Limited- and Full-Information Estimation and Goodness-of-Fit Testing in 2^n Contingency Tables: A Unified Framework

Albert MAYDEU-OLIVARES and Harry JOE

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 classes of quadratic form statistics based on the residuals of margins or multivariate moments up to order r. These classes of test statistics are asymptotically chi-squared distributed under the null hypothesis. Further, the marginal residuals are useful for diagnosing lack of fit of parametric models. We show that when r is small ($r = 2, 3$), the proposed statistics have better small-sample properties and are asymptotically more powerful than X^2 for some useful multivariate binary models. Related to these 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.

KEY WORDS: High-dimensional contingency table; Item response modeling; Limited information; Low-dimensional margin; Multivariate Bernoulli distribution; Quadratic form statistics.

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 from, for instance, 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. For instance, Chang, D'Zurilla, and Maydeu-Olivares (1994) considered modeling the responses of 393 individuals to the Beck Hopelessness Scale (BHS) (Beck, Weissman, Lester, and Trexler 1974), a set of $n = 20$ true-or-false questions used to predict depression, suicidal ideation, and suicidal intent. There are 2^{20} ($>10^6$) cells in the contingency table.

A researcher confronted with 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, binary contingency tables most often 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 null distribution. This problem can be overcome by generating the empirical sampling distribution of the statistic using the parametric bootstrap method (e.g., Collins, Fidler, Wugalter, and Long 1993; Bartholomew and Tzamourani 1999). However, this approach may be very time-consuming if the researcher is interested in comparing the fit of several models.

If, as is often the case, the overall tests suggests significant misfit, then a second challenge that a researcher must confront is to identify the source of the misfit. Inspecting cell residuals is often not very useful toward this aim. It is difficult to find trends when inspecting these residuals, and even for moderate n the number of residuals that need to be inspected is too large. And perhaps most important, Bartholomew and Tzamourani (1999) pointed 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. Key references in this literature include works by Reiser (1996), Reiser and Lin (1999), Reiser and VandenBergh (1994), Bartholomew and Tzamourani (1999), and Bartholomew and Leung (2002).

A third challenge that 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 and Knott 1999), which are extremely popular in the social sciences. If the distribution of the latent traits is assumed to be multivariate normal, as is most often the case, then computing the binary pattern probabilities becomes 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 using estimation methods that use information only 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 use only 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 1997, chap. 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 treatment of limited- and full-information estimation and testing in 2^n contingency tables. We provide such a framework in this article under multivariate Bernoulli (MVB) sampling. In Section 2 we provide a convenient representation of the MVB distribution using its

Albert Maydeu-Olivares is Associate Professor, Faculty of Psychology, University of Barcelona, Barcelona, Spain, and Marketing Department, Instituto de Empresa, Madrid, Spain (E-mail: *amaydeu@ub.edu*). Harry Joe is Professor, Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T IZ2 (E-mail: *harry@stat.ubc.ca*). This research was supported by grants BSO2000-0661 and BSO2003-08507 from the Spanish Ministry of Science and Technology, a Distinguished Young Investigator Award of the Catalan Government, and an NSERC Canada grant. The authors thank the referees for comments leading to improvements in the article.

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214504000002069

joint moments. From the asymptotic distribution of sample joint moments (marginal proportions), we obtain the asymptotic distribution of marginal residuals. In Section 3 we propose a family of limited information quadratic form statistics based on these marginal residuals to assess the goodness of fit of simple null hypotheses. These statistics are asymptotically chi-squared distributed under the null hypothesis, 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. We consider two classes of estimators: minimum variance full-information estimators, such as maximum likelihood, and consistent and asymptotically normal estimators, including limited-information estimators. We propose a family of limited-information goodness-of-fit test statistics whose members are asymptotically chi-squared distributed for both classes of estimators. 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 we propose a family of limited-information estimators that is closely linked to our proposed family of limited-information goodness-of-fit tests. 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. In Section 6 we include an example of binary item response data from Bartholomew and Knott (1999) and a summary from the BHS to illustrate our results. Finally, in Section 7 we provide conclusions and a discussion of further research.

2. MULTIVARIATE BERNOULLI DISTRIBUTIONS AND ASYMPTOTIC DISTRIBUTION OF SAMPLE MOMENTS

In this section we characterize the MVB distribution in terms of multivariate moments and define the notation used in the remainder of the article. Consider an n-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ of Bernoulli random variables, with $\dot{\pi}_i = \Pr(Y_i = 1)$, $i = 1, \ldots, n$, and joint distribution

$$\begin{aligned} \pi_{\mathbf{y}} &= \Pr(Y_i = y_i, i = 1, \ldots, n), \\ \mathbf{y} &= (y_1, \ldots, y_n), \qquad y_i \in \{0, 1\}. \end{aligned} \tag{1}$$

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 way of ordering the elements of $\boldsymbol{\pi}(\boldsymbol{\theta})$ is by order of the values of $\mathbf{y}'\mathbf{1} = 0, 1, \ldots, n$, and by lexicographic ordering within a constant sum. An example with $n = 3$ is given later.

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 $\dot{\boldsymbol{\pi}}_1' = (\dot{\pi}_1, \ldots, \dot{\pi}_n)'$, $\dot{\boldsymbol{\pi}}_2$ is the $\binom{n}{2}$-dimensional vector of bivariate noncentral moments with elements $\mathrm{E}(Y_i Y_j) = \Pr(Y_i = 1, Y_j = 1) = \dot{\pi}_{ij}$, $j < i$, 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 1's and 0's, 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 earlier. For example, for $n = 3$, we have

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

The first column of $\mathbf{T}$ is a column of 0's, so we can partition $\mathbf{T} = (\mathbf{0} \quad \dot{\mathbf{T}})$ and write $\dot{\boldsymbol{\pi}} = \dot{\mathbf{T}}\check{\boldsymbol{\pi}}$ with $\boldsymbol{\pi} = \binom{\pi_{0\cdots0}}{\check{\boldsymbol{\pi}}}$. Because $\pi_{0\cdots0} = 1 - \mathbf{1}'\check{\boldsymbol{\pi}}$, 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 MVB distribution up to order $r \le n$, 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 (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. Then we have

$$\sqrt{N}(\dot{\mathbf{p}} - \dot{\boldsymbol{\pi}}) = \mathbf{T}\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}). \tag{2}$$

Because

$$\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}),$$

where $\boldsymbol{\Gamma} = \mathbf{D} - \boldsymbol{\pi}\boldsymbol{\pi}'$ and $\mathbf{D} = \mathrm{diag}(\boldsymbol{\pi})$ (Agresti 1990), it follows from (2) that

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

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 $N\,\mathrm{var}(\dot{p}_a) = \dot{\pi}_a(1 - \dot{\pi}_a)$ and $N\,\mathrm{cov}(\dot{p}_a, \dot{p}_b) = \dot{\pi}_{a\cup b} - \dot{\pi}_a\dot{\pi}_b$, so that, for example, when $n \ge 3$, for $i \ne j$, $j = k$, $N\,\mathrm{var}(\dot{p}_{ij}) = \dot{\pi}_{ij}(1 - \dot{\pi}_{ij})$ 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, and 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, with dimension $s = s(r) = \sum_{i=1}^{r} \binom{n}{i}$. Then we have

$$\sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_r) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Xi}_r), \qquad \boldsymbol{\Xi}_r = \mathbf{T}_r\boldsymbol{\Gamma}\mathbf{T}_r'. \tag{3}$$

Because $\mathbf{T}_r$ 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 hypothesis $H_0: \boldsymbol{\pi} = \boldsymbol{\pi}_0$ versus $H_1: \boldsymbol{\pi} \neq \boldsymbol{\pi}_0$. 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-squared distribution (e.g., Koehler and Larntz 1980), with X^2 having a better small-sample performance than G^2.

