of $\boldsymbol{\Sigma}_r(\hat{\boldsymbol{\theta}})$. This is asymptotically χ^2 with degrees of freedom equal to the rank of $\boldsymbol{\Sigma}_r$, which is between $s-q$ and s. With $r=2$, this is the statistic proposed by Reiser (1996). However, from studying $\boldsymbol{\Sigma}_r$ for some MVB models, we discovered that it sometimes has a small non-zero singular value, so that computation of $\widehat{\boldsymbol{\Sigma}}_r^{+}$ is not always stable. Hence, below we propose an alternative quadratic form statistic, with degree of freedom $s-q \leq \text{rank}\,(\boldsymbol{\Sigma}_r)$, based on a matrix that has $\boldsymbol{\Sigma}_r$ as a generalized inverse.

Consider a $s \times (s-q)$ orthogonal complement to $\boldsymbol{\Delta}_r$, say $\boldsymbol{\Delta}_r^{(c)}$, such that $\boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Delta}_r = \mathbf{0}$. Then, from (4.2), $\sqrt{N}\,\boldsymbol{\Delta}_r^{(c)\prime}\hat{\mathbf{e}}_r = \boldsymbol{\Delta}_r^{(c)\prime}\sqrt{N}\,\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})\big)$ has asymptotic covariance matrix

$$\boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Sigma}_r\boldsymbol{\Delta}_r^{(c)} = \boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r^{(c)}. \tag{4.4}$$

Thus,

$$\sqrt{N}\,\boldsymbol{\Delta}_r^{(c)\prime}\hat{\mathbf{e}}_r \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r^{(c)}\big). \tag{4.5}$$

Next, let

$$\mathbf{C}_r = \mathbf{C}_r(\boldsymbol{\theta}) = \boldsymbol{\Delta}_r^{(c)}\big(\boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r^{(c)}\big)^{-1}\boldsymbol{\Delta}_r^{(c)\prime}$$

and note that $\mathbf{C}_r$ is invariant to the choice of orthogonal complement (if $\boldsymbol{\Delta}_r^{(c)}$ is a full rank orthogonal complement, then so $\boldsymbol{\Delta}_r^{(c)}\mathbf{A}$ for a nonsingular matrix $\mathbf{A}$). It is straightforward to verify that $\mathbf{C}_r = \mathbf{C}_r\boldsymbol{\Sigma}_r\mathbf{C}_r$, that is, $\boldsymbol{\Sigma}_r$ is a generalized inverse of $\mathbf{C}_r$. Letting $\widehat{\mathbf{C}}_r = \mathbf{C}_r(\hat{\boldsymbol{\theta}})$, then we define

$$M_r = M_r(\hat{\boldsymbol{\theta}}) = N\hat{\mathbf{e}}_r'\widehat{\boldsymbol{\Delta}}_r^{(c)}\Big([\widehat{\boldsymbol{\Delta}}_r^{(c)}]'\widehat{\boldsymbol{\Xi}}_r\widehat{\boldsymbol{\Delta}}_r^{(c)}\Big)^{-1}[\widehat{\boldsymbol{\Delta}}_r^{(c)}]'\hat{\mathbf{e}}_r = N\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})\big)'\widehat{\mathbf{C}}_r\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})\big). \tag{4.6}$$

From (4.5) and Slutsky's theorem,

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

where the degrees of freedom are obtained from a result in Rao (1973: p. 30) using the fact that $\boldsymbol{\Delta}_r^{(c)}$ is of full column rank $s-q$ and hence $\mathbf{C}_r$ is also of rank $s-q$. Furthermore, using another result in Rao (1973: p. 77), $\mathbf{C}_r$ can be alternatively written as

$$\mathbf{C}_r = \mathbf{C}_r(\boldsymbol{\theta}) = \boldsymbol{\Xi}_r^{-1} - \boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r\big(\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r\big)^{-1}\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}. \tag{4.7}$$

Consider now the boundary case of this family of test statistics, M_n. From the results in the Appendix, M_n can be written as a quadratic form in the cell residuals as $M_n = N(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}))'\widehat{\mathbf{U}}(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}))$, and $M_n = X^2 - N(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}))'\widehat{\mathbf{V}}(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}))$, with $\widehat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}})$, where $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{D}^{-1}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1}$. But $(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}))'\widehat{\mathbf{D}}^{-1}\widehat{\boldsymbol{\Delta}}$ is the score vector or gradient in maximum

likelihood estimation, so that it is zero for the MLE, or $M_n = X^2$ when $\hat{\boldsymbol{\theta}}$ is the MLE. But for other minimum variance asymptotically normal estimators, $M_n \leq X^2$ and M_n, X^2 are equivalent only asymptotically.

Similar to L_r, M_r is invariant to the relabeling of the categories indexed by 0 and 1 provided that one stays inside the same parametric model (proof outlined in the Appendix).

To illustrate the finite sample performance of M_r, consider the following model with $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$, and multivariate binary probabilities

$$\pi_{\mathbf{y}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \Pr(Y_1 = y_1, \ldots, Y_n = y_n) = \int_{-\infty}^{\infty} \prod_{j=1}^{n} \frac{e^{(\alpha_j + \beta_j x) y_j}}{1 + e^{\alpha_j + \beta_j x}} \, \phi(x) \, dx, \tag{4.8}$$

where $\phi(x)$ is the standard normal density. This is the logit-normit model (Bartholomew & Knott 1999), and it is also known as two-parameter logistic model with a normally distributed latent trait (e.g., Lord & Novick 1968).

Table 3 has the mean, variance, and empirical rejection rates at $\alpha = 0.20, 0.10, 0.05, 0.01$ for M_2, M_3 and X^2 with maximum likelihood estimation of a logit-normit model for a 5-variable model and an 8-variable model with $N = 100$ and $N = 1000$. Numerical optimization used a quasi-Newton routine with analytic derivatives. Computations used 48-point Gauss-Hermite quadrature for the integrals (4.8) and their derivatives with respect to α_i, β_i; this is computationally faster, and matched computations of MLEs to four decimal places when Romberg integration was used with accuracy 10^{-6} integrals in (4.8) and their derivatives. The tabulated results are based on the simulations for which the iterations for maximum likelihood estimation converged; see comments in Bartholomew & Knott (1999) regarding non-convergence. As can be seen in this table, similar to L_r versus X^2, the M_r statistics have small sample distributions closer to the asymptotic one in sparse high-dimensional case, especially in the extreme upper tail; in particular, asymptotic critical values of X^2 are not reliable in this case.

4.2 Consistent and asymptotically normal estimators

In this subsection we consider limited information testing of composite hypotheses when the model parameters are estimated using some alternative consistent estimator $\tilde{\boldsymbol{\theta}}$. Other simpler estimation methods, such as the limited information estimation methods in Section 5, must be considered when the n-dimensional probabilities may be too difficult to compute.

We assume that $\tilde{\boldsymbol{\theta}}$ satisfies

$$\sqrt{N}\,(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{H}\sqrt{N}\,(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1), \tag{4.9}$$

for some $q \times 2^n$ matrix $\mathbf{H}$. Some special cases of limited information estimators $\tilde{\boldsymbol{\theta}}$ (based on low-dimensional margins) are given in Section 5.

We derive the asymptotic distribution of the vector of cell residuals $\tilde{\mathbf{e}} = \mathbf{p} - \boldsymbol{\pi}(\tilde{\boldsymbol{\theta}})$ for (4.9). 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 $\sqrt{N}\,\tilde{\mathbf{e}} = (\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1)$, and the asymptotic covariance matrix of $\sqrt{N}\,\tilde{\mathbf{e}}$ is $\widetilde{\boldsymbol{\Sigma}} = (\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})\boldsymbol{\Gamma}(\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})'$.

