An Extended GFfit Statistic Defined on Orthogonal Components of Pearson's Chi-Square

Published Version:
Reiser M., C.S. (2023). An Extended GFfit Statistic Defined on Orthogonal Components of Pearson's Chi-Square. PSYCHOMETRIKA, 88(1 (March)), 208-240 [10.1007/s11336-022-09866-6].

Availability:
This version is available at: https://hdl.handle.net/11585/880293 since: 2023-03-06

Published:

DOI: http://doi.org/10.1007/s11336-022-09866-6

AN EXTENDED GFFIT STATISTIC DEFINED ON ORTHOGONAL COMPONENTS OF PEARSON'S CHI-SQUARE

MARK REISER

SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES, ARIZONA STATE UNIVERSITY, USA

SILVIA CAGNONE

DEPARTMENT OF STATISTICAL SCIENCES, UNIVERSITY OF BOLOGNA, ITALY

JUNFEI ZHU

SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES, ARIZONA STATE UNIVERSITY, USA

First author contact: SoMSS, Arizona State University, Tempe, Arizona 85287. E-Mail: mark.reiser@asu.edu

AN EXTENDED GFFIT STATISTIC DEFINED ON ORTHOGONAL COMPONENTS OF PEARSON'S CHI-SQUARE

Abstract

The Pearson and likelihood ratio statistics are commonly used to test goodness of fit for models applied to data from a multinomial distribution. The goodness-of-fit test based on Pearson's chi-squared statistic is sometimes considered to be a global test that gives little guidance to the source of poor fit when the null hypothesis is rejected, and it has also been recognized that the global test can often be outperformed in terms of power by focused or directional tests. For the cross-classification of a large number of manifest variables, the *GFfit* statistic focused on second-order marginals for variable pairs i, j has been proposed as a diagnostic to aid in finding the source of lack of fit after the model has been rejected based on a more global test. When data are from a table formed by the cross-classification of a large number of variables, the common global statistics may also have low power and inaccurate Type I error level due to sparseness in the cells of the table. The sparseness problem is rarely encountered with the $GFfit$ statistic because it is focused on the lower-order marginals. In this paper, a new and extended version of the *GFfit* statistic is proposed by decomposing the Pearson statistic from the full table into orthogonal components defined on marginal distributions and then defining the new version, $GFfit_{\perp}^{(ij)}$, as a partial sum of these orthogonal components. While the emphasis is on lower-order marginals, the new version of $GFfit_{\perp}^{(ij)}$ is also extended to higher-order tables so that the $GFfit_{\perp}$ statistics sum to the Pearson statistic. As orthogonal components of the Pearson X^2 statistic, the $GFfit_{\perp}^{(ij)}$ have advantages over other lack-of-fit diagnostics that are currently available for cross-classified tables: the $GFfit_{\perp}^{(ij)}$ have higher power to detect lack of fit while maintaining good Type I error control even if the joint frequencies are very sparse, as will be shown in simulation results; theoretical results will establish that the $GFfit_{\perp}^{(ij)}$ have known degrees of freedom and are asymptotically independent statistics with known joint distribution, a property which facilitates less conservative control of false discovery rate (FDR) or familywise error rate (FWER) in a high-dimensional table which would produce a large number of bivariate lack-of-fit diagnostics. $GFfit_{\perp}^{(ij)}$ are also computationally stable. The extended $GFfit_{\perp}^{(ij)}$ statistic can be applied to a variety of models for cross-classified tables. An application of the new *GFfit* statistic as a diagnostic for a latent variable model is presented.

Key Words: multivariate discrete distribution, overlapping cells, orthogonal components, composite null hypothesis

Contents

1. Introduction

The fit of a multinomial model is often assessed by testing the composite null hypothesis $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$, where $\boldsymbol{\pi}$ is a T-dimensional vector of multinomial probabilities, and $\boldsymbol{\pi}(\boldsymbol{\beta})$ is a vector of the multinomial probabilities as a function of unique model parameters in the g-dimensional vector $\boldsymbol{\beta}$. For q variables, each with c categories, $T = c^q$. When the model parameters $\boldsymbol{\beta}$ are unknown and estimated, the null hypothesis $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$ is often tested with the Pearson-Fisher statistic:

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

where

$$z_s = \left(\pi_s(\hat{\boldsymbol{\beta}})\right)^{-\frac{1}{2}} \left(\hat{\mathrm{p}}_s - \pi_s(\hat{\boldsymbol{\beta}})\right), \ s = 1, \dots, c^q, \tag{1.2}$$

and where

$$\begin{aligned}
\hat{\mathrm{p}}_s &= \frac{n_s}{n} \text{ is element } s \text{ of } \hat{\mathbf{p}}, \text{ the vector of multinomial proportions,}\\
n_s &= \text{element } s \text{ of } \mathbf{n}, \text{ the vector of observed frequencies,}\\
n &= \text{total sample size} = \sum_{s=1}^{T} n_s,\\
\hat{\boldsymbol{\beta}} &= \text{parameter estimator vector,}\\
\pi_s(\boldsymbol{\beta}) &= \text{the expected proportion for cell } s, \text{and}\\
\pi_s(\hat{\boldsymbol{\beta}}) &= \text{estimated expected proportion for cell } s.
\end{aligned}$$