The poor approximation of X^2 to its reference asymptotic distribution in sparse 2^n tables can be attributed to the fact that the mean and variance of its reference asymptotic distribution are $2^n - 1$ and $2(2^n - 1)$, 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 and 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 with $\pi_c \ll 2^{-n}$, the type I error X^2 will be larger than the α level based on its asymptotic critical value.

We show in the Appendix that X^2 is a member (with $r = n$) of the family of test statistics

$$L_r = N(\mathbf{p}_r - \boldsymbol{\pi}_r)' \boldsymbol{\Xi}_r^{-1} (\mathbf{p}_r - \boldsymbol{\pi}_r), \qquad r = 1, \ldots, n. \quad (4)$$

That is, X^2 can be written as the weighted discrepancy between the sample and expected joint moments of the MVB distribution,

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

But because large samples are needed to accurately estimate the high-order joint sample moments in X^2, as an alternative to X^2 in sparse tables we propose using L_r, with r depending on the size n of the model relative to sample size N. From (3), under H_0, the quadratic form statistic 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-squared for small N even when there are some small probabilities π_c.

If the L_r test suggests significant misfit, then 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 $\sqrt{N}(\dot{p}_a - \dot{\pi}_a)/\sqrt{\xi_{aa}}$, where 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 $X^2 = L_n$, Table 1 summarizes simulated type I errors using the asymptotic $\alpha = .05$–level critical values. For null MVB distributions, we use examples from the exchangeable beta–binomial MVB model with Bernoulli parameter η and dependence parameter γ [see Joe 1997, sec. 7.1, and our eq. (6)].

Table 1. Type I Errors Using Asymptotic $\alpha = .05$-Level Critical Values for X^2, L_1, L_2, and L_3

(η, γ)	n	N	X^2	L_1	L_2	L_3
(.5, .5)	5	100	.054	.049	.051	.055
	5	1,000	.053	.053	.051	.052
	10	100	.230	.051	.055	.084
	10	1,000	.089	.051	.049	.055
(.8, .5)	5	100	.071	.053	.057	.066
	5	1,000	.056	.049	.054	.053
	10	100	.326	.056	.081	.142
	10	1,000	.140	.052	.053	.065

NOTE: 10^4 replications; MVB probabilities from model (6). The number of cells is 32 and 1,024 for $n = 5$ and 10.

Table 1 has two different null MVB distributions; the one based on $(\eta, \gamma) = (.8, .5)$ has much smaller π_c values than the one based on $(\eta, \gamma) = (.5, .5)$. Table 1 clearly demonstrates the theory referred to earlier. Note that the asymptotic critical values for L_1 and L_2 are quite good even for small $N/2^n$ ratios; they are not as good for L_3 and are much worse for $X^2 = L_n$ as sparseness increases. Thus, for increasingly large and sparse tables, lower r values must be used to ensure accurate type I errors, particularly in models with some small probabilities.

Bartholomew and 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

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

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-squared distributed even in the case of simple null hypotheses. Bartholomew and Leung (2002) used the first three moments of this statistic to approximate its sampling distribution using a chi-squared 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. This is a standard approach for a power comparison; it avoids the simulations needed to get finite-sample critical values of X^2 for power calculations for a sequence of sample sizes.

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

$$H_{1N}: \boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{\epsilon}/\sqrt{N}. \quad (5)$$

Let $\boldsymbol{\delta} = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} \boldsymbol{\epsilon}$. Under (5), from Bishop, Fienberg, and Holland (1975, p. 471), we have

$$\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}_0) \xrightarrow{d} \mathrm{N}(\boldsymbol{\delta}, \mathbf{D}_0 - \boldsymbol{\pi}_0\boldsymbol{\pi}_0') \qquad \text{and}$$

$$\sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_{0r})] \xrightarrow{d} \mathrm{N}(\mathbf{T}_r\boldsymbol{\delta}, \boldsymbol{\Xi}_{r0}),$$

where $\boldsymbol{\Xi}_{r0} = \mathbf{T}_r(\mathbf{D}_0 - \boldsymbol{\pi}_0\boldsymbol{\pi}_0')\mathbf{T}_r'$. Therefore, under (5), the limiting distributions of X^2 and L_r are noncentral chi-squared distributed 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 (100α)th percentile from the chi-squared 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) in 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

$$\pi_{\mathbf{y}} = \pi_{\mathbf{y}}(\eta, \gamma) = \frac{\prod_{i=0}^{k-1}(\eta + i\gamma)\prod_{i=0}^{n-k-1}[1-\eta+i\gamma]}{\prod_{i=0}^{n-1}(1+i\gamma)}, \qquad k = 0, \ldots, n, \qquad y_1 + \cdots + y_n = k. \quad (6)$$

Table 2 gives a representative summary of the asymptotic power results. For (6), $\boldsymbol{\theta} = (\eta, \gamma)'$; 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 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 expect more asymptotic power when more information is used (higher r), but note that all of the information in the beta–binomial MVB distribution can be summarized in the bivariate margins ($r = 2$).

We did a power analysis for another model to show that X^2 sometimes has more asymptotic power. We considered an MVB distribution with higher-order dependence parameters; one simple model for this is the Bahadur representation [see (7.21) in Joe 1997] in the exchangeable case with up to third-order terms. This model has one univariate parameter, one bivariate parameter, and one trivariate parameter. In this case L_2 and L_3 sometimes have more power than X^2 but not always; among our arbitrary choices of parameter vectors and directions of departures from the null, X^2 was most powerful approximately 50% of the time. Also, L_3 is sometimes more powerful than L_2 and is definitely more powerful if the local alternative makes no change to the univariate and bivariate parameters.

Table 2. Power of X^2, L_1, L_2, and L_3 at Level $\alpha = .05$ for a Sequence of Local Alternatives

n	η	γ	ϵ_1	ϵ_2	X^2	L_1	L_2	L_3
5	.5	0	1.0	1.0	.890	.952	.966	.920
5	.5	.1	1.0	1.0	.648	.858	.809	.700
5	.5	.3	1.0	1.0	.398	.697	.553	.443
5	.6	.3	1.0	1.0	.441	.718	.600	.488
5	.2	.3	1.0	1.0	.554	.896	.722	.606
5	.5	0	0	2.0	.972	.050	.995	.983
5	.5	.1	0	2.0	.608	.050	.774	.661
5	.5	.3	0	2.0	.202	.050	.287	.223
5	.2	.3	0	2.0	.158	.050	.212	.173
10	.5	0	.5	.5	.121	.542	.561	.296
10	.5	.1	.5	.5	.073	.295	.197	.118
10	.5	.3	.5	.5	.060	.177	.106	.078
10	.6	.3	.5	.5	.061	.184	.114	.081
10	.2	.3	.5	.5	.063	.272	.126	.087
10	.5	0	0	1.0	.256	.050	.952	.708
10	.5	.1	0	1.0	.083	.050	.278	.153
10	.5	.3	0	1.0	.057	.050	.089	.069
10	.2	.3	0	1.0	.056	.050	.078	.065

NOTE: MVB probabilities from model (6); the number of cells is 32 for $n = 5$ and 1,024 for $n = 10$.

The results of the power comparisons and small-sample behavior show the usefulness of the class of L_r statistics for the case of an 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 $L_n = X^2$, because the asymptotic chi-squared approximation is valid for much smaller N.

4. LIMITED–INFORMATION TESTS OF COMPOSITE NULL HYPOTHESES

In the preceding section we considered goodness-of-fit tests that can be used 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 consider the hypotheses $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$ versus $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$, and we study the analogs of the L_r statistics in (4) when parameters are estimated via maximum likelihood or another estimation method. To do this, throughout this section we assume 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 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 (MLE) or the minimum chi-squared estimator.