Next we consider moments up to order r only, where $r \geq r_0$. 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 (4.3)) $\sqrt{N}\,\tilde{\mathbf{e}}_r \xrightarrow{d} N(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_r)$, with

$$\widetilde{\boldsymbol{\Sigma}}_r = (\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{H})\boldsymbol{\Gamma}(\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{H})'. \tag{4.10}$$

To test composite null hypotheses with this class of estimators we may use the $M_r = M_r(\tilde{\boldsymbol{\theta}})$ statistic (4.6) with $\tilde{\boldsymbol{\theta}}$ in place of $\hat{\boldsymbol{\theta}}$. This is because if $\boldsymbol{\Delta}_r^{(c)}$ is a $s \times (s-q)$ orthogonal complement to $\boldsymbol{\Delta}_r$, then $\sqrt{N}\,\boldsymbol{\Delta}_r^{(c)\prime}\tilde{\mathbf{e}}_r = \boldsymbol{\Delta}_r^{(c)\prime}\sqrt{N}\left(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})\right)$ has asymptotic covariance matrix

$$\boldsymbol{\Delta}_r^{(c)\prime}\widetilde{\boldsymbol{\Sigma}}_r\boldsymbol{\Delta}_r^{(c)} = \boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r^{(c)},$$

the same as the right-hand side of (4.4).

Thus, we have shown that M_r is asymptotically χ^2_{s-q} if $\tilde{\boldsymbol{\theta}}$ is any consistent estimator of $\boldsymbol{\theta}$. In particular we have shown that the full information test statistic $M_n = M_n(\tilde{\boldsymbol{\theta}})$ is asymptotically $\chi^2_{2^n-1-q}$ for this large class of consistent estimators. Previously, there had not been any goodness-of-fit statistic that is asymptotically chi-square for any consistent estimator of $\boldsymbol{\theta}$. Note that with $X^2(\tilde{\boldsymbol{\theta}})$ representing the X^2 statistic based on $\tilde{\boldsymbol{\theta}}$, the results in the Appendix, with $\tilde{\boldsymbol{\theta}}$ replacing $\hat{\boldsymbol{\theta}}$, imply that $M_n(\tilde{\boldsymbol{\theta}}) \leq X^2(\tilde{\boldsymbol{\theta}})$; that is, for a consistent estimator that is not the MLE, the asymptotic distribution of $X^2(\tilde{\boldsymbol{\theta}})$ is stochastically larger than $\chi^2_{2^n-1-q}$.

4.3 Asymptotic distribution under local alternatives and power comparison of X^2 and M_r

Similar to Section 3.2, we can compare the asymptotic power of X^2 and M_r under a sequence of local alternatives. There are several ways to specify the null and alternative hypotheses, and we will take the special case where the null hypothesis is a nested model with parameters to be estimated, since if fitting models to categorical data one often checks if a simpler (nested) version of a model explains the data adequately.

We let $\boldsymbol{\pi}(\boldsymbol{\theta})$ denote a MVB model. For the submodel or nested model, we suppose the parametrization is of the form $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ where $\boldsymbol{\theta}_2 = \beta\mathbf{1}$.

For testing, the hypotheses are

$$H_0 : (\boldsymbol{\theta}_1', \beta\mathbf{1}')' \quad \text{vs} \quad H_1 : (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'. \tag{4.11}$$

For a sequence of local alternatives, we take $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}'_{10}, \beta_0 \mathbf{1}')'$ as a 'true' model, and let $\boldsymbol{\theta}_{1N} = (\boldsymbol{\theta}'_{10}, \beta_0 \mathbf{1}' + w_N \boldsymbol{\gamma}')'$ be the sequence of alternative parameter values, with $\sqrt{N}\, w_N \to \epsilon$. $\boldsymbol{\gamma}$ is a nonconstant vector that sums to 0 (for identifiability). Let $\boldsymbol{\theta}_0^* = (\boldsymbol{\theta}'_{10}, \beta_0)'$, and $\boldsymbol{\theta}^* = (\boldsymbol{\theta}'_1, \beta)'$ and let $\hat{\boldsymbol{\theta}}_N$ (same dimension as $\boldsymbol{\theta}_0^*$) be the MLE (or an asymptotic minimum variance estimator) based on the null model, assuming a random sample of size N from $\boldsymbol{\pi}(\boldsymbol{\theta}_{1N})$. Under the above sequence of local alternatives, $\hat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0^*$ and $\Sigma_r(\hat{\boldsymbol{\theta}}_N) \xrightarrow{p} \Sigma_r(\boldsymbol{\theta}_0^*)$. For the vector of residuals,

$$\sqrt{N}\,(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}_N)) = \sqrt{N}\,\Big\{[\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta}_N)] + [\boldsymbol{\pi}_r(\boldsymbol{\theta}_N) - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}_N)]\Big\}.$$

Taking expected values the first term is zero in expectation, and expanding the second term leads to:

$$\begin{aligned}
\sqrt{N}\,\mathrm{E}\,[\boldsymbol{\pi}_r(\boldsymbol{\theta}_N) - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}_N)] &= \sqrt{N}\,[\boldsymbol{\pi}_r(\boldsymbol{\theta}_N) - \boldsymbol{\pi}_r(\boldsymbol{\theta}_0)] - \sqrt{N}\,\mathrm{E}\,[\boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}_N) - \boldsymbol{\pi}_r(\boldsymbol{\theta}_0)] \\
&= \sqrt{N}\,\Big[\frac{\partial \boldsymbol{\pi}_r(\boldsymbol{\theta}_0)}{\partial(\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)}(\boldsymbol{\theta}_N - \boldsymbol{\theta}_0) - \frac{\partial \boldsymbol{\pi}_r}{\partial(\boldsymbol{\theta}'_1, \beta)}\mathrm{E}\,(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*) + o_p(||\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*||)\Big] \\
&\approx \sqrt{N}\,\Big[\frac{\partial \boldsymbol{\pi}_r(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'_2} w_N \boldsymbol{\gamma}\Big] - \epsilon \frac{\partial \boldsymbol{\pi}_r}{\partial(\boldsymbol{\theta}'_1, \beta)}\boldsymbol{\zeta} + o_p(1) \\
&\to \epsilon\Big[\frac{\partial \boldsymbol{\pi}_r(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'_2}\boldsymbol{\gamma} - \frac{\partial \boldsymbol{\pi}_r}{\partial(\boldsymbol{\theta}'_1, \beta)}\boldsymbol{\zeta}\Big] \stackrel{\text{def}}{=} \boldsymbol{\delta}_r, \qquad (4.12)
\end{aligned}$$

where from the Appendix,

$$\epsilon\boldsymbol{\zeta} = \lim \sqrt{N}\,\mathrm{E}\,(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*) = \epsilon[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0^*)]^{-1} \sum_{\mathbf{y}} \frac{\partial \log \pi_{\mathbf{y}}(\boldsymbol{\theta}_{10}, \beta_0 \mathbf{1})}{\partial(\boldsymbol{\theta}'_1, \beta)'} \cdot \boldsymbol{\gamma}' \frac{\partial \pi_{\mathbf{y}}(\boldsymbol{\theta}_{10}, \beta_0 \mathbf{1})}{\partial \boldsymbol{\theta}_2}, \qquad (4.13)$$

and $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0^*)$ is the Fisher information matrix for the model $\boldsymbol{\pi}(\boldsymbol{\theta})$ under the null hypothesis. Note that $\boldsymbol{\delta}_r = \mathbf{T}_r \boldsymbol{\delta}$ where $\boldsymbol{\delta}$ is computed like $\boldsymbol{\delta}_r$ with $\boldsymbol{\pi}$ replacing $\boldsymbol{\pi}_r$ in (4.12).