The goodness-of-fit test based on Pearson's chi-squared statistic is sometimes considered to be a global test that gives little guidance to the source of poor fit when the null hypothesis is rejected, and it has also been recognized that the global test can often be outperformed in terms of power by focused or directional tests. For a multidimensional contingency table, Joreskog and Moustaki (2001) proposed the following *GFfit* statistic focused on second-order marginals for variable pairs i, j as a diagnostic to aid in finding the source of model lack of fit after the model has been rejected based on a more global test:

$$n \sum_{ab} \frac{(\hat{p}^{(ij)}_{ab} - \hat{\pi}^{(ij)}_{ab})^2}{\hat{\pi}^{(ij)}_{ab}} \tag{1.3}$$

where $i = 1, \dots, q-1$; $j = i+1, \dots, q$; $a = 1, \dots, c$; $b = 1, \dots, c$; $\hat{p}^{(ij)}_{ab}$ is the observed proportion for cell a, b in the i, j marginal table; and $\hat{\pi}^{(ij)}_{ab}$ is the expected or fitted proportion for cell a, b in the i, j marginal table. Cagnone and Mignani (2007) worked with a somewhat different formulation of the $GFfit^{(ij)}$ statistic, which is reviewed in Section 4, and obtained the asymptotic distribution of the statistic when applied to ordinal variables. To avoid possible confusion, the 2001 version in expression 1.3 will be referred to as X^2_{ij} in the remainder of this paper. If a $GFfit$ statistic is too large, it suggests that the association between variables i and j is not well fit by the model. The focus on second-order marginals for cross-classified tables is supported by the findings of Salomaa (1990).

In this paper, a new and extended version of the *GFfit* statistic is proposed and validated. Our approach, in the tradition of Lancaster (1969), Miralev (1987), and Rayner and Best (1989), decomposes the Pearson statistic from a full cross-classified table, where variables have $c \geq 2$ categories, into orthogonal components. The new version of the statistic, $GFfit_{\perp}^{(ij)}$, is defined as a partial sum of orthogonal components obtained on marginal distributions of the cross-classified table. The new version is also extended to higher-order tables, and $GFfit_{\perp}$ statistics are shown to sum to the Pearson statistic. As orthogonal components of the Pearson X^2 statistic, the $GFfit_{\perp}^{(ij)}$ have advantages over other lack-of-fit diagnostics that are currently available for cross-classified tables: the $GFfit_{\perp}^{(ij)}$ have higher power to detect lack of fit while maintaining good Type I error control even if the joint frequencies are very sparse, as will be shown in simulation results; theoretical results will establish that the $GFfit_{\perp}^{(ij)}$ have known degrees of freedom and are asymptotically independent statistics with known joint distribution, a property which facilitates less conservative control of false discovery rate (FDR) or familywise error rate (FWER) in a high-dimensional table which would produce a large number of bivariate lack-of-fit diagnostics. $GFfit_{\perp}^{(ij)}$ are also computationally stable.

When data are a table of counts formed by the cross-classification of a large number of variables, Pearson's chi-square and the likelihood ratio statistic may have low power and inaccurate Type I error level due to sparseness (Koehler, 1986; Koehler & Larntz, 1980; Agresti & Yang, 1987). In order to overcome the effects of sparse frequencies, several omnibus statistics have been proposed that focus on lower-order marginal distributions of the joint variables rather than the full joint distribution. Reiser (1996, 2008) developed omnibus test statistics focused on lower-order marginals for models fit to cross-classified binary variables by using orthogonal components of the Pearson-Fisher statistic. We also extend Reiser's test statistics for binary variables to the case of cross-classified tables with $c \geq 2$ categories, and we show that the $GFfit_{\perp}^{(ij)}$ statistics are orthogonal components of these omnibus statistics on lower-order marginals. While a test for a model based on lower-order marginals is more focused than X^2_{PF}, it is still omnibus in the sense that an entire set of lower-order marginals is included, and there still may be a lack of guidance to the source of poor fit when the null hypothesis is rejected. The $GFfit_{\perp}$ statistics may be employed as a diagnostic on bivariate tables after a model is rejected by a test using the global X^2_{PF} or after if it is rejected by an omnibus test on lower-order marginals. The sparseness problem is rarely encountered with the new $GFfit_{\perp}^{(ij)}$ statistic itself because it is also focused on the lower-order marginals.