4.1 Maximum Likelihood and Asymptotic Minimum Variance Estimators

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

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{B}\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1), \qquad \mathbf{B} = \boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1}, \quad (7)$$

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

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

$$\boldsymbol{\Sigma}_r = \mathbf{T}_r\boldsymbol{\Sigma}\mathbf{T}_r' = \boldsymbol{\Xi}_r - \boldsymbol{\Delta}_r\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}_r' \quad (8)$$

and

$$\boldsymbol{\Delta}_r = \frac{\partial\boldsymbol{\pi}_r(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} = \mathbf{T}_r\frac{\partial\boldsymbol{\pi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} = \mathbf{T}_r\boldsymbol{\Delta} \quad (9)$$

is an $s \times q$ matrix.

For an index a that is a subset of $\{1, \ldots, n\}$ of size $\leq r$, 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 for assessing 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. If a model has many parameters, then different parameter vectors could lead the same margins of order r, and the model would be not identified from the joint moments up to order r. 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. For our theory, we make the assumption that $\boldsymbol{\Delta}_r$ is of full column rank. Note that this assumption implies that $q \leq s$. However, this assumption is introduced only for succinctness in our presentation. When this assumption does not hold, limited-information statistics based on lower-order marginals could still be considered (see Reiser 1996). Also, for our theory, we exclude the case $q = s$. (This happens for example with a log-linear model with interactions terms up to rth order only, in which case our statistic would become 0.)

We could consider the statistic

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

where $\widehat{\boldsymbol{\Sigma}}_r^{+}$ is the Moore–Penrose inverse of $\boldsymbol{\Sigma}_r(\hat{\boldsymbol{\theta}})$. Under H_0, this is asymptotically chi-squared distributed 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 nonzero singular value (not due to sparseness), so that computation of $\widehat{\boldsymbol{\Sigma}}_r^{+}$ is not always stable. Hence we propose an alternative quadratic form statistic, with degrees of freedom $s - q \leq \operatorname{rank}(\boldsymbol{\Sigma}_r)$, based on a matrix that has $\boldsymbol{\Sigma}_r$ as a generalized inverse.

As in Browne (1984), consider an $s \times (s-q)$ 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 (8), $\sqrt{N}\boldsymbol{\Delta}_r^{(c)\prime}\hat{\mathbf{e}}_r = \boldsymbol{\Delta}_r^{(c)\prime}\sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}))$ has asymptotic covariance matrix

$$\boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Sigma}_r\boldsymbol{\Delta}_r^{(c)} = \boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r^{(c)}. \tag{10}$$

Thus,

$$\sqrt{N}\boldsymbol{\Delta}_r^{(c)\prime}\hat{\mathbf{e}}_r \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, \boldsymbol{\Delta}_r^{(c)\prime}\boldsymbol{\Xi}_r\boldsymbol{\Delta}_r^{(c)}\big). \tag{11}$$

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 is $\boldsymbol{\Delta}_r^{(c)}\mathbf{A}$ 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 $\widehat{\mathbf{C}}_r = \mathbf{C}_r(\hat{\boldsymbol{\theta}})$, we then define

$$\begin{aligned} M_r = M_r(\hat{\boldsymbol{\theta}}) &= N\hat{\mathbf{e}}_r'\widehat{\boldsymbol{\Delta}}_r^{(c)}\big(\big[\widehat{\boldsymbol{\Delta}}_r^{(c)}\big]'\widehat{\boldsymbol{\Xi}}_r\widehat{\boldsymbol{\Delta}}_r^{(c)}\big)^{-1}\big[\widehat{\boldsymbol{\Delta}}_r^{(c)}\big]'\hat{\mathbf{e}}_r \\ &= N(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}))'\widehat{\mathbf{C}}_r(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})). \end{aligned} \tag{12}$$

From (11) and Slutsky's theorem, under H_0,

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

where the degrees of freedom are obtained from a result of 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 a result of Khatri (1966) [or Rao (1973, p. 77)], $\mathbf{C}_r$ can be alternatively written as

$$\mathbf{C}_r = \mathbf{C}_r(\boldsymbol{\theta}) = \boldsymbol{\Xi}_r^{-1} - \boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r(\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}\boldsymbol{\Delta}_r)^{-1}\boldsymbol{\Delta}_r'\boldsymbol{\Xi}_r^{-1}. \tag{13}$$

Now consider 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 0 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 and X^2 are equivalent only asymptotically, with $M_n \leq X^2$.

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. (The proof is outlined in the App.)

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:

$$\begin{aligned} \pi_{\mathbf{y}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \Pr(Y_1 = y_1, \ldots, Y_n = y_n) \\ &= \int_{-\infty}^{\infty} \prod_{j=1}^{n} \frac{e^{(\alpha_j + \beta_j x)y_j}}{1 + e^{\alpha_j + \beta_j x}} \phi(x)\, dx, \end{aligned} \tag{14}$$

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

Table 3 gives the means, variances, and empirical rejection rates at $\alpha = .20, .10, .05, .01$ for M_2, M_3, and X^2 with maximum likelihood estimation of a logit-normit model for a five-variable model and an eight-variable model with $N = 100$ and $N = 1{,}000$. Numerical optimization used a quasi-Newton routine with analytic derivatives. Computations used 48-point Gauss–Hermite quadrature for the integrals (14) and their derivatives with respect to α_i and β_i; this is computationally faster, and it matched computations of MLEs to four decimal places when Romberg integration was used with accuracy 10^{-6} in (14) and their derivatives. The tabulated results are based on the simulations for which the iterations for maximum likelihood estimation converged; see comments of Bartholomew and Knott (1999) regarding nonconvergence. 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 the 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

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

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

$$\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{H}\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1) \tag{15}$$

Table 3. Small-Sample Distribution for X^2, M_2, and M_3; MVB Probabilities for a Logit-Normit Latent Trait Model: Mean, Variance, and Exceedances of Asymptotic Upper .2, .1, .05, and .01 Quantiles

n	N	Statistic	df	Mean	Variances	$\alpha = .2$	$\alpha = .1$	$\alpha = .05$	$\alpha = .01$
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	1,000	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	1,000	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	2,500	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

NOTE: 10^4 replications. Convergence rates were 63% for $n = 8$ and $N = 100$, 69% for $n = 5$ and $N = 100$, and >90% for other cases. $(\alpha; \beta) = (-1, -.5, 0, .5, 1; 1, 1.3, 1.6, 1.9, 2.2)$ for $n = 5$; $(\alpha; \beta) = (-1, -.5, .5, 1, -1, -.5, .5, 1;\ .5, .9, 1.3, 1.6, 1.6, 1.3, .9, .5)$ for $n = 8$.

for some $q \times 2^n$ matrix $\mathbf{H}$; see Section 5 for some examples. We derive the asymptotic distribution of the vector of cell residuals, $\tilde{\mathbf{e}} = \mathbf{p} - \boldsymbol{\pi}(\tilde{\boldsymbol{\theta}})$, for (15). Note that $\boldsymbol{\pi}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\pi}(\boldsymbol{\theta}) = \boldsymbol{\Delta}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(N^{-1/2}) = \boldsymbol{\Delta}\mathbf{H}(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(N^{-1/2})$. Because $\mathbf{p} - \boldsymbol{\pi}(\tilde{\boldsymbol{\theta}}) = [\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})] - [\boldsymbol{\pi}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\pi}(\boldsymbol{\theta})]$, we have that $\sqrt{N}\tilde{\mathbf{e}} = (\mathbf{I} - \boldsymbol{\Delta}\mathbf{H})(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1)$, 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 $r \geq r_0$ (with r_0 as defined in Sec. 4.1). Let the vector of residuals of the moments be $\tilde{\mathbf{e}}_r = \mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})$. Because $\tilde{\mathbf{e}}_r = \mathbf{T}_r\tilde{\mathbf{e}}$, the asymptotic distribution of these marginal residuals is [using (9)] $\sqrt{N}\tilde{\mathbf{e}}_r \xrightarrow{d} \mathrm{N}(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_r)$, with

$$\widetilde{\boldsymbol{\Sigma}}_r = (\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{H})\boldsymbol{\Gamma}(\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{H})'. \tag{16}$$

To test composite null hypotheses with this class of estimators, we may use the $M_r = M_r(\tilde{\boldsymbol{\theta}})$ statistic (12) with $\tilde{\boldsymbol{\theta}}$ in place of $\hat{\boldsymbol{\theta}}$. This is because if $\boldsymbol{\Delta}_r^{(c)}$ is an $s \times (s - q)$ orthogonal complement to $\boldsymbol{\Delta}_r$, then $\sqrt{N}\boldsymbol{\Delta}_r^{(c)\prime}\tilde{\mathbf{e}}_r = \boldsymbol{\Delta}_r^{(c)\prime}\sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}}))$ 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 side of (10).