Under the sequence of local alternatives,

$$\sqrt{N}\,(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}_N)) \xrightarrow{d} N(\boldsymbol{\delta}_r, \boldsymbol{\Sigma}_r).$$

For the comparison with the usual chi-square statistic,

$$\sqrt{N}\,(\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_N)) \xrightarrow{d} N(\boldsymbol{\delta}, \boldsymbol{\Sigma})$$

using an argument analogous to above.

Using standard results for non-central distributions (e.g., Rao 1973), noncentrality parameters for X^2 and M_r ($r \geq r_0$) are $\boldsymbol{\delta}' \mathbf{D}_0^{-1} \boldsymbol{\delta}$ [$\mathbf{D}_0 = \mathrm{diag}\,(\boldsymbol{\pi}(\theta_{10}, \beta_0 \mathbf{1}))$)] and $\boldsymbol{\delta}'_r \mathbf{C}_r \boldsymbol{\delta}_r$ respectively, and the degrees of freedom are $2^n - 1 - q$ and $s - q$ respectively. The power calculations are then like in Section 3.2. Also, the power under local alternatives can be computed in a similar way for other consistent estimators. If the estimator is written as a solution to a set of estimating equations

$\sum_{i=1}^{N} \boldsymbol{\psi}(\boldsymbol{\theta}, \mathbf{y}_i)$ (Godambe, 1991), then in (A4), the inverse information matrix is replaced by $-\mathbf{D}_{\psi}(\boldsymbol{\theta})$ where $\mathbf{D}_{\psi} = \mathrm{E}\,[\partial\boldsymbol{\psi}/\partial\boldsymbol{\theta}']$, and $\partial\ell/\partial\boldsymbol{\theta}$ is replaced by $\boldsymbol{\psi}$.

To illustrate our discussion, for the logit-normit model (4.8) with $H_0 : \boldsymbol{\beta} = \beta\mathbf{1}$, the power for X^2 and M_r $(r = 2, 3)$ were computed under sequences of local alternatives. The model under the null hypothesis is referred to in the educational testing literature as one-parameter logistic (or Rasch) model with a normally distributed latent trait (e.g., Thissen 1982). Some representative results are given in Table 4. These show that both M_2 and M_3 are more powerful than X^2, with M_2 being the most powerful of the three. Note that model (4.8) is determined from the univariate and bivariate moments for $n \geq 3$. As a check on the asymptotic power results, simulations were performed to compared the power for finite N. The relative comparisons were analogous to those in Table 4; the rate of convergence to the asymptotic power as N increases, depends on the null parameter vector and direction of local alternative.

In summary, for this commonly used model for multivariate binary data we have shown that the newly proposed M_r statistics have more power than the X^2 statistic.

5 Limited information estimation

In this section, we consider consistent estimators that are limited information estimators, that is, they are based on low-dimensional margins. A simple class of such estimators are based on weighted least squares (WLS) of the moment residuals up to order r. The results of Section 4.2 apply to these estimators.

Consider the estimator $\tilde{\boldsymbol{\theta}}$ that is the minimum of

$$F_r = F_r(\boldsymbol{\theta}) = (\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta}))'\widehat{\mathbf{W}}(\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta})), \tag{5.1}$$

where $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W} = \mathbf{W}(\boldsymbol{\theta})$ positive definite matrix. Obvious choices for $\widehat{\mathbf{W}}$ in (5.1) are $\widehat{\mathbf{W}} = \mathbf{I}$, $\widehat{\mathbf{W}} = (\,\mathrm{diag}\,(\widehat{\boldsymbol{\Xi}}_r))^{-1}$, and $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_r^{-1}$, where $\widehat{\boldsymbol{\Xi}}_r$ indicates that $\boldsymbol{\Xi}_r$ is consistently evaluated using sample proportions. Alternatively, we could also minimize

$$F_r(\boldsymbol{\theta}) = (\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta}))'\mathbf{W}(\boldsymbol{\theta})(\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta})). \tag{5.2}$$

If $r \geq r_0$ and $\boldsymbol{\Delta}_r$ is of full rank q, and some other mild regularity conditions are satisfied (e.g., Browne, 1984; Satorra, 1989; Ferguson, 1996), then $\tilde{\boldsymbol{\theta}}$ is consistent and

$$\sqrt{N}\,(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{K}\sqrt{N}\,(\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta})) + o_p(1) = \mathbf{K}\mathbf{T}_r\sqrt{N}\,(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1), \tag{5.3}$$

where $\mathbf{K} = (\boldsymbol{\Delta}_r'\mathbf{W}\boldsymbol{\Delta}_r)^{-1}\boldsymbol{\Delta}_r'\mathbf{W}$. Note that (5.3) has the form of (4.9). Furthermore, we have

$$\sqrt{N}\,(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{K}\boldsymbol{\Xi}_r\mathbf{K}') \tag{5.4}$$

and

$$\sqrt{N}\,(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})) \xrightarrow{d} N(\mathbf{0}, (\mathbf{I} - \boldsymbol{\Delta}_r\mathbf{K})\boldsymbol{\Xi}_r(\mathbf{I} - \boldsymbol{\Delta}_r\mathbf{K})'), \tag{5.5}$$

since from (4.10), $\widetilde{\boldsymbol{\Sigma}}_r = (\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{K}\mathbf{T}_r)\boldsymbol{\Gamma}(\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{K}\mathbf{T}_r)' = (\mathbf{I} - \boldsymbol{\Delta}_r\mathbf{K})\boldsymbol{\Xi}_r(\mathbf{I} - \boldsymbol{\Delta}_r\mathbf{K})'$.

For the special case, where $\mathbf{W}(\boldsymbol{\theta}) = \boldsymbol{\Xi}_r(\boldsymbol{\theta})$, with $\widehat{\mathbf{W}}$ in (5.1) corresponding to $\widehat{\boldsymbol{\Xi}}_r^{-1}$, there are some simplifications of the results. Equations (5.4) and (5.5) simplify to

$$\sqrt{N}\,(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, (\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r)^{-1}), \quad \sqrt{N}\,(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})) \xrightarrow{d} N\big(\mathbf{0},\, \widetilde{\boldsymbol{\Sigma}}_r = \boldsymbol{\Xi}_r - \boldsymbol{\Delta}_r(\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r)^{-1}\boldsymbol{\Delta}_r'\big), \tag{5.6}$$

and we obtain the optimal estimator within the class of the form of weighted least squares in the residuals of moments up to order r. In this case, we can also define a simpler form Q_r in place of $M_r(\tilde{\boldsymbol{\theta}})$ in (4.6) that looks more like L_r in (3.1):

$$Q_r = N\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})\big)'\widehat{\boldsymbol{\Xi}}_r^{-1}\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})\big). \tag{5.7}$$

From the theory of quadratic forms on normal random variables (Rao, 1973: Section 3b.4) and Slutsky's theorem, Q_r is asymptotically χ^2 since $\boldsymbol{\Xi}_r^{-1}\widetilde{\boldsymbol{\Sigma}}_r = \mathbf{I} - \boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r\big(\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r\big)^{-1}\boldsymbol{\Delta}_r'$ (with $\widetilde{\boldsymbol{\Sigma}}_r$ in (5.6)) is idempotent.