The new version of *GFfit* is a member of a class of lack-of-fit diagnostics derived from an underlying multinomial distribution that can be applied to a variety of models for multidimensional contingency tables, including item response models, latent class models, log-linear models, and longitudinal models. Other members of this class of diagnostics include M_{ij} (Maydue-Olivares and Joe, 2006), $\bar{\bar{X}}^2_{ij}$ (Muthén & Asparouhov (2010), and X^2_{ij}, the Pearson X^2 statistic applied to bivariate marginal tables. X^2_{ij} is the same statistic as the original $GFfit^{(ij)}$. Liu & Maydeu-Olivares (2014) advocate the need for a goodness-of-fit diagnostic to be well approximated by a known reference distribution when the fitted model is correctly specified. $GFfit^{(ij)}_{\perp}$, M_{ij}, and $\bar{\bar{X}}^2_{ij}$ have known asymptotic distributions, although X^2_{ij} does not. In addition, the joint distribution of a set of diagnostics should be considered for the purpose of maintaining Type I error level when assessing a set of diagnostics for a model. While we will show that $GFfit^{(ij)}_{\perp}$ has known joint distribution function, the joint distribution of $\bar{\bar{X}}^2_{ij}$ is unknown, and the joint distribution of M_{ij} has not been given.

We compare the performance of $GFfit^{(ij)}_{\perp}$ to these other three diagnostics in terms of Type I error and power. Results for both asymptotic power and simulations are presented, and an application of the new *GFfit* statistic as a diagnostic for a latent variable model is also presented. Orthogonal components of Pearson's statistic have a sequential nature. In the application we demonstrate that occasionally it may be possible to use substantive theory to select an order that will moderately increase power for the first few $GFfit^{(ij)}_{\perp}$, but power results using theory and simulations show that an arbitrary order has modest, if any, effect on the power of $GFfit^{(ij)}_{\perp}$ to detect lack of fit when compared to other diagnostics that are not order dependent. Furthermore, we will demonstrate that the correction for multiple testing to maintain Type I error level is much more conservative for the diagnostics that have unknown joint distribution function compared to the adjustment that can be used for $GFfit^{(ij)}_{\perp}$.

2. Linear Combinations of Joint Frequencies

A traditional goodness-of-fit approach for a multinomial model fit to a cross-classified table uses the joint frequencies to calculate a test statistic. This section presents a transformation from joint proportions or frequencies to marginal proportions which are used to develop statistics on lower-order marginals and the new version of *GFfit* which is presented in Section 4.1.

2.1. Marginal Probabilities

Marginal proportions for a contingency table can be obtained by a linear combination of the joint proportions. The relationship can be shown by using zeros and 1's to code the levels of categorical response variables, $Y_i, i = 1, 2, \ldots, q$, where $q \geq 2$ and Y_i has $c \geq 2$ response categories. A $q(c-1)$-dimensional vector of zeros and 1's, can indicate a specific cell from the contingency table formed by the cross-classification of q response variables. Then a $T = c^q$-dimensional set of response patterns can be generated by varying the levels of the q^{th} variable most rapidly, the $q^{th} - 1$ variable next, etc. Define $\boldsymbol{V}$ as the T by $q(c-1)$ matrix with response patterns as rows.
For $q = 3$ and $c = 2$, $\boldsymbol{V}$ is familiar:

$$\boldsymbol{V} = \begin{pmatrix} 0\,0\,0 \\ 0\,0\,1 \\ 0\,1\,0 \\ 0\,1\,1 \\ 1\,0\,0 \\ 1\,0\,1 \\ 1\,1\,0 \\ 1\,1\,1 \end{pmatrix} . \tag{2.1}$$

For $q = 3$ and $c = 3$,

$$\boldsymbol{V}_{27 \text{ x } 6} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} . \tag{2.2}$$

Although the order of the columns of $\boldsymbol{V}$ is arbitrary, the ordering described above is a natural ordering that will produce rows such that the binary numbers represented by the patterns are in ascending order. More detail on generating the matrix $\boldsymbol{V}$ is given in Appendix A.

Formal definitions of first- and second-order marginal proportions are given in Section 9.1 of Appendix A. The summation across the frequencies associated with the response patterns to obtain the marginal proportions represents a linear transformation of the frequencies in the multinomial vector $\boldsymbol{\pi}$ which can be implemented via multiplication by a certain incidence matrix, denoted here generically by the symbol $\mathbf{H}$. The symbol $\mathbf{H}_{[t]}$ denotes the transformation matrix that would produce marginals of order t. The symbol $\mathbf{H}_{[t:u]}$, $t \leq u \leq q$, denotes the transformation matrix that would produce marginals from order t up to and including order u.

Furthermore, $\mathbf{H}_{[t]} \equiv \mathbf{H}_{[t:t]}$, and $\mathbf{H} \equiv \mathbf{H}_{[t:u]}$. More detail on generating the matrix $\mathbf{H}$ is given in Appendix A. For first-order marginal proportions, $\mathbf{H}_{[1]} = \boldsymbol{V}'$.