Thus we have shown that under H_0, M_r is asymptotically χ^2_{s-q} if $\tilde{\boldsymbol{\theta}}$ is any $\sqrt{N}$-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. 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 $\sqrt{N}$-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, we can compare the asymptotic power of X^2 and M_r under a sequence of local alternatives. There are several ways of specifying the null and alternative hypotheses; we take the special case where the null hypothesis is a nested model with parameters to be estimated, because in fitting models to categorical data, one often checks whether a simpler (nested) version of a model explains the data adequately.

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

For testing, the hypotheses are

$$H_0 : (\boldsymbol{\theta}_1', \beta\mathbf{1}')' \qquad \text{versus} \qquad H_1 : (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'. \tag{17}$$

For a sequence of local alternatives, we take $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}', \beta_0\mathbf{1}')'$ as a "true" model and let $\boldsymbol{\theta}_N = (\boldsymbol{\theta}_{10}', \beta_0\mathbf{1}' + w_N\boldsymbol{\gamma}')'$ 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 $\boldsymbol{\pi}(\boldsymbol{\theta}_N)$. Under the foregoing sequence of local alternatives, $\hat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}_0^*$ and $\boldsymbol{\Sigma}_r(\hat{\boldsymbol{\theta}}_N) \xrightarrow{p} \boldsymbol{\Sigma}_r(\boldsymbol{\theta}_0^*)$. 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 0 in expectation, and expanding the second term leads to

$$\begin{aligned}
&\sqrt{N}\mathrm{E}[\boldsymbol{\pi}_r(\boldsymbol{\theta}_N) - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}_N)] \\
&\quad = \sqrt{N}[\boldsymbol{\pi}_r(\boldsymbol{\theta}_N) - \boldsymbol{\pi}_r(\boldsymbol{\theta}_0)] - \sqrt{N}\mathrm{E}[\boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}_N) - \boldsymbol{\pi}_r(\boldsymbol{\theta}_0)] \\
&\quad = \sqrt{N}\left[\frac{\partial\boldsymbol{\pi}_r(\boldsymbol{\theta}_0)}{\partial(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')}(\boldsymbol{\theta}_N - \boldsymbol{\theta}_0)\right. \\
&\qquad\qquad \left. - \frac{\partial\boldsymbol{\pi}_r(\boldsymbol{\theta}_0^*)}{\partial(\boldsymbol{\theta}_1', \beta)}\mathrm{E}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*) + o_p(\|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*\|)\right] \\
&\quad = \sqrt{N}\left[\frac{\partial\boldsymbol{\pi}_r(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_2'}w_N\boldsymbol{\gamma}\right] - \epsilon\frac{\partial\boldsymbol{\pi}_r(\boldsymbol{\theta}_0^*)}{\partial(\boldsymbol{\theta}_1', \beta)}\boldsymbol{\zeta} + o_p(1) \\
&\quad \to \epsilon\left[\frac{\partial\boldsymbol{\pi}_r(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_2'}\boldsymbol{\gamma} - \frac{\partial\boldsymbol{\pi}_r(\boldsymbol{\theta}_0^*)}{\partial(\boldsymbol{\theta}_1', \beta)}\boldsymbol{\zeta}\right] \stackrel{\text{def}}{=} \boldsymbol{\delta}_r,
\end{aligned} \tag{18}$$

where, from the Appendix,

$$\epsilon\boldsymbol{\zeta} = \lim \sqrt{N}\mathrm{E}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*) = \epsilon[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0^*)]^{-1} \sum_{\mathbf{y}} \frac{\partial \log \pi_{\mathbf{y}}(\boldsymbol{\theta}_{10}, \beta_0\mathbf{1})}{\partial(\boldsymbol{\theta}_1', \beta)'} \cdot \boldsymbol{\gamma}' \frac{\partial \pi_{\mathbf{y}}(\boldsymbol{\theta}_{10}, \beta_0\mathbf{1})}{\partial \boldsymbol{\theta}_2}, \tag{19}$$

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 (18).

Under the sequence of local alternatives,

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

For the comparison with the usual chi-squared statistic,

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

using an argument analogous to that earlier.

Using standard results for noncentral 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 $\boldsymbol{\delta}_r'\mathbf{C}_r\boldsymbol{\delta}_r$, and the degrees of freedom are $2^n - 1 - q$ and $s - q$. The power calculations are then like those in Section 3.2. 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 (A.4) 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 (14) with $H_0 : \boldsymbol{\beta} = \beta\mathbf{1}$, the powers for X^2 and M_r $(r = 2, 3)$ were computed under sequences of local alternatives. The model under the null hypothesis is known in the educational testing literature as a one-parameter logistic (or Rasch) model with a normally distributed latent trait (e.g., Thissen 1982). Table 4 gives some representative results showing that both M_2 and M_3 are more powerful than X^2, with M_2 the most powerful of the three. Note that model (14) is identified from the univariate and bivariate moments for $n \geq 3$. As a check on the asymptotic power results, we performed simulations to compare 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 for small r have more power than the X^2 statistic. The technique used in this section can be used more generally to assess the power of M_r for other full and nested models. Future investigations of power properties of M_r for other models should aid the development of guidelines for models in which M_r is highly effective.

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 is 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

$$F_r = F_r(\boldsymbol{\theta}) = (\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta}))'\widehat{\mathbf{W}}(\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta})), \tag{20}$$

where $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W} = \mathbf{W}(\boldsymbol{\theta})$, a positive-definite matrix. Obvious choices for $\widehat{\mathbf{W}}$ in (20) 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

$$F_r(\boldsymbol{\theta}) = (\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta}))'\mathbf{W}(\boldsymbol{\theta})(\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta})). \tag{21}$$

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

$$\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{K}\sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_r(\boldsymbol{\theta})) + o_p(1) = \mathbf{K}\mathbf{T}_r\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + o_p(1), \tag{22}$$

where $\mathbf{K} = (\boldsymbol{\Delta}_r'\mathbf{W}\boldsymbol{\Delta}_r)^{-1}\boldsymbol{\Delta}_r'\mathbf{W}$. Note that (22) has the form of (15). Furthermore, we have

$$\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{K}\boldsymbol{\Xi}_r\mathbf{K}') \tag{23}$$

and

$$\sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})) \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, (\mathbf{I} - \boldsymbol{\Delta}_r\mathbf{K})\boldsymbol{\Xi}_r(\mathbf{I} - \boldsymbol{\Delta}_r\mathbf{K})'\big), \tag{24}$$

because, from (16), $\widetilde{\boldsymbol{\Sigma}}_r = (\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{K}\mathbf{T}_r)\boldsymbol{\Gamma}(\mathbf{T}_r - \boldsymbol{\Delta}_r\mathbf{K}\mathbf{T}_r)' = (\mathbf{I} - \boldsymbol{\Delta}_r\mathbf{K})\boldsymbol{\Xi}_r(\mathbf{I} - \boldsymbol{\Delta}_r\mathbf{K})'$.