Another way to show this asymptotic result, with the degrees of freedom, is as follows. (5.7) can be considered as a special case of

$$M_r' = M_r'(\tilde{\boldsymbol{\theta}}) = N\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})\big)'\widehat{\mathbf{C}}_r\big(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})\big), \tag{5.8}$$

where $\widehat{\mathbf{C}}_r$ is $\mathbf{C}_r(\boldsymbol{\theta})$ given by (4.7) evaluating all the derivative matrices using consistent parameter estimates and consistently estimating the marginal probabilities in $\boldsymbol{\Xi}_r$ using sample proportions. By Slutsky's theorem and the results of Section 4, M_r' is asymptotically χ^2_{s-q}. The estimator obtained by minimizing (5.1) satisfies $(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}))'\widehat{\mathbf{W}}\boldsymbol{\Delta}_r = \mathbf{0}'$ from the gradient of (5.1), and for $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_r^{-1}$, (5.8) becomes (5.7) as the second term (after substitution for $\widehat{\mathbf{C}}_r$) becomes zero. Hence $NF_r(\tilde{\boldsymbol{\theta}}) = Q_r = M_r'$ when $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_r^{-1}$.

As special cases of the theory laid out in this section we find that minimizing F_n with $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_n^{-1}$ is equivalent to minimizing the minimum modified chi-square function $N\sum_{c=1}^{2^n}(p_c - \pi_c)^2/p_c$. Also, Christoffersson (1975) minimized F_2 with $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_2^{-1}$ to estimate the normit-normit (aka multidimensional normal ogive) latent trait model (see Bartholomew & Knott, 1999). In general, limited information methods such as Christoffersson's are computationally attractive to estimate models, such as the multidimensional normal ogive model, for which computing cell probabilities is difficult.

However, for large n such as $n > 25$, Christoffersson's estimator becomes unattractive since a large weight matrix needs to be inverted. Furthermore, large samples may be needed to estimate the

fourth order probabilities involved in $\boldsymbol{\Xi}_2$ using sample proportions. Alternatively, one could minimize F_2 in (5.1) with $\widehat{\mathbf{W}} = (\operatorname{diag}(\widehat{\boldsymbol{\Xi}}_2))^{-1}$ or $\widehat{\mathbf{W}} = \mathbf{I}$, or (5.2) with $\mathbf{W}(\boldsymbol{\theta}) = (\operatorname{diag}(\boldsymbol{\Xi}_2(\boldsymbol{\theta}))^{-1}$. These estimators are extremely attractive from a computational viewpoint but they are not asymptotically efficient even within the class of estimators relying only on univariate and bivariate information.

It is interesting to compare the asymptotic efficiency of alternative members of this class of estimators. In Table 5 we provide some results for model (4.8) comparing the asymptotic relative efficiency (ARE) of estimators relative to the MLE, for the weighted residual moments least squares F_r with $\widehat{\mathbf{W}} = \mathbf{I}$ $(r = 2, 3)$, F_r with $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_2^{-1}$ $(r = 2, 3)$, and F_n with $\widehat{\mathbf{W}} = \mathbf{I}$. The AREs in Table 5 are based on the average of 100 sets of parameters for (4.8), with the α_i's random with Uniform$(-2, 2)$ distribution, and the β_i's random with Uniform$(1, 2)$ distribution. Relative efficiencies were calculated based on diagonal entries and determinants of asymptotic covariance matrices. The matrices involved in the calculations in Table 5 are:

(a) the asymptotic covariance matrix of the MLE is $\boldsymbol{\mathcal{I}}^{-1}$ from (4.1),

(b) with $\widehat{\mathbf{W}} = \mathbf{I}$ for unweighted least squares (ULS), the asymptotic covariance matrix of $\tilde{\boldsymbol{\theta}}$ is $(\boldsymbol{\Delta}_r'\boldsymbol{\Delta}_r)^{-1}\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r(\boldsymbol{\Delta}_r'\boldsymbol{\Delta}_r)^{-1}$,

(c) with $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_r^{-1}$, the asymptotic covariance matrix of $\tilde{\boldsymbol{\theta}}$ is $(\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r)^{-1}$.

Note that the estimators in (b) are highly efficient, and the WLS estimators in (c) with $r = 2, 3$ are extremely highly efficient with efficiency in the 0.99–1.00$^-$ range. For (c), the efficiency is shown as 0.99 in Table 5 in the one case of $n = 5$, $r = 2$ only; the $r = 3$ and $n = 8$ cases are not displayed as the efficiency summary is the same as for $n = 5$, $r = 2$. Note that ULS with $r = n$ has worse efficiency than ULS with $r = 2, 3$. The $r = n$ case is probably worse because it will weight the small n-dimensional probabilities the same as the larger ones. For $r = 2, 3$, the marginal probabilities tend not be vary as much. We also did finite sample (N in the range of hundreds to thousands) comparisons of the estimators in (b) and (c), and the comparisons are similar to the asymptotic efficiencies. The MLE is only marginally better in terms of mean squared error.

6 Numerical example

A common task in the Social Sciences is to measure unobservable constructs such as cognitive abilities, personality traits, or social attitudes by administering a set items written to be indicators of the unobservable constructs (see Bartholomew, 1988). We now provide an example where a sample of individuals were asked to respond to a set of items using two categories. Their responses were collected in a 2^n contingency table. These contingency tables are then modeled using a latent trait model, with the latent trait being the unobservable construct being measured.

This example for the Social Life Feelings scale is taken from Bartholomew & Knott (1999, pp. 97–98), who used data from an original study by Schuessler (1982). The data consist of the responses of 1490 German respondents to $n = 5$ binary questions intended to measure economic self-determination. Bartholomew & Knott (1999) fitted a logit-normit model (4.8) to these data using maximum likelihood.

To illustrate the use of limited information estimation, in Table 6 we provide our maximum likelihood and bivariate ULS ($r = 2$) estimates. Our MLE parameter estimates and standard errors agree with those reported by Bartholomew & Knott (1999). In terms of model fit, we obtained the results provided in Table 7. The M_r statistics based on MLEs and bivariate ULS are similar, and lead to the same conclusions. Note that $X^2 = M_5$ with $r = n = 5$ for maximum likelihood estimation only, from results in Section 4. Unlike Bartholomew & Knott (1999) we have not pooled cells in computing X^2. Nevertheless, our P-value agrees with the reported by these authors. Furthermore, there is agreement between the results obtained using limited information and full information tests.

Clearly, the model does not fit well in this situation and we proceed to identify the source of the misfit using the maximum likelihood estimates. From the standardized cell residuals, the binary patterns that show significant misfit are (10011), (00111), (10110), (11110), and (11111). These residuals suggest that the model does not fit well for item 4. However, the standardized marginal residuals up to third order (see Section 4.1) present a very different picture. Significant marginal residuals are obtained for (1,5), (3,5), (1,2,4), (1,2,5), (1,3,5), and (1,4,5). They clearly suggest that the model does not fit well for item 5. To verify both conjectures, we fitted a logit-normit model to these data to all 5 combinations of 4 items (with 7 degrees of freedom). The results are presented in the second part of Table 7. They clearly indicate that economic self determination is best measured by the first four items of these scale, as suggested by the marginal residuals.

7 Discussion and conclusions

The most serious challenge faced by a researcher confronted with modeling 2^n contingency tables for large n is how to test the goodness-of-fit of the model, as the empirical distribution of the usual goodness-of-fit statistics is not well approximated by its asymptotic distribution in large and sparse tables. In the past, two general solutions have been proposed to overcome this problem: resampling methods and pooling cells. Resampling methods may be too time consuming when fitting models that are computationally intensive, whereas pooling cells in large and sparse tables may not make best use of the multivariate structure and may yield statistics with unknown sampling distribution. Here we have proposed an alternative approach: testing whether the model reproduces the low order moments of the MVB distribution. This amounts to pooling cells in a systematic way, so that the

resulting statistics have a known (asymptotic) distribution.