For higher-order marginal proportions, the rows of $\mathbf{H}$ can be obtained as Hadamard products among the columns of $\boldsymbol{V}$, as shown in Section 9.3 of Appendix A. The second-order marginal proportions for variables Y_i and Y_j can be obtained by employing the matrix $\mathbf{H}_{[2]}$, and the third-order marginal proportions for variables Y_i, Y_j, and Y_k can be obtained by employing the matrix $\mathbf{H}_{[3]}$. Then, for example,

$$\mathbf{H}_{[1:3]} = \begin{pmatrix} \mathbf{H}_{[1]} \\ \cdots \\ \mathbf{H}_{[2]} \\ \cdots \\ \mathbf{H}_{[3]} \end{pmatrix}. \tag{2.3}$$

A general matrix $\mathbf{H}_{[t:u]}$ to obtain marginals of any order can be defined in a similar fashion by using Hadamard products among the columns of $\boldsymbol{V}$. $\mathbf{H}_{[1:q]}$ gives a mapping from the c^q joint proportions to the set of $(c^q - 1)$ linearly independent marginal proportions:

$$\boldsymbol{\Pi} = \mathbf{H}_{[1:q]}\boldsymbol{\pi}, \tag{2.4}$$

where

$$\begin{aligned}\boldsymbol{\Pi} = (\pi^{(1)}(2),\ \pi^{(1)}(3), \ldots, \pi^{(1)}(c),\ \pi^{(2)}(2),\ \pi^{(2)}(3), \ldots, \pi^{(2)}(c), \ldots, \pi^{(q)}(c),\\ \pi^{(12)}(2,2), \pi^{(12)}(2,3), \ldots \pi^{(q-1,q)}(c,c),\ \pi^{(1,1,2)}(2,2,2), \ldots, \pi^{(q-2,q-1,q)}(c,c,c)\\ \ldots \pi^{(1,2,3\ldots q)}(c,c,c,\ldots c))'\end{aligned} \tag{2.5}$$

is the vector of linearly independent marginal proportions (Bartholomew, 1987). As constructed, the first column of $\mathbf{H}_{[1:q]}$ is a column of 0's, and can be omitted along with the first element of $\boldsymbol{\pi}$. So define $\ddot{\mathbf{H}}_{[1:q]} = (\mathbf{h}_2\ \mathbf{h}_3\ \cdots \mathbf{h}_{c^q})$ and $\ddot{\boldsymbol{\pi}} = (\pi_2,\ \pi_3, \cdots, \pi_{c^q})$, where $\mathbf{h}_f$ is column $f,\ f = 2, \cdots, c^q$, from $\mathbf{H}_{[1:q]}$, and π_f is element f of $\boldsymbol{\pi}$. $\ddot{\mathbf{H}}_{[1:q]}$ is now a $c^q - 1$ by $c^q - 1$ full rank matrix. The location of the column of 0's in $\mathbf{H}_{[1:q]}$ is arbitrary, but it appears in column 1 under the ordering defined above. Then

$$\boldsymbol{\Pi} = \ddot{\mathbf{H}}_{[1:q]}\ddot{\boldsymbol{\pi}} \tag{2.6}$$

is equivalent to expression 2.4, and the transformation from joint proportions to marginal proportions can be seen to be a one-to-one transformation from $\Re^{c^q-1} \longrightarrow \Re^{c^q-1}$. Throughout this paper, $\mathbf{H}_{[1:q]}\boldsymbol{\pi}$ could be replaced by $\ddot{\mathbf{H}}_{[1:q]}\ddot{\boldsymbol{\pi}}$.

3. Orthogonal Components of X^2_{PF}

Orthogonal components of X^2_{PF} defined on marginal proportions were presented by Reiser (2008) for $q \geq 3$ categorical variables with $c = 2$ categories. In this section, we extend those results to the case where $c \geq 2$ categories. This extension of orthogonal components to variables with multiple categories is straightforward.

3.1. Equivalence of X^2_{PF} and a Quadratic Form on Marginals

Define the unstandardized residual $\mathrm{r}_s = \hat{\mathrm{p}}_s - \pi_s(\hat{\boldsymbol{\beta}})$, and denote the T-dimensional vector of unstandardized residuals as $\mathbf{r}$ with element r_s. A vector of simple residuals for marginals of any order may be defined such that $\boldsymbol{e} = \mathbf{H}(\hat{\mathrm{p}} - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})) = \mathbf{H}\mathbf{r}$. Let X^2_{T-g-1} be a quadratic form statistic defined on marginal proportions. As mentioned earlier, g is the number of model parameters to be estimated. This section gives the conditions under which X^2_{T-g-1} is equivalent to X^2_{PF}. Define the two quadratic form statistics as follows:

$$X^2_{PF} = n\mathbf{r}'D(\boldsymbol{\pi}(\hat{\boldsymbol{\beta}}))^{-1}\mathbf{r}, \text{ and} \tag{3.1}$$

$$X^2_{T-g-1} = n\mathbf{r}'\mathbf{H}'\widehat{\boldsymbol{\Omega}}_{\mathrm{e}}^{-1}\mathbf{H}\mathbf{r} = \boldsymbol{e}'\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}^{-1}\boldsymbol{e} \tag{3.2}$$

where,

$$n^{\frac{1}{2}}\boldsymbol{r} \xrightarrow{d} MVN(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{r}}) \tag{3.3}$$

$$\boldsymbol{\Omega}_{\boldsymbol{r}} = \ (D(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{G}') \tag{3.4}$$

$$\boldsymbol{\Omega}_{\boldsymbol{e}} = \mathbf{H}\boldsymbol{\Omega}_{\boldsymbol{r}}\mathbf{H}' \tag{3.5}$$

$$D(\boldsymbol{\pi}) = \text{diagonal matrix with } (s, s) \text{ element equal to } \pi_s(\boldsymbol{\beta}) \tag{3.6}$$

$$\mathbf{A} = D(\boldsymbol{\pi})^{-1/2}\frac{\partial\boldsymbol{\pi}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}, \ \mathbf{G} = \ \frac{\partial\boldsymbol{\pi}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}, \tag{3.7}$$

$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}} = n^{-1}\widehat{\boldsymbol{\Omega}}_{\mathrm{e}}$, and $\widehat{\boldsymbol{\Omega}}_{\mathrm{e}} = \boldsymbol{\Omega}_{\mathrm{e}}$ evaluated at $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$. See Haberman (1993) for (3.3) and (3.4). $\mathbf{H} = \mathbf{H}_{T-g-1}$, is a $T - g - 1$ by c^q partition of $\mathbf{H}_{[1:q]}$, as defined in Section 2, with g rows deleted in order to render $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$ full rank by accounting for estimated parameters of the model $\boldsymbol{\pi}(\boldsymbol{\beta})$. For some models, any g rows can be deleted. For other models, such as a hierarchical log-linear model where some marginals are exactly fit, the g rows to be deleted are determined in whole or in part by features of the model. Alternatively, it is possible to define and calculate X^2_{T-g-1} without deleting rows from $\mathbf{H}_{[1:q]}$, but then a generalized inverse would be needed for $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$. More details on deleting rows are given in Section 4.1. We use the expected information matrix $(\mathbf{A}'\mathbf{A})$ because we are interested in components of the X^2_{PF} statistic.

Consider the full rank matrix $\mathbf{H} = \mathbf{H}_{T-g-1}$, as defined above. Then define $\mathbf{H}^* = \boldsymbol{F}'\mathbf{H}$, where $\boldsymbol{F}$ is the upper triangular matrix such that $\boldsymbol{F}'\boldsymbol{\Omega}_{\boldsymbol{e}}\boldsymbol{F} = \boldsymbol{I}$. $\boldsymbol{F} = (\boldsymbol{C}')^{-1}$, where $\boldsymbol{C}$ is the

Cholesky factor of $\boldsymbol{\Omega_e}$. Premultiplication by $(\boldsymbol{C}')^{-1}$ orthonormalises the matrix $\mathbf{H}$ relative to the matrix $(D(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{G}')$. $\mathbf{H}^*$ has full row rank.

Then from Reiser (2008),

$$X^2_{T-g-1} = n\mathbf{r}'\mathbf{H}'\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{-1}\widehat{\boldsymbol{\Omega}}_{\mathbf{e}}^{-1}(\widehat{\mathbf{F}}')^{-1}\widehat{\mathbf{F}}'\mathbf{H}\mathbf{r} = n\mathbf{r}'\mathbf{H}'\widehat{\mathbf{F}}(\widehat{\mathbf{F}}'\widehat{\boldsymbol{\Omega}}_{\mathbf{e}}\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}'\mathbf{H}\mathbf{r} = n\mathbf{r}'(\widehat{\mathbf{H}}^*)'\widehat{\mathbf{H}}^*\mathbf{r}, \qquad (3.8)$$

where $\widehat{\mathbf{H}}^* = \mathbf{H}^*(\hat{\boldsymbol{\beta}})$. Reiser (2008) proves that $X^2_{PF} = X^2_{T-g-1}$ if $\mathbf{H}$ is a $T-g-1$ by T matrix with rank= $(T-g-1)$ for the case $T = 2^q$. $\mathbf{H}_{[1:q]}$ with g rows removed as specified above satisfies this rank condition. For extending results on components to the case $c \geq 2$, the proof for $X^2_{PF} = X^2_{T-g-1}$ remains valid when $T = c^q$. If $\mathbf{H} = \mathbf{H}_{[T-g-1]}$, then the null hypotheses $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$ and $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$ are equivalent, and $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$ becomes $H_o\colon \boldsymbol{\Pi} = \boldsymbol{\Pi}(\boldsymbol{\beta})$, where $\boldsymbol{\Pi}$ is the vector of marginal proportions as defined in Section 2. If g rows are not removed from $\mathbf{H}_{[1:q]}$ and a generalized inverse is used for $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$, then g of the elements of $\widehat{\mathbf{H}}^*\mathbf{r}$ will be identically equal to zero.