Table 4. Power of X^2, M_2, and M_3 at Level $\alpha = .05$ for a Sequence of Local Alternatives, Model (14) and Hypothesis (17), $\epsilon = 10$

n	$\boldsymbol{\alpha}$	β	$\boldsymbol{\gamma}$	X^2	M_2	M_3
5	−1, −.5, 0, .5, 1	1.0	−.6, −.3, 0, .3, .6	.131	.136	.104
	−1, −.5, 0, .5, 1	1.5	−.6, −.3, 0, .3, .6	.118	.120	.095
	−1, −.5, 0, .5, 1	2.0	−.6, −.3, 0, .3, .6	.097	.098	.081
	−1, −.5, 0, .5, 1	1.0	0, −.6, .3, −.6, .9	.220	.358	.251
	−1, −.5, 0, .5, 1	1.5	0, −.6, .3, −.6, .9	.192	.311	.219
	−1, −.5, 0, .5, 1	2.0	0, −.6, .3, −.6, .9	.147	.230	.165
8	−1, −.5, .5, 1, −1, −.5, .5, 1	1.0	−.6, −.3, .3, .6, .6, .3, −.3, −.6	.122	.286	.163
	−1, −.5, .5, 1, −1, −.5, .5, 1	1.5	−.6, −.3, .3, .6, .6, .3, −.3, −.6	.106	.229	.136
	−1, −.5, .5, 1, −1, −.5, .5, 1	2.0	−.6, −.3, .3, .6, .6, .3, −.3, −.6	.087	.165	.106
	−1, −.5, .5, 1, −1, −.5, .5, 1	1.0	−.6, −.3, .3, .9, .3, −.3, .6, −.9	.176	.489	.270
	−1, −.5, .5, 1, −1, −.5, .5, 1	1.5	−.6, −.3, .3, .9, .3, −.3, .6, −.9	.146	.392	.216
	−1, −.5, .5, 1, −1, −.5, .5, 1	2.0	−.6, −.3, .3, .9, .3, −.3, .6, −.9	.112	.270	.155

For the special case where $\mathbf{W}(\boldsymbol{\theta}) = \boldsymbol{\Xi}_r^{-1}(\boldsymbol{\theta})$, with $\widehat{\mathbf{W}}$ in (20) corresponding to $\widehat{\boldsymbol{\Xi}}_r^{-1}$, there are some simplifications of the results. Equations (23) and (24) simplify to

$$\begin{aligned}&\sqrt{N}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (\boldsymbol{\Delta}_r' \boldsymbol{\Xi}_r^{-1} \boldsymbol{\Delta}_r)^{-1}),\\ &\sqrt{N}(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})) \\ &\quad \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_r = \boldsymbol{\Xi}_r - \boldsymbol{\Delta}_r(\boldsymbol{\Delta}_r' \boldsymbol{\Xi}_r^{-1} \boldsymbol{\Delta}_r)^{-1} \boldsymbol{\Delta}_r'\big),\end{aligned} \tag{25}$$

and we obtain the optimal estimator within the class of the form of WLS 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 (12) that looks more like L_r in (4),

$$Q_r = N(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}}))' \widehat{\boldsymbol{\Xi}}_r^{-1} (\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})). \tag{26}$$

From the theory of quadratic forms on normal random variables (Rao 1973, sec. 3b.4) and Slutsky's theorem, Q_r is asymptotically chi-squared distributed, because $\boldsymbol{\Xi}_r^{-1} \widetilde{\boldsymbol{\Sigma}}_r = \mathbf{I} - \boldsymbol{\Xi}_r^{-1} \boldsymbol{\Delta}_r (\boldsymbol{\Delta}_r' \boldsymbol{\Xi}_r^{-1} \boldsymbol{\Delta}_r)^{-1} \boldsymbol{\Delta}_r'$ [with $\widetilde{\boldsymbol{\Sigma}}_r$ in (25)] is idempotent.

Another way to show this asymptotic result, with the degrees of freedom, is as follows. Equation (26) can be considered a special case of

$$M_r' = M_r'(\tilde{\boldsymbol{\theta}}) = N(\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}}))' \widehat{\mathbf{C}}_r (\mathbf{p}_r - \boldsymbol{\pi}_r(\tilde{\boldsymbol{\theta}})), \tag{27}$$

where $\widehat{\mathbf{C}}_r$ is $\mathbf{C}_r(\boldsymbol{\theta})$ given by (13) evaluating all of 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} under H_0. The estimator obtained by minimizing (20) satisfies $(\mathbf{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}))' \widehat{\mathbf{W}} \boldsymbol{\Delta}_r = \mathbf{0}'$ from the gradient of (20), and for $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_r^{-1}$, (27) becomes (26) as the second term (after substitution for $\widehat{\mathbf{C}}_r$) becomes 0. Hence $NF_r(\tilde{\boldsymbol{\theta}}) = Q_r = M_r'$ when $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_r^{-1}$.

Limited-information methods have been considered mainly for estimating normit-normit (see Bartholomew and Knott 1999) and related latent trait models. Computing cell probabilities is difficult in these models, because they may involve high-dimensional normal integrals. However, these models can be estimated from the univariate and bivariate Bernoulli sample moments avoiding altogether the need of high-dimensional normal integrals (see Takane and de Leeuw 1987). The use of limited-information methods to estimate the normit-normit model (also known as the multidimensional normal ogive model) was first proposed by Christoffersson (1975), who minimized F_2 with $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_2^{-1}$. He also showed that for this particular estimator, $NF_r(\tilde{\boldsymbol{\theta}})$ is asymptotically chi-squared distributed. Other multistage approaches based on the information contained in the univariate and bivariate margins of the table have been proposed to estimate latent trait models (see Jöreskog 1994; Lee, Poon, and Bentler 1995; Maydeu-Olivares 2001; Muthén 1978, 1984, 1993).

Here we have placed Christoffersson's (1975) results in the general context of the family of estimators (20). Within this general framework, we find that minimizing F_n with $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Xi}}_n^{-1}$ is equivalent to minimizing the minimum modified chi-squared function $N \sum_{c=1}^{2^n} (p_c - \pi_c)^2 / p_c$. However, for large n such as $n > 25$, Christoffersson's estimator becomes unattractive, because a large weight matrix must be inverted. Furthermore, large samples may be needed to estimate the fourth-order probabilities involved in $\boldsymbol{\Xi}_2$ using sample proportions. Alternatively, we could minimize F_2 in (20) with $\widehat{\mathbf{W}} = (\mathrm{diag}(\widehat{\boldsymbol{\Xi}}_2))^{-1}$ or $\widehat{\mathbf{W}} = \mathbf{I}$, or (21) with $\mathbf{W}(\boldsymbol{\theta}) = (\mathrm{diag}(\boldsymbol{\Xi}_2(\boldsymbol{\theta}))^{-1}$. These estimators are extremely attractive from a computational standpoint, 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. Table 5 provides some results for model (14) 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 (14); for $n = 5, 8$, with the α_i's random with uniform$(-2, 2)$ distribution and the β_i's random with uniform$(1, 2)$ distribution. AREs were calculated based on diagonal entries and determinants of asymptotic covariance matrices. The matrices involved in the calculations in Table 5 are as follows:

(a) The asymptotic covariance matrix of the MLE is $\boldsymbol{\mathcal{I}}^{-1}$ from (7).