To this aim, we have proposed two families of test statistics, L_r and M_r where r denotes the highest order at which testing is performed. L_r is a family of test statistics suitable for testing parametric hypotheses with a priori determined parameter values, whereas M_r is a family of test statistics suitable for testing parametric hypotheses where the parameters are to be estimated from the data.

In large and sparse 2^n tables L_r for small r ($r = 1, 2, 3$) should be employed instead of X^2 as the former have more precise empirical Type I errors and may be asymptotically more powerful than the latter. Similarly, with estimated model parameters, M_r for small r should be used to test composite parametric hypotheses instead of X^2, as the former have more precise empirical Type I errors and may be asymptotically more powerful than the latter.

If the model is identified from the margins up to order r and if it is estimated using a consistent and asymptotically normal estimator, M_r is asymptotically $\chi^2_{s(r)-q}$, with degrees of freedom equal to the total number of multivariate moments used for testing minus the number of parameters being estimated. This is a remarkable result as we are not aware of any goodness-of-fit statistic for 2^n tables whose asymptotic distribution has been described under such general conditions. A special case of M_r is M_n. This is a full information statistic that can be used to assess the goodness-of-fit to the table cells under the same conditions stated above. For minimum variance consistent and asymptotically normal estimators, M_n is asymptotically equal to X^2. In particular, in the case of maximum likelihood estimation, $M_n = X^2$.

After assessing the overall goodness-of-fit of a model, if this is poor, it is necessary to determine the source of the misfit. We propose using marginal residuals which are asymptotically standard normal. As our numerical example illustrate, the use of these residuals can be much more informative than the use of cell residuals.

With high-dimensional sparse contingency tables for which maximum likelihood estimation may not be computationally feasible, limited information estimators are often used in Psychometrics to estimate normit-normit and related latent trait models, generally using a multi-stage approach that makes use of the information contained in the univariate and bivariate margins of the table (see Christoffersson, 1975; Jöreskog, 1994; Lee, Poon & Bentler, 1995; Maydeu-Olivares, 2001, 2002; Muthén, 1978, 1984, 1993). Such popular software packages as LISREL (Jöreskog & Sörbom, 2001), EQS (Bentler, 1995) and MPLUS (Muthén & Muthén, 2001) can be used to estimate these models using these sequential limited information estimators. Here we have provided a full information test statistic, M_n, which can be used to assess the goodness of fit of models estimated using these sequential procedures. Also, we have considered a class of one-stage estimators obtained by minimizing

F_r in (5.1), which includes both limited and full information estimators. This class of estimators is related to the class of goodness-of-fit test statistics M_r.

As n gets larger, there are computational details that have to be considered to manage the computations within available computer memory. In future research, we will provide other related approaches that are computationally simpler. Also, we have not covered here sparse multidimensional tables in which the categorical variables take more than two values. Our results extend readily to this case, which we will discuss in a separate report.

Appendix

$L_n = X^2$, and $M_n(\hat{\boldsymbol{\theta}}) \leq X^2(\hat{\boldsymbol{\theta}})$ with equality for MLE

We claim that $X^2 = N(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}})'\boldsymbol{\Xi}^{-1}(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}})$, which is the definition of L_n since $\boldsymbol{\pi}_n = \dot{\boldsymbol{\pi}}$ and $\boldsymbol{\Xi}_n = \boldsymbol{\Xi}$. To see this, let $\dot{\mathbf{e}} = \dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}$, $\check{\mathbf{e}} = \check{\mathbf{p}} - \check{\boldsymbol{\pi}}$ and $\mathbf{e} = \mathbf{p} - \boldsymbol{\pi}$. Since $\dot{\mathbf{e}} = \dot{\mathbf{T}}\check{\mathbf{e}}$,

$$N(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}})'\boldsymbol{\Xi}^{-1}(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}) = N\check{\mathbf{e}}'\dot{\mathbf{T}}'\boldsymbol{\Xi}^{-1}\dot{\mathbf{T}}\check{\mathbf{e}}'. \tag{A1}$$

Letting $\check{\mathbf{D}} = \operatorname{diag}(\check{\boldsymbol{\pi}})$, $\boldsymbol{\Xi} = \dot{\mathbf{T}}(\check{\mathbf{D}} - \check{\boldsymbol{\pi}}\check{\boldsymbol{\pi}}')\dot{\mathbf{T}}'$, and

$$\boldsymbol{\Xi}^{-1} = (\dot{\mathbf{T}})'^{-1}(\check{\mathbf{D}}^{-1} + \mathbf{1}D_0^{-1}\mathbf{1}')\dot{\mathbf{T}}^{-1}, \tag{A2}$$

where $D_0 = \pi_{0\cdots0}$. Thus, (A1) is the same as $N(\check{\mathbf{e}}'\check{\mathbf{D}}^{-1}\check{\mathbf{e}} + \check{\mathbf{e}}'\mathbf{1}D_0^{-1}\mathbf{1}'\check{\mathbf{e}})$. Since $\mathbf{e}$ can be partitioned as $\mathbf{e}' = (e_0, \check{\mathbf{e}}')'$ where $e_0 = -\mathbf{1}'\check{\mathbf{e}}$, then (A1) becomes

$$N(\check{\mathbf{e}}'\check{\mathbf{D}}\check{\mathbf{e}} + D_0^{-1}e_0^2) = N\mathbf{e}'\mathbf{D}^{-1}\mathbf{e} = X^2.$$

For M_n, let $\dot{\mathbf{e}} = \dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}}) = \mathbf{p}_n - \boldsymbol{\pi}_n(\hat{\boldsymbol{\theta}})$, $\check{\mathbf{e}} = \check{\mathbf{p}} - \check{\boldsymbol{\pi}}(\hat{\boldsymbol{\theta}})$, and $\hat{\mathbf{e}} = \mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ for an estimator $\hat{\boldsymbol{\theta}}$, so that

$$M_n = N\dot{\mathbf{e}}'\widehat{\mathbf{C}}_n\dot{\mathbf{e}}, \quad \widehat{\mathbf{C}}_n = \mathbf{C}_n(\hat{\boldsymbol{\theta}}), \quad \mathbf{C}_n = \boldsymbol{\Xi}^{-1} - \boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_n\big(\boldsymbol{\Delta}_n'\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_n\big)^{-1}\boldsymbol{\Delta}_n'\boldsymbol{\Xi}^{-1}.$$

We claim that

$$M_n = N\hat{\mathbf{e}}'\widehat{\mathbf{U}}(\hat{\mathbf{e}}), \quad \widehat{\mathbf{U}} = \mathbf{U}(\hat{\boldsymbol{\theta}}), \quad \text{where } \mathbf{U} = \mathbf{U}(\boldsymbol{\theta}) = \mathbf{D}^{-1} - \mathbf{D}^{-1}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1},$$

so that

$$M_n = X^2(\hat{\boldsymbol{\theta}}) - N\hat{\mathbf{e}}'\widehat{\mathbf{V}}\hat{\mathbf{e}}, \quad \widehat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}}), \quad \text{where } \mathbf{V}(\boldsymbol{\theta}) = \mathbf{D}^{-1}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1}.$$