3.2. Orthogonal Components

Define

$$\hat{\boldsymbol{\gamma}} = n^{-\frac{1}{2}}\widehat{\boldsymbol{F}}'\mathbf{H}\mathbf{r} = n^{-\frac{1}{2}}\widehat{\mathbf{H}}^*\mathbf{r} \qquad (3.9)$$

where $\widehat{\boldsymbol{F}}$ is the matrix $\boldsymbol{F}$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. Then from expression 3.8,

$$X^2_{T-g-1} = \hat{\boldsymbol{\gamma}}'\hat{\boldsymbol{\gamma}} = \sum_{\ell=1}^{j=T-g-1} \hat{\gamma}^2_\ell. \qquad (3.10)$$

From properties of the Cholesky factor, $\hat{\gamma}^2_\ell$, $\ell = 1, \ldots, (T-g-1)$, are orthogonal components of X^2_{PF} when $c \geq 2$, and they have a sequential property. Orthonormalization of goodness-of-fit diagnostics using the Cholesky factor has been considered for other statistical models. Houseman et al. (2004), Jacqmin-Gadda et al. (2007), and Verbeke and Molenberghs (2009) have proposed and studied the *Cholesky residual.* $\widehat{\mathbf{H}}^*\mathbf{r}$ is essentially a vector of Cholesky residuals on the marginals.

Under the regularity conditions given by Birch (1964), and assuming the model $\boldsymbol{\pi}(\boldsymbol{\beta})$ is identified, $n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}} \xrightarrow{P} \boldsymbol{\Omega_e}$ and $\boldsymbol{e} \xrightarrow{d} MVN(\boldsymbol{\xi}, \boldsymbol{\Sigma_e})$, where $\boldsymbol{\xi} = \mathbf{H}(\boldsymbol{\pi} - \boldsymbol{\pi}(\boldsymbol{\beta}))$. Then, since $\boldsymbol{e}$ is a linear combination of the elements of $\mathbf{r}$, $\widehat{\mathbf{H}}^*\mathbf{r}$ has asymptotic covariance matrix $\boldsymbol{F}'\boldsymbol{\Omega_e}\boldsymbol{F} = \boldsymbol{I}_{T-g-1}$, and the elements $\hat{\gamma}^2_\ell \xrightarrow{d} \chi^2_1$ and are asymptotically independent random variables. Consequently, partial sums of of the $\hat{\gamma}^2_\ell$ have an asymptotic chi-square distribution with degrees of freedom equal to the number of non-zero components in the sum. However, as with X^2_{PF}, the asymptotic distribution for $\hat{\gamma}^2_\ell$ may fail if the marginal frequencies used to calculate $\hat{\gamma}^2_\ell$ are sparse.

$n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$ is estimated from the fitted joint proportions. In the sparseness case, which may be present if q is large relative to n, an important result from Simonoff (1986) is applicable to the distribution of $\hat{\gamma}^2_\ell$ on lower-order marginals (Reiser, 2019). Simonoff (1986) defines the sparse asymptotic framework as $n, K \to \infty$, with $0 < \nu_1 < n/K < \nu_2 < \infty$, and $\exists M \in (1, \infty)$ such that $0 < (MK)^{-1} < \pi_s < M/K < 1 \;\forall\; s$. Assuming $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + O_p(n^{-\frac{1}{2}})$; if $\boldsymbol{\pi}(\boldsymbol{\beta})$ has bounded second partial derivatives with respect to $\boldsymbol{\beta}$, $sup_s \left|\pi_s(\hat{\boldsymbol{\beta}})/\pi_s - 1\right| = O_p(n^{-\frac{1}{2}})$. So, even under sparseness conditions for the joint frequencies, $\pi_s(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \pi_s$, $\boldsymbol{\pi}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{\pi}$, $n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}} \xrightarrow{P} \boldsymbol{\Omega}_{\boldsymbol{e}}$, as required for the theoretical result on the asymptotic distribution. Then in the case of sparse joint frequencies, the asymptotic distribution of $\hat{\gamma}^2_\ell$ from lower-order marginals that are not sparse is still χ^2_1. The result is relevant because with a simpler estimator for $\boldsymbol{\Sigma}_{\boldsymbol{e}}$ using observed proportions, as in Christoffersson (1975), the asymptotic distribution for a statistic on lower-order marginals will fail in the sparseness condition (Reiser and Vandenberg, 1994).