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

(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 that the WLS estimators in (c) with $r = 2, 3$ are very highly efficient with efficiency in the .99–1.00$^-$ range. Note that ULS

Table 5. Comparison of AREs for WLS/ULS Estimators With Maximum Likelihood; Average Over 100 Simulations

n	*Estimator*	*Quantity*	*avg(ARE)*	*SD(ARE)*	*min(ARE)*
5	ULS($r = 2$)	α_i	.96	.06	.70
		β_i	.93	.07	.67
		$\det^{1/10}$	.96	.02	.92
5	ULS($r = 3$)	α_i	.94	.07	.65
		β_i	.87	.06	.63
		$\det^{1/10}$	.93	.02	.88
5	ULS($r = n$)	α_i	.78	.13	.35
		β_i	.74	.14	.35
		$\det^{1/10}$	.80	.05	.70
5	WLS($r = 2$)	α_i	.99	.01	.98
		β_i	.99	.01	.97
		$\det^{1/10}$	.99	.01	.99
8	ULS($r = 2$)	α_i	.94	.06	.65
		β_i	.89	.08	.57
		$\det^{1/20}$	.93	.02	.89
8	ULS($r = 3$)	α_i	.91	.10	.59
		β_i	.81	.07	.54
		$\det^{1/20}$	.88	.02	.84
8	ULS($r = n$)	α_i	.62	.14	.16
		β_i	.62	.16	.19
		$\det^{1/20}$	.65	.04	.57

NOTE: The α_i's are random with uniform$(-2, 2)$ distribution, and the β_i's are random with uniform$(1, 2)$ distribution. AREs were calculated based on diagonal entries and determinants of asymptotic covariance matrices. For WLS, .99 means $\geq .99$, and the AREs of WLS($r = 2$) for $n = 8$ and WLS($r = 3$) for $n = 5, 8$ are at least as good as WLS($r = 2$) for $n = 5$. Also the AREs for WLS($r = n$) are 1.

with $r = n$ has worse efficiency than ULS with $r = 2, 3$. The $r = n$ case is probably worse, because it weights the small n-dimensional probabilities the same as the larger ones. For $r = 2, 3$, the marginal probabilities tend not to 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 found that the comparisons are similar to the AREs; the MLE is only marginally better in terms of mean squared error.

6. NUMERICAL EXAMPLES

A common task in the social sciences is to measure unobservable constructs, such as cognitive abilities, personality traits, and social attitudes, by administering a set of items written to be indicators of the unobservable constructs (see Bartholomew 1998). The BHS dataset that we described in Section 1 is an example of this practice. Before we present brief results for the BHS example, we present results for a nonsparse dataset where the asymptotic p values of X^2 are likely to be accurate.

6.1 The Social Life Feelings Scale 10

Our nonsparse example is taken from Bartholomew and Knott (1999, pp. 97–98), who used data from an original study reported by Schuessler (1982). The data consist of the responses of $N = 1{,}490$ German respondents to $n = 5$ binary questions intended to measure economic self-determination (the Social Life Feelings Scale 10). Their responses were collected in a $2^5 = 32$ contingency table. Bartholomew and Knott (1999) used maximum likelihood to estimate a logit-normit latent trait model (14), where the latent trait is the unobservable construct being measured.

To illustrate the use of limited-information estimation, Table 6 provides our maximum likelihood and bivariate ULS ($r = 2$) estimates. Our MLE parameter estimates and standard errors (SEs) agree with those reported by Bartholomew and 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 and Knott (1999), we did not pool cells in computing X^2. Nevertheless, our p value agrees with those reported by these authors.

Clearly, the model does not fit well in this situation, and we proceed to identify the source of the misfit using the MLEs. From the standardized cell residuals, the binary patterns that show significant misfit are (1 0 0 1 1), (0 0 1 1 1), (1 0 1 1 0), (1 1 1 1 0), and (1 1 1 1 1). These residuals suggest that the model does not fit well for item 4. However, the standardized marginal residuals up to third order (see Sec. 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 $(1, 4, 5)$, clearly suggesting that the model does not fit well for item 5. To verify both conjectures, we fitted a logit-normit model to all five combinations of four items. The results, presented in the second part of Table 7, clearly indicate that economic self-determination is best measured by the first four items of this scale, as suggested by the marginal residuals.

Table 6. Values of MLEs and Bivariate ULS Estimators for the Data Example From Bartholomew and Knott (1999)

	MLE		ULS($r = 2$)	
Parameter	Estimate	SE	Estimate	SE
α_1	−2.35	.13	−2.57	.18
α_2	.80	.06	.80	.06
α_3	.99	.09	1.00	.10
α_4	−.67	.13	−.63	.11
α_5	−1.10	.07	−1.10	.08
β_1	1.20	.15	1.44	.20
β_2	.71	.09	.73	.09
β_3	1.53	.17	1.56	.18
β_4	2.55	.41	2.34	.35
β_5	.92	.10	.93	.11

Table 7. Values of Goodness-of-Fit Statistics for the Data Example From Bartholomew and Knott (1999)

Estimator	Statistic	Value	df	p value
MLE	X^2	38.9	21	.01
MLE	M_2	15.7	5	.01
MLE	M_3	27.9	15	.02
ULS($r = 2$)	M_5	41.3	21	.01
ULS($r = 2$)	M_2	16.5	5	.01
ULS($r = 2$)	M_3	29.1	15	.02
MLE	X^2 (item 1 deleted)	17.7	7	.01
MLE	X^2 (item 2 deleted)	12.0	7	.10
MLE	X^2 (item 3 deleted)	15.2	7	.03
MLE	X^2 (item 4 deleted)	19.4	7	.01
MLE	X^2 (item 5 deleted)	6.0	7	.55
MLE	M_2 (item 1 deleted)	3.9	2	.14
MLE	M_2 (item 2 deleted)	10.6	2	.01
MLE	M_2 (item 3 deleted)	7.7	2	.02
MLE	M_2 (item 4 deleted)	9.1	2	.01
MLE	M_2 (item 5 deleted)	1.9	2	.38

NOTE: For MLE estimation, $X^2 = M_5$.

6.2 Beck Hopelessness Scale

As suggested by Beck et al. (1974), for 12 of the variables on the BHS a 1 was assigned if the respondent endorsed the item, and 0 was assigned otherwise. The remaining nine items were inverse-coded; 0 was assigned if the item was endorsed, and 1 was assigned otherwise. With this coding, the correlations among all binary variables are positive. For these data, the bivariate marginal tables are not sparse, but some trivariate marginal tables have some small counts. We estimated a logit-normit model to these data using maximum likelihood and bivariate ULS estimation. We found that M_2 was 231.5 with the former and 239.2 with the latter, with 170 degrees of freedom, so the model does not fit well ($p < .002$). There were 20 univariate and bivariate significant residuals ($\alpha = .05$). Item 20 had the largest univariate residual and was involved in the largest bivariate residual. Thus the residual analysis suggests deleting this item. After removing this item, the model fits better: $M_2 = 183.7$ and 187.3 on 152 degrees of freedom for MLE and ULS, $p \approx .03$. For this model, the MLE of the intercepts (α's) ranged from −6.2 (SE = 1.2) to .21 (.12), and the MLE of the slopes (β's) ranged from .32 (.15) to 4.2 (1.0).

7. DISCUSSION AND CONCLUSIONS

A 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, because 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 to this problem have been proposed: 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, limited-information testing.

Our approach is based in the observation that MVB models can be equivalently specified as a set of restrictions on the 2^n cell probabilities or on the $2^n - 1$ set of joint moments of the distribution. We propose using only a subset of the moment restrictions to test the fully specified model (hence the term "limited-information testing"). Because we advocate choosing a set of low-order joint moments for testing, our approach amounts to pooling cells in a systematic way, so that the resulting statistics have a known (asymptotic) distribution.

Toward 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, and 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 used instead of X^2, because 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, because the former have more precise empirical type I errors and may be asymptotically more powerful than the latter. Theoretically, the asymptotic variances of L_r and M_r are influenced by the smallest marginal probability of dimension $\min\{2r, n\}$. This property, combined with our simulation results, suggest that the asymptotic null distribution of M_r and L_r can be acceptable if the rth-order margins are not sparse, and that larger sample sizes are needed as r increases for the null asymptotics to be valid. Note that L_r and M_r have no power to distinguish among models with the same margins up to order r but different higher-order margins.