Let hats on matrices denoting evaluation at $\hat{\boldsymbol{\theta}}$. For the proof of the claim, from the above algebraic result for X^2 and L_n, $\dot{\mathbf{e}}'\widehat{\boldsymbol{\Xi}}^{-1}\dot{\mathbf{e}} = \hat{\mathbf{e}}'\widehat{\mathbf{D}}^{-1}\hat{\mathbf{e}}$. With partitioning of $\boldsymbol{\Delta}' = \begin{pmatrix}\boldsymbol{\Delta}_0' & \check{\boldsymbol{\Delta}}'\end{pmatrix}$, we have $\boldsymbol{\Delta}_n = \dot{\mathbf{T}}\check{\boldsymbol{\Delta}}$. Thus, from (A2), $\boldsymbol{\Delta}_n'\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_n$ in the definition of $\mathbf{C}_n$ for M_n equals $\check{\boldsymbol{\Delta}}'(\check{\mathbf{D}}^{-1} + \mathbf{1}D_0^{-1}\mathbf{1}')\check{\boldsymbol{\Delta}}$, evaluated at $\hat{\boldsymbol{\theta}}$. But since $\mathbf{1}'\boldsymbol{\Delta} = \mathbf{0}'$, $\mathbf{1}'\check{\boldsymbol{\Delta}} = -\boldsymbol{\Delta}_0$, and $\boldsymbol{\Delta}_n'\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_n = \boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta}$ at $\hat{\boldsymbol{\theta}}$. Similarly, since $\dot{\mathbf{e}} = \dot{\mathbf{T}}\check{\mathbf{e}}$,

$$\dot{\mathbf{e}}'\widehat{\boldsymbol{\Xi}}^{-1}\widehat{\boldsymbol{\Delta}}_n = \check{\mathbf{e}}'\dot{\mathbf{T}}'\dot{\mathbf{T}}'^{-1}\Big(\widehat{\check{\mathbf{D}}}^{-1} + \mathbf{1}\hat{D}_0^{-1}\mathbf{1}'\Big)\dot{\mathbf{T}}^{-1}\dot{\mathbf{T}}\widehat{\check{\boldsymbol{\Delta}}} = \check{\mathbf{e}}'\widehat{\check{\mathbf{D}}}^{-1}\widehat{\check{\boldsymbol{\Delta}}} + e_0\hat{D}_0^{-1}\widehat{\boldsymbol{\Delta}}_0' = \hat{\mathbf{e}}'\widehat{\mathbf{D}}^{-1}\widehat{\boldsymbol{\Delta}}. \tag{A3}$$

where $e_0 = -\mathbf{1}'\check{\mathbf{e}}$. Hence the claim is established.

Finally, (A3) is $\mathbf{0}'$ if $\hat{\boldsymbol{\theta}}$ is the MLE since it is the vector of score equations that the MLE satisfies. So $M_n = X^2$ for the MLE.

Invariance to 0-1 labeling

For any statistical procedure with binary data, it is important to check on the effect of the labeling of categories. We first prove the invariance for L_r. If the 0-1 labeling is reversed, then $\boldsymbol{\pi}$ (in the ordering described in Section 2) is completely reversed, that is, the probability vector becomes $\boldsymbol{\Lambda\pi}$, where $\boldsymbol{\Lambda}$ is a $2^n \times 2^n$ matrix which has 1s in the $(i, 2^n - i)$ positions for all i and 0s elsewhere. Let $\mathbf{e} = \mathbf{p} - \boldsymbol{\pi}$ and $\mathbf{e}_r = \mathbf{p}_r - \boldsymbol{\pi}_r$. Under the relabeling, $\mathbf{e}_r = \mathbf{T}_r\mathbf{e} \to \mathbf{T}_r\boldsymbol{\Lambda}\mathbf{e} = \boldsymbol{\Lambda}_r^*\mathbf{T}_r\mathbf{e}$, where $\boldsymbol{\Lambda}_r^*$ is a $s(r) \times s(r)$ matrix, with entries in $\{-1, 0, 1\}$, such that $\boldsymbol{\Lambda}_r^*\boldsymbol{\Lambda}_r^* = \mathbf{I}$. The entries of $\boldsymbol{\Lambda}_r^*$ come from the expansion of $\mathrm{E}\left[\prod_j (1 - Y_{i_j})\right]$, in terms of the MVB moments, over different subsets $\{i_1, \ldots, i_k\}$ of size 1 to r; the factor of 1 cancels from the differencing of $\mathbf{p}$ and $\boldsymbol{\pi}$. If the relabeling is done twice, then we have

$$\mathbf{T}_r\mathbf{e} = \mathbf{T}_r\boldsymbol{\Lambda\Lambda}\mathbf{e} = \boldsymbol{\Lambda}_r^*\mathbf{T}_r\boldsymbol{\Lambda}\mathbf{e} = \boldsymbol{\Lambda}_r^*\boldsymbol{\Lambda}_r^*\mathbf{T}_r\mathbf{e},$$

which shows that $\boldsymbol{\Lambda}_r^*\boldsymbol{\Lambda}_r^* = \mathbf{I}$.

Furthermore with the relabeling, $\boldsymbol{\Gamma} = \mathrm{diag}\,(\boldsymbol{\pi}) - \boldsymbol{\pi\pi}' \to \boldsymbol{\Lambda\Gamma\Lambda}'$, $\boldsymbol{\Xi}_r = \mathbf{T}_r\boldsymbol{\Gamma}\mathbf{T}_r' \to \boldsymbol{\Lambda}_r^*\boldsymbol{\Xi}_r\boldsymbol{\Lambda}_r^{*\prime}$ and

$$\mathbf{e}_r'\boldsymbol{\Xi}^{-1}\mathbf{e}_r \to \mathbf{e}_r'\boldsymbol{\Lambda}_r^{*\prime}(\boldsymbol{\Lambda}_r^{*\prime})^{-1}\boldsymbol{\Xi}_r^{-1}(\boldsymbol{\Lambda}_r^*)^{-1}\boldsymbol{\Lambda}_r^*\mathbf{e}_r = \mathbf{e}_r'\boldsymbol{\Xi}^{-1}\mathbf{e}_r,$$

which establishes the invariance.

For the relabeling for a parametric MVB family and M_r, suppose the relabeling changes $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_\Lambda$ with invertible Jacobian $\mathbf{J} = \partial\boldsymbol{\theta}/\partial\boldsymbol{\theta}_\Lambda$. We just summarize the effect of the relabeling on all of the matrices and vectors in M_r:

$$\boldsymbol{\Delta} \to \boldsymbol{\Lambda\Delta}\mathbf{J}', \quad \boldsymbol{\Delta}_r \to \boldsymbol{\Lambda}_r^*\boldsymbol{\Delta}_r\mathbf{J}', \quad \mathbf{C}_r \to (\boldsymbol{\Lambda}_r^{*\prime})^{-1}\mathbf{C}_r(\boldsymbol{\Lambda}_r^*)^{-1},$$

$$\hat{\boldsymbol{\theta}} \to \hat{\boldsymbol{\theta}}_\Lambda, \quad \mathbf{p} - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}) \to \boldsymbol{\Lambda}_r^*[\mathbf{p} - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})].$$

It follows that M_r is invariant to the 0-1 relabeling.

Local alternatives: expected value of MLE

Consider a parametric family is $f(y; \boldsymbol{\theta})$ which can be continuous or discrete; f is a density relative to measure ν (Lebesgue or counting measure). This subsection concerns a limit of the expected value of the maximum likelihood estimator (MLE) for a sequence of local alternatives when the null hypothesis is a nested submodel of a certain form. The usual regularity conditions are assumed to hold. The technique of derivation can be used for other forms of nested model (e.g., some of the parameters fixed under H_0) but we cannot obtain a result to be used for all forms of nested submodels.

For the submodel, we suppose the parametrization is of the form $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ where $\boldsymbol{\theta}_2 = \beta\mathbf{1}$. We obtain the maximum likelihood estimator based on the submodel, and derive its distribution under local alternatives in the full model. That is, the hypotheses are

$$H_0 : (\boldsymbol{\theta}_1', \beta\mathbf{1}')' \quad \text{vs} \quad H_1 : (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'.$$

For a sequence of local alternatives, we take $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}', \beta_0\mathbf{1}')'$ as a 'true' model, and let $\boldsymbol{\theta}_{1N} = (\boldsymbol{\theta}_{10}', \beta_0\mathbf{1}' + w_N\boldsymbol{\gamma}')'$ be the sequence of alternative parameter values. $\boldsymbol{\gamma}$ is a nonconstant vector that sums to 0 (for identifiability).