A computational method from Reiser (2008) produces very reliable results by obtaining orthogonal components as the sequential sum of squares from a weighted orthogonal regression. The method is applicable when $c \geq 2$ and is used to calculate components for the simulations in Section 6 and the application in Section 7. The regression coefficients themselves have no meaning, and the regression is simply a method to obtain orthogonal components, an alternative to applying the Cholesky factor or Gram-Schmidt orthogonalization.

3.3. A Lower-Order Omnibus Test Statistic Based Orthogonal Components of X^2_{PF}

Given the traditional null hypothesis H_o: $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$, linear combinations of $\boldsymbol{\pi}$ may be tested under the null hypothesis H_o: $\mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$, where $\mathbf{H}$ contains coefficients for the linear combinations and may specify combinations that form marginal proportions as defined in Section 2. If $\mathbf{H}$ has rank less than $R = T - g - 1$, then H_o: $\mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$ might specify a test that is more focused such as test on lower-order marginals. For example, the null hypothesis H_o: $\mathbf{H}_{[2]}\boldsymbol{\pi} = \mathbf{H}_{[2]}\boldsymbol{\pi}(\boldsymbol{\beta})$ would specify a test that is focused on the second-order marginals but still omnibus in the sense that it would incorporate all of the second-order marginals. Such a focused test may be used for the purpose of higher power, and it also may be used for a test on lower-order marginals when joint frequencies are sparse because the asymptotic chi-square approximation for the test statistic is still valid for a test on lower-order marginals that are not sparse.

Consider the test statistic $X^2_{[t:u]}$ from Reiser (2008), but now for $c \geq 2$ categories, an extension that only requires incorporating an $\mathbf{H}$ matrix for $c \geq 2$ categories as defined in Section 2. $X^2_{[t:u]}$ is a test statistic that can be computed from the partial sum of orthogonal components for marginals from order t to order u. For example, if $\mathbf{H} = \mathbf{H}_{[2]}$,

$$X^2_{[2]} = \sum_{\ell=1}^{\ell=\binom{q}{2}(c-1)^2} \hat{\gamma}^2_\ell \tag{3.11}$$

is sum of the components for second-order marginals. The null hypothesis H_o: $\mathbf{H}_{[2]}\boldsymbol{\pi} = \mathbf{H}_{[2]}\boldsymbol{\pi}(\boldsymbol{\beta})$ would be tested using the statistic $X^2_{[2]}$. Section 3.2 discusses deleting g rows from $\mathbf{H}_{[1:q]}$ in order to show that X^2_{PF} is equivalent to X^2_{T-g-1}. Deleting rows of $\mathbf{H}_{[1:q]}$ is not relevant for $X^2_{[2]}$ because the definition uses only the $\mathbf{H}_{[2]}$ partition of $\mathbf{H}_{[1:q]}$, which means that more than g rows have already been deleted from $\mathbf{H}_{[1:q]}$. No additional rows are deleted unless a bivariate table happens to be exactly fit by the model, which has an extremely low probability of occurring by chance. If a bivariate table is exactly fit and appropriate rows are not deleted, then some components will be identically equal to zero.

Under the null hypothesis stated above, an asymptotic central chi-square distribution for $X^2_{[t:u]}$ follows from results in Section 3.2. Degrees of freedom are determined from the number of nonzero components in the partial sum that produces the statistic. Components identically equal to zero may be produced by restrictions on the marginals included in $X^2_{[t:u]}$ or linear dependencies of rows in $\mathbf{H}_{[t:u]}$ on the model matrix for $\boldsymbol{\pi}(\boldsymbol{\beta})$, both of which would be known from theory. Degrees of freedom, then, are known from theory and do not need to be estimated. More details on the degrees of freedom are given in Reiser (2008). In Section 4.1, we will show that $X^2_{[t:u]}$ statistics are sums of certain $GFfit_{\perp}$ statistics. Other test statistics on lower-order marginals have been developed by Bartholomew & Leung (2002), Tollenaar and Mooijaart (2003), and Maydeu-Olivares and Joe (2005, 2006), but these statistics are not components of the Pearson-Fisher statistic. Statistical tests using lower-order marginals have very good performance for Type I error rate when the full data table is sparse.

Although the topic of this paper is the $GFfit$ statistic, a small Monte Carlo simulation study for $X^2_{[2]}$ using the generalized linear latent variable model (GLLVM) is presented in Appendix B because $X^2_{[2]}$ calculated from orthogonal components applied to ordinal variables has not been previously studied. Results show that $X^2_{[2]}$ has Type I error at the nominal level when applied to multi-category variables even when joint frequencies are sparse and has high power to reject a false null hypothesis when lack of fit is in the second-order marginals.