If the model is identified from the margins up to order r ($\boldsymbol{\Delta}_r$ has full column rank q) and $s(r) > q$, and if a consistent and asymptotically normal estimator is used, then 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. 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 earlier. 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, then if this is poor, it is necessary to determine the source of the misfit. Following Reiser (1996), we suggest using marginal residuals that are asymptotically standard normal. As our numerical example illustrates, the use of these residuals can be much more informative than using cell residuals.

In practice, we recommend using M_2 for testing composite hypotheses when the model is identified using only univariate and bivariate information. Only up to bivariate sample moments and up to four-way model probabilities are involved in the computation of M_2. As a result, its asymptotic distribution under the null and alternative hypotheses should be well approximated with samples of a few hundred observations even for large models. This is the case in our simulations for the logit-normit model where, in addition, highest power is obtained with M_2. Additional Monte Carlo simulations are needed to determine the sample size needed for M_2 to yield accurate type I errors for different parametric models and for an increasing number of observed variables. Also, additional work is needed to investigate the use of a quadratic form statistic like M_r when $s > q$ but $\boldsymbol{\Delta}_r$ is not of full column rank q (i.e., the model is not identified from MVB moments up to order r).

In this article we have also considered limited-information estimators. In psychometrics, multistage estimators that use the information contained in the univariate and bivariate margins of the table are often used to estimate models for which computing cell probabilities is difficult. Popular software packages such as LISREL (Jöreskog and Sörbom 2001), EQS (Bentler 1995), and MPLUS (Muthén and Muthén 2001) implement these estimators to estimate normit-normit and related latent trait models. 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 (20), 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.

In choosing among limited-information estimators, for computational reasons we recommend estimators based on univariate and bivariate information. Our small-sample simulations with the logit-normit model suggest that the MLE is only marginally better in terms of mean squared error than the bivariate ULS estimator. Similar simulation results have been obtained by Finger (2002) using the normit-normit model. More research is needed using a variety of models to investigate the empirical behavior of limited-information estimators relative to MLEs.

As n gets larger, certain computational details must be considered to manage the computations within available computer memory. In future research, we will provide other related approaches that are computationally simpler. Also, here we have not covered 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: PROOFS

A.1 $L_n = X^2$ and $M_n(\hat{\theta}) \leq X^2(\hat{\theta})$ With Equality for the Maximum Likelihood Estimator

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 because $\boldsymbol{\pi}_n = \dot{\boldsymbol{\pi}}$ 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 $\mathbf{e} = \mathbf{p} - \boldsymbol{\pi}$. Because $\dot{\mathbf{e}} = \dot{\mathbf{T}}\check{\mathbf{e}}$,

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

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

$$\boldsymbol{\Xi}^{-1} = ((\dot{\mathbf{T}})')^{-1} (\check{\mathbf{D}}^{-1} + \mathbf{1}D_0^{-1}\mathbf{1}')\dot{\mathbf{T}}^{-1}, \tag{A.2}$$

where $D_0 = \pi_{0\ldots0}$. Thus (A.1) is the same as $N(\check{\mathbf{e}}'\check{\mathbf{D}}^{-1}\check{\mathbf{e}} + \check{\mathbf{e}}'\mathbf{1}D_0^{-1}\mathbf{1}'\check{\mathbf{e}})$. Because $\mathbf{e}$ can be partitioned as $\mathbf{e}' = (e_0, \check{\mathbf{e}}')'$, where $e_0 = -\mathbf{1}'\check{\mathbf{e}}$, then (A.1) 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

$$\begin{aligned} M_n &= N\dot{\mathbf{e}}'\widehat{\mathbf{C}}_n\dot{\mathbf{e}}, \qquad \widehat{\mathbf{C}}_n = \mathbf{C}_n(\hat{\boldsymbol{\theta}}), \\ \mathbf{C}_n &= \boldsymbol{\Xi}^{-1} - \boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_n(\boldsymbol{\Delta}_n'\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_n)^{-1}\boldsymbol{\Delta}_n'\boldsymbol{\Xi}^{-1} \\ &= \boldsymbol{\Delta}_n^{(c)}\left(\boldsymbol{\Delta}_n^{(c)\prime}\boldsymbol{\Xi}\boldsymbol{\Delta}_n^{(c)}\right)^{-1}\boldsymbol{\Delta}_n^{(c)\prime}. \end{aligned}$$

We claim that

$$M_n = N\hat{\mathbf{e}}'\widehat{\mathbf{U}}(\hat{\mathbf{e}}), \qquad \widehat{\mathbf{U}} = \mathbf{U}(\hat{\boldsymbol{\theta}}),$$

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

$$M_n = X^2(\hat{\boldsymbol{\theta}}) - N\hat{\mathbf{e}}'\widehat{\mathbf{V}}\hat{\mathbf{e}}, \qquad \widehat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}}),$$

where

$$\mathbf{V}(\boldsymbol{\theta}) = \mathbf{D}^{-1}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{D}^{-1}.$$

Let hats on matrices denote evaluation at $\hat{\boldsymbol{\theta}}$. For the proof of the claim, from the foregoing 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 $\boldsymbol{\Delta}' = (\,\boldsymbol{\Delta}_0' \quad \check{\boldsymbol{\Delta}}'\,)$, we have $\boldsymbol{\Delta}_n = \dot{\mathbf{T}}\check{\boldsymbol{\Delta}}$. Thus, from (A.2), $\boldsymbol{\Delta}_n'\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_n$ 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 because $\mathbf{1}'\boldsymbol{\Delta} = \mathbf{0}'$, $\mathbf{1}'\check{\boldsymbol{\Delta}} = -\boldsymbol{\Delta}_0$, and $\boldsymbol{\Delta}_n'\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_n = \boldsymbol{\Delta}'\mathbf{D}^{-1}\boldsymbol{\Delta}$ at $\hat{\boldsymbol{\theta}}$. Similarly, because $\dot{\mathbf{e}} = \dot{\mathbf{T}}\check{\mathbf{e}}$,

$$\begin{aligned} \dot{\mathbf{e}}'\widehat{\boldsymbol{\Xi}}^{-1}\widehat{\boldsymbol{\Delta}}_n &= \check{\mathbf{e}}'\dot{\mathbf{T}}'\dot{\mathbf{T}}'^{-1}\left(\widehat{\check{\mathbf{D}}}^{-1} + \mathbf{1}\hat{D}_0^{-1}\mathbf{1}'\right)\dot{\mathbf{T}}^{-1}\dot{\mathbf{T}}\widehat{\check{\boldsymbol{\Delta}}} \\ &= \check{\mathbf{e}}'\widehat{\check{\mathbf{D}}}^{-1}\widehat{\check{\boldsymbol{\Delta}}} + e_0\hat{D}_0^{-1}\widehat{\boldsymbol{\Delta}}_0' = \hat{\mathbf{e}}'\widehat{\mathbf{D}}^{-1}\widehat{\boldsymbol{\Delta}}, \end{aligned} \tag{A.3}$$

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

Finally, (A.3) is $\mathbf{0}'$ if $\hat{\boldsymbol{\theta}}$ is the MLE, because it is the vector of score equations that the MLE satisfies. So $M_n = X^2$ for the MLE.