Let $\boldsymbol{\theta}_0^* = (\boldsymbol{\theta}_{10}', \beta_0)'$ and $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1', \beta)'$ and we write the density for H_0 as $f^*(;\boldsymbol{\theta}^*) = f(\cdot;(\boldsymbol{\theta}_1', \beta\mathbf{1}')')$. Let $\ell(\boldsymbol{\theta}^*; y) = \log f^*(y;\boldsymbol{\theta}^*)$, $\dot{\ell} = \partial\ell/\partial\boldsymbol{\theta}^*$, $\ddot{\ell} = \partial^2\ell/\partial\boldsymbol{\theta}^*\partial\boldsymbol{\theta}^{*\prime}$.

Suppose that the MLE $\hat{\boldsymbol{\theta}}_N^*$ is a solution of $L(\boldsymbol{\theta}^*) = \sum_{i=1}^N \dot{\ell}(\boldsymbol{\theta}^*, y_{iN}) = 0$, where $y_{1N}, \ldots, y_{NN}$ is a random sample from $f(\cdot;\boldsymbol{\theta}_{1N})$. Take an expansion of L about $\boldsymbol{\theta}_0^*$ to get

$$\mathbf{0} = \sum \dot{\ell}(\hat{\boldsymbol{\theta}}_N; y_{iN}) \approx \sum \dot{\ell}(\boldsymbol{\theta}_0^*; y_{iN}) + \sum \ddot{\ell}(\boldsymbol{\theta}_0^*; y_{iN})(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*) + o_p(||\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*||)$$

or

$$\sqrt{N}\,(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*) \approx \left[-N^{-1}\sum \ddot{\ell}(\boldsymbol{\theta}_0^*; y_{iN})\right]^{-1} N^{-1/2}\sum \dot{\ell}(\boldsymbol{\theta}_0^*; y_{iN}) + o_p(1).$$

Under the sequence of local alternatives,

$$-N^{-1}\sum \ddot{\ell}(\boldsymbol{\theta}_0^*; y_{iN}) \xrightarrow{p} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0^*),$$

where $\boldsymbol{\mathcal{I}}$ is the Fisher information matrix for the model $f^*(\cdot;\boldsymbol{\theta}^*)$. Hence,

$$\sqrt{N}\,\mathrm{E}\,(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*) \approx [\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0^*)]^{-1}\sqrt{N}\,\mathrm{E}\,[\dot{\ell}(\boldsymbol{\theta}_0^*; Y_{1N})]. \tag{A4}$$

Taking an expansion of $f(y;(\boldsymbol{\theta}_{10}', \beta_0\mathbf{1}' + w_N\boldsymbol{\gamma}')')$ about $\boldsymbol{\theta}_2$ leads to

$$\mathrm{E}\,[\dot{\ell}(\boldsymbol{\theta}_0^*; y_{1N})] \approx \int \dot{\ell}(\boldsymbol{\theta}_0^*; y)\left[f(y;(\boldsymbol{\theta}_{10}', \beta_0\mathbf{1}')') + w_N\boldsymbol{\gamma}'\frac{\partial f}{\partial\boldsymbol{\theta}_2}\right]d\nu(y) = w_N\int \dot{\ell}(\boldsymbol{\theta}_0^*; y)\,\boldsymbol{\gamma}'\frac{\partial f}{\partial\boldsymbol{\theta}_2}\,d\nu(y).$$

Finally, if $\sqrt{N}\,w_N \to \epsilon$, then (A4) becomes (as $N \to \infty$)

$$\sqrt{N}\,\mathrm{E}\,(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*) \to \epsilon[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0^*)]^{-1}\int \dot{\ell}(\boldsymbol{\theta}_0^*; y)\,\boldsymbol{\gamma}'\frac{\partial f}{\partial\boldsymbol{\theta}_2}\,d\nu(y). \tag{A5}$$

For a discrete model (ν corresponding to counting measure), write $f(y;\theta) = \pi_y(\theta)$, where y may be a vector, e.g., binary vector of dimension n. Then (A5) becomes (4.13).

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

[2] Bartholomew, D.J. (1998). Scaling unobservable constructs in the social sciences. *Applied Statistics*, **47**, 1–13.

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

[4] 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.

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

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

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

[8] Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, **37**, 62–83.

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

[10] 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.

[11] Ferguson, T.S. (1996). *A Course in Large Sample Theory*. London: Chapman & Hall.

[12] Godambe, V.P. (ed.) (1991). *Estimating Functions*. Oxford: Oxford University Press.

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

[14] Jöreskog, K.G. & Srbom, D. (2001). *LISREL 8*. Chicago, IL: Scientific Software.

[15] 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.

[16] 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.

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

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

[19] Maydeu-Olivares, A. (2002). Limited information estimation and testing of Thurstonian models for preference data. *Mathematical Social Sciences*, **43**, 467–483.

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

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

[22] 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.

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

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

[25] Read, T.R.C. and Cressie, N.A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer.

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

[27] Reiser, M. & Lin (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.

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

[29] Satorra, A. (1989). Alternative test criteria in covariance structure analysis: a unified approach. *Psycholometrika*, **54**, 131–151.

[30] Schuessler, K.F. (1982). *Measuring Social Life Feelings*. San Francisco: Jossey-Bass.

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

[32] Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, **47**, 175–186.

Table 1:
Type I errors (based on 10^4 simulations) using asymptotic $\alpha = 0.05$ level critical values for X^2, L_1, L_2, L_3; MVB probabilities from model (3.3)

(η, γ)	n	N	X^2	L_1	L_2	L_3
(0.5,0.5)	5	100	0.054	0.049	0.051	0.055
	5	1000	0.053	0.053	0.051	0.052
	10	100	0.230	0.051	0.055	0.084
	10	1000	0.089	0.051	0.049	0.055
(0.8,0.5)	5	100	0.071	0.053	0.057	0.066
	5	1000	0.056	0.049	0.054	0.053
	10	100	0.326	0.056	0.081	0.142
	10	1000	0.140	0.052	0.053	0.065

Table 2:
Power of X^2, L_1, L_2, L_3 at level $\alpha = 0.05$ for a sequence of local alternatives, model (3.3)

m	η	γ	ϵ_1	ϵ_2	X_2	L_1	L_2	L_3
5	0.5	0.0	1.0	1.0	0.890	0.952	0.966	0.920
5	0.5	0.1	1.0	1.0	0.648	0.858	0.809	0.700
5	0.5	0.3	1.0	1.0	0.398	0.697	0.553	0.443
5	0.6	0.3	1.0	1.0	0.441	0.718	0.600	0.488
5	0.2	0.3	1.0	1.0	0.554	0.896	0.722	0.606
5	0.5	0.0	0.0	2.0	0.972	0.050	0.995	0.983
5	0.5	0.1	0.0	2.0	0.608	0.050	0.774	0.661
5	0.5	0.3	0.0	2.0	0.202	0.050	0.287	0.223
5	0.2	0.3	0.0	2.0	0.158	0.050	0.212	0.173
10	0.5	0.0	0.5	0.5	0.121	0.542	0.561	0.296
10	0.5	0.1	0.5	0.5	0.073	0.295	0.197	0.118
10	0.5	0.3	0.5	0.5	0.060	0.177	0.106	0.078
10	0.6	0.3	0.5	0.5	0.061	0.184	0.114	0.081
10	0.2	0.3	0.5	0.5	0.063	0.272	0.126	0.087
10	0.5	0.0	0.0	1.0	0.256	0.050	0.952	0.708
10	0.5	0.1	0.0	1.0	0.083	0.050	0.278	0.153
10	0.5	0.3	0.0	1.0	0.057	0.050	0.089	0.069
10	0.2	0.3	0.0	1.0	0.056	0.050	0.078	0.065