$X^2_{[t:u]}$ is essentially a version of the score statistic from Rayner and Best (1989). Glas (1988) developed omnibus test statistics specific to the Rasch model based on the underlying multinomial model for the response patterns. His R_2 statistic focused on second-order marginals, and Maydeu-Olivares and Montano (2013) compared R_2 to the M_2 statistic from Maydeu-Olivares and Joe (2006).

4. *GFfit*$^{(ij)}$

Joreskog and Moustaki (2001) proposed the *GFfit* statistic as a lack-of-fit diagnostic to be employed after a global goodness-of-fit test indicates that a proposed model does not fit. Joreskog and Moustaki had X^2_{PF} in mind as the global test statistic, but a lack-of-fit diagnostic might also be useful if the model does not fit by a lower-order omnibus statistic such as $X^2_{[2]}$. Cagnone and Mignani (2007) defined $GFfit^{(ij)}$ as a special case of $X^2_{[t:u]}$, which is a somewhat different formulation than Joreskog and Moustaki (2001) because it incorporates the covariance matrix. Instead of $\mathbf{H}_{[2]}$, Cagnone and Mignani define $GFfit^{ij}$ using a $c^2q(q-1)/2$ by T matrix $\boldsymbol{M}_{[2]}$. Details of generating $\boldsymbol{M}_{[2]}$ are given in Appendix A.

Then using $\boldsymbol{M}_{[2]}$, the Cagnone and Mignani version of $GFfit^{(ij)}$ is

$$GFfit^{(ij)} = \boldsymbol{e}'(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}^{(ij)})^{+}\boldsymbol{e} \tag{4.1}$$

where $\boldsymbol{B}^{+}$ indicates the Moore-Penrose generalized inverse of matrix $\boldsymbol{B}$, and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}^{(ij)} = n^{-1}\boldsymbol{\Omega}_{\boldsymbol{e}}^{ij}$ with $\boldsymbol{\Omega}_{\boldsymbol{e}}^{(ij)}$ evaluated at the MLE $\hat{\boldsymbol{\beta}}$, and now

$$\boldsymbol{\Omega}_{\boldsymbol{e}}^{(ij)} = \boldsymbol{M}_{[2]}^{(ij)}(D(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' - \mathbf{G}(\mathrm{A}'\mathrm{A})^{-1}\mathbf{G}')(\boldsymbol{M}_{[2]}^{(ij)})', \tag{4.2}$$

where $\mathrm{M}_{[2]}^{(ij)}$ is a partition of the general matrix $\boldsymbol{M}_{[2]}$ such that

$$\mathrm{M}_{[2]}^{(ij)} = \begin{pmatrix} \boldsymbol{m}'_{d+1} \\ \boldsymbol{m}'_{d+2} \\ \vdots \\ \boldsymbol{m}'_{d+c^2} \end{pmatrix}_{[2]} \tag{4.3}$$

where $d = (q(i-1) - i(i-1)/2)c^2$.

Linear dependencies exist among the rows of $\boldsymbol{M}_{[2]}$; $\mathbf{H}_{[2]}$ consists of the linearly independent rows of $\boldsymbol{M}_{[2]}$ such that $\mathbf{H}_{[2]} = \boldsymbol{A}\boldsymbol{M}_{[2]}$, where matrix $\boldsymbol{A}$ is given in Appendix A.

4.1. Extended $GFfit_{\perp}^{(ij)}$ Statistic Using Orthogonal Components

Returning to the full row-rank matrix $\mathbf{H}$, the new version $GFfit_{\perp}^{(ij)}$ can be defined using the $\mathbf{H}_{[2]}$ partition of $\mathbf{H}$. Define

$$\mathbf{H}_{[2]}^{(ij)} = \begin{pmatrix} \boldsymbol{h}'_{d+1} \\ \boldsymbol{h}'_{d+2} \\ \vdots \\ \boldsymbol{h}'_{d+(c-1)^2} \end{pmatrix}_{[2]} \tag{4.4}$$

where now $d = (q(i-1) - i(i-1)/2)(c-1)^2$, and now define an orthogonal components version of $GFfit^{(ij)}$ as the sum of second-order components pertaining to variable i and variable j:

$$GFfit_{\perp}^{(ij)} = \sum_{\ell=d+1}^{\ell=d+(c-1)^2} \hat{\gamma}_{\ell}^{2}. \tag{4.5}$$

Since the $\hat{\gamma}_\ell^2$ are asymptotically independent χ_1^2, $GFfit_\perp^{(ij)}$ are asymptotically distributed chi-square on $(c-1)^2$ degrees of freedom, assuming that no marginal in the $c \times c$ table is exactly fit by the model. $GFfit_\perp^{(ij)}$ can be defined and calculated directly without decomposing X_{PF}^2 into $T-g-1$ components:

$$GFfit_\perp^{(ij)} = n^{-1}\mathbf{r}'\widehat{\mathbf{