A.2 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 Sec. 2) is completely reversed; that is, the probability vector becomes $\boldsymbol{\Lambda}\boldsymbol{\pi}$, where $\boldsymbol{\Lambda}$ is a $2^n \times 2^n$ matrix that has 1's in the $(i, 2^n - i)$ positions for all i and 0's 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 an $s(r) \times s(r)$ 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 $\mathrm{E}[\prod_j(1 - Y_{i_j})]$, 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

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

which shows that $\boldsymbol{\Lambda}_r^*\boldsymbol{\Lambda}_r^* = \mathbf{I}$. Furthermore with the relabeling, $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' \to \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Lambda}'$, $\boldsymbol{\Xi}_r = \mathbf{T}_r\boldsymbol{\Gamma}\mathbf{T}_r' \to \boldsymbol{\Lambda}_r^*\boldsymbol{\Xi}_r\boldsymbol{\Lambda}_r^{*\prime}$, 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 that the relabeling changes $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_{\boldsymbol{\Lambda}}$ with invertible Jacobian $\mathbf{J} = \partial\boldsymbol{\theta}/\partial\boldsymbol{\theta}_{\boldsymbol{\Lambda}}$. We just summarize the effect of the relabeling on all of the matrices and vectors in M_r,

$$\begin{aligned} &\boldsymbol{\Delta} \to \boldsymbol{\Lambda}\boldsymbol{\Delta}\mathbf{J}', \qquad \boldsymbol{\Delta}_r \to \boldsymbol{\Lambda}_r^*\boldsymbol{\Delta}_r\mathbf{J}', \qquad \mathbf{C}_r \to (\boldsymbol{\Lambda}_r^{*\prime})^{-1}\mathbf{C}_r(\boldsymbol{\Lambda}_r^*)^{-1}, \\ &\hat{\boldsymbol{\theta}} \to \hat{\boldsymbol{\theta}}_{\boldsymbol{\Lambda}}, \qquad \mathbf{p} - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}) \to \boldsymbol{\Lambda}_r^*[\mathbf{p} - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})]. \end{aligned}$$

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

A.3 Local Alternatives: Expected Value of the Maximum Likelihood Estimator

Consider a parametric family $f(y; \boldsymbol{\theta})$, which can be continuous or discrete, where f is a density relative to measure ν (Lebesgue or counting measure). This section concerns a limit of the expected value of the 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 that the parametrization is of the form $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, where $\boldsymbol{\theta}_2 = \beta\mathbf{1}$. We obtain the MLE based on the submodel and derive its distribution under local alternatives in the full model. That is, the hypotheses are

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

For a sequence of local alternatives, we take $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}', \beta_0\mathbf{1}')'$ as a "true" model, and let $\boldsymbol{\theta}_N = (\boldsymbol{\theta}_{10}', \beta_0\mathbf{1}' + w_N\boldsymbol{\gamma}')'$ be the sequence of alternative parameter values. Here $\boldsymbol{\gamma}$ 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 write the density for H_0 as $f^*(\cdot; \boldsymbol{\theta}^*) = f(\cdot; (\boldsymbol{\theta}_1', \beta\mathbf{1}')')$. Let $\ell(\boldsymbol{\theta}^*; y) = \log f^*(y; \boldsymbol{\theta}^*)$, $\dot{\ell} = \partial\ell/\partial\boldsymbol{\theta}^*$, and $\ddot{\ell} = \partial^2\ell/\partial\boldsymbol{\theta}^*\,\partial\boldsymbol{\theta}^{*\prime}$.

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 $f(\cdot; \boldsymbol{\theta}_N)$. Take an expansion of L about $\boldsymbol{\theta}_0^*$ to get

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

or

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

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, under the usual regularity conditions for asymptotic maximum likelihood,

$$\sqrt{N}\mathrm{E}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*) = [\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0^*)]^{-1}\sqrt{N}\mathrm{E}[\dot{\ell}(\boldsymbol{\theta}_0^*; Y_{1N})] + o_p(1). \tag{A.4}$$

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

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

Finally, if $\sqrt{N}w_N \to \epsilon$, then (A.4) becomes (as $N \to \infty$)

$$\sqrt{N}\mathrm{E}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0^*) \to \epsilon[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0^*)]^{-1}\int \dot{\ell}(\boldsymbol{\theta}_0^*; y)\,\boldsymbol{\gamma}'\frac{\partial f}{\partial\boldsymbol{\theta}_2}\,d\nu(y). \tag{A.5}$$

For a discrete model (ν corresponding to counting measure), write $f(y;\theta) = \pi_y(\theta)$, where y may be a vector (e.g., a binary vector of dimension n). Then (A.5) becomes (19).

[Received February 2004. Revised November 2004.]

REFERENCES

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

Bartholomew, D. J. (1998), "Scaling Unobservable Constructs in the Social Sciences," *Applied Statistics*, 47, 1–13.

Bartholomew, D. J., and Knott, M. (1999), *Latent Variable Models and Factor Analysis* (2nd ed.), London: Arnold.

Bartholomew, D. J., and 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., and Tzamourani, P. (1999), "The Goodness of Fit of Latent Trait Models in Attitude Measurement," *Sociolological Methods and Research*, 27, 525–546.

Beck, A. T., Weissman, A., Lester, D., and Trexler, L. (1974), "The Measurement of Pessimism: The Hopelessness Scale," *Journal of Consulting and Clinical Psychology*, 42, 861–865.

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

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

Browne, M. W. (1984), "Asymptotically Distribution-Free Methods for the Analysis of Covariance Structures," *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.

Chang, E. C., D'Zurilla, T. J., and Maydeu-Olivares, A. (1994), "Assessing the Dimensionality of Optimism and Pessimism Using a Multimeasure Approach," *Cognitive Therapy and Research*, 18, 143–160.

Christoffersson, A. (1975), "Factor Analysis of Dichotomized Variables," *Psychometrika*, 40, 5–32.

Collins, L. M., Fidler, P. L., Wugalter, S. E., and Long, J. (1993), "Goodness-of-Fit Testing for Latent Class Models," *Multivariate Behavioral Research*, 28, 375–389.

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

Finger, M. S. (2002), "A Comparison of Full-Information and Unweighted Least Squares Limited-Information Item Parameter Estimation Methods Used With the Two-Parameter Normal Ogive Model," presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Godambe, V. P. (ed.) (1991), *Estimating Functions*, Oxford, U.K.: Oxford University Press.

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

Jöreskog, K. G. (1994), "On the Estimation of Polychoric Correlations and Their Asymptotic Covariance Matrix," *Psychometrika*, 59, 381–389.

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

Khatri, C. G. (1966), "A Note on a MANOVA Model Applied to Problems in Growth Curve," *Annals of the Institute of Statistical Mathematics*, 18, 75–86.

Koehler, K., and Larntz, K. (1980), "An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials," *Journal of the American Statistical Association*, 75, 336–344.

Lee, S. Y., Poon, W. Y., and 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., and Novick, M. R. (1968), *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

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.

Muthén, B. (1978), Contributions to Factor Analysis of Dichotomous Variables, *Psychometrika*, 43, 551–560.

——— (1984), "A General Structural Equation Model With Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators," *Psychometrika*, 49, 115–132.

——— (1993), "Goodness of Fit With Categorical and Other Nonnormal Variables," in *Testing Structural Equation Models*, eds. K. A. Bollen and J. S. Long, Newbury Park, CA: Sage, pp. 205–234.

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

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

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

Reiser, M. (1996), "Analysis of Residuals for the Multinomial Item Response Model," *Psychometrika*, 61, 509–528.

Reiser, M., and Lin, Y. (1999), "A Goodness-of-Fit Test for the Latent Class Model When Expected Frequencies Are Small," in *Sociological Methodology 1999*, eds. M. Sobel and M. Becker, Boston: Blackwell, pp. 81–111.

Reiser, M., and 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.

Satorra, A. (1989), "Alternative Test Criteria in Covariance Structure Analysis: A Unified Approach," *Psychometrika*, 54, 131–151.

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

Takane, Y., and de Leeuw, J. (1987), "On the Relationship Between Item Response Theory and Factor Analysis of Discretized Variables," *Psychometrika*, 52, 393–408.

Teugels, J. L. (1990), "Some Representations of the Multivariate Bernoulli and Binomial Distributions," *Journal of Multivariate Analysis*, 32, 256–268.

Thissen, D. (1982), "Marginal Maximum Likelihood Estimation for the One-Parameter Logistic Model," *Psychometrika*, 47, 175–186.