Table 3:
Small sample distribution (based on convergent cases from 10^4 simulations) for X^2, M_2, M_3; MVB probabilities from model (4.8); mean, variance and exceedances of asymptotic upper 0.2, 0.1, 0.05, 0.01 quantiles. $(\boldsymbol{\alpha};\boldsymbol{\beta}) = (-1,-.5,0,.5,1;1,1.3,1.6,1.9,2.2)$ for $n = 5$; $(\boldsymbol{\alpha};\boldsymbol{\beta}) = (-1,-.5,.5,1,-1,-.5,.5,1;.5,.9,1.3,1.6,1.6,1.3,.9,.5)$ for $n = 8$. Convergence rates were 63% for $n = 8, N = 100$ and 69% for $n = 5, N = 100$, and over 90% for other cases.

n	N	statistic	df	mean	var.	$\alpha = .2$	$\alpha = .1$	$\alpha = .05$	$\alpha = .01$
5	100	X^2	21	21	104	.21	.14	.10	.05
		M_2	5	4.9	8.6	.18	.09	.04	.006
		M_3	15	15	33	.19	.10	.06	.02
5	1000	X^2	21	21	46	.20	.11	.06	.02
		M_2	5	5.0	10	.20	.10	.05	.009
		M_3	15	15	30	.20	.10	.05	.01
8	100	X^2	239	235	2×10^5	.22	.20	.19	.16
		M_2	20	20	40	.20	.11	.06	.012
		M_3	76	76	300	.25	.18	.13	.06
8	1000	X^2	239	240	1×10^4	.27	.23	.21	.17
		M_2	20	20	39	.20	.09	.05	.009
		M_3	76	76	160	.19	.10	.05	.015
8	2500	X^2	239	240	5×10^3	.27	.22	.18	.12
		M_2	20	20	41	.20	.10	.05	.009
		M_3	76	76	160	.19	.10	.05	.009

Table 4:
Power of X^2, M_2, M_3 at level $\alpha = 0.05$ for a sequence of local alternatives, model (4.8) and hypothesis (4.11), $\epsilon = 10$.

n	$\boldsymbol{\alpha}$	β	γ	X^2	M_2	M_3
5	-1,-0.5,0,0.5,1	1.0	-0.6,-0.3,0,0.3,0.6	0.131	0.136	0.104
5	-1,-0.5,0,0.5,1	1.5	-0.6,-0.3,0,0.3,0.6	0.118	0.120	0.095
5	-1,-0.5,0,0.5,1	2.0	-0.6,-0.3,0,0.3,0.6	0.097	0.098	0.081
5	-1,-0.5,0,0.5,1	1.0	0,-0.6,0.3,-0.6,0.9	0.220	0.358	0.251
5	-1,-0.5,0,0.5,1	1.5	0,-0.6,0.3,-0.6,0.9	0.192	0.311	0.219
5	-1,-0.5,0,0.5,1	2.0	0,-0.6,0.3,-0.6,0.9	0.147	0.230	0.165
8	-1,-0.5,0.5,1,-1,-0.5,0.5,1	1.0	-0.6,-0.3,0.3,0.6,0.6,0.3,-0.3,-0.6	0.122	0.286	0.163
8	-1,-0.5,0.5,1,-1,-0.5,0.5,1	1.5	-0.6,-0.3,0.3,0.6,0.6,0.3,-0.3,-0.6	0.106	0.229	0.136
8	-1,-0.5,0.5,1,-1,-0.5,0.5,1	2.0	-0.6,-0.3,0.3,0.6,0.6,0.3,-0.3,-0.6	0.087	0.165	0.106
8	-1,-0.5,0.5,1,-1,-0.5,0.5,1	1.0	-0.6,-0.3,0.3,0.9, 0.3,-0.3,0.6,-0.9	0.176	0.489	0.270
8	-1,-0.5,0.5,1,-1,-0.5,0.5,1	1.5	-0.6,-0.3,0.3,0.9,0.3,-0.3,0.6,-0.9	0.146	0.392	0.216
8	-1,-0.5,0.5,1,-1,-0.5,0.5,1	2.0	-0.6,-0.3,0.3,0.9,0.3,-0.3,0.6,-0.9	0.112	0.270	0.155

Table 5:
Comparison of asymptotic relative efficiencies (ARE) for WLS/ULS estimators with maximum likelihood, average over 100 simulations with α_i's random with Uniform$(-2, 2)$ distribution, and β_i's random with Uniform$(1, 2)$ distribution. Relative efficiencies with calculated based on diagonal entries and determinants of asymptotic covariance matrices.

n	estimator	quantity	avg(ARE)	SD(ARE)	min(ARE)
5	ULS($r = n$)	α_i	0.78	0.13	0.35
		β_i	0.74	0.14	0.35
		$\det^{1/10}$	0.80	0.05	0.70
5	ULS($r = 2$)	α_i	0.96	0.06	0.70
		β_i	0.93	0.07	0.67
		$\det^{1/10}$	0.96	0.02	0.92
5	ULS($r = 3$)	α_i	0.94	0.07	0.65
		β_i	0.87	0.06	0.63
		$\det^{1/10}$	0.93	0.02	0.88
5	WLS($r = 2$)	α_i	0.99	0.01	0.98
		β_i	0.99	0.01	0.97
		$\det^{1/10}$	0.99	0.01	0.99
8	ULS($r = n$)	α_i	0.62	0.14	0.16
		β_i	0.62	0.16	0.19
		$\det^{1/20}$	0.65	0.04	0.57
8	ULS($r = 2$)	α_i	0.94	0.06	0.65
		β_i	0.89	0.08	0.57
		$\det^{1/20}$	0.93	0.02	0.89
8	ULS($r = 3$)	α_i	0.91	0.10	0.59
		β_i	0.81	0.07	0.54
		$\det^{1/20}$	0.88	0.02	0.84

Table 6:
Values of MLEs and bivariate ULS estimators for the data example from Bartholomew and Knott (1999, pp. 97-98)

parameter	MLE		ULS($r = 2$)	
	estimate	se	estimate	se
α_1	-2.35	0.13	-2.57	0.18
α_2	0.80	0.06	0.80	0.06
α_3	0.99	0.09	1.00	0.10
α_4	-0.67	0.13	-0.63	0.11
α_5	-1.10	0.07	-1.10	0.08
β_1	1.20	0.15	1.44	0.20
β_2	0.71	0.09	0.73	0.09
β_3	1.53	0.17	1.56	0.18
β_4	2.55	0.41	2.34	0.35
β_5	0.92	0.10	0.93	0.11

Table 7:
Values of goodness-of-fit statistics for the data example from Bartholomew and Knott (1999, pp. 97-98)

estimator	statistic	value	df	p-value
MLE	X^2	38.9	21	0.010
MLE	M_2	15.7	5	0.008
MLE	M_3	27.9	15	0.022
MLE	X^2 (item 1 deleted)	17.9	7	0.013
MLE	X^2 (item 2 deleted)	12.0	7	0.101
MLE	X^2 (item 3 deleted)	15.3	7	0.032
MLE	X^2 (item 4 deleted)	19.4	7	0.007
MLE	X^2 (item 5 deleted)	6.0	7	0.540
ULS($r = 2$)	M_5	41.3	21	0.005
ULS($r = 2$)	M_2	16.5	5	0.006
ULS($r = 2$)	M_3	29.1	15	0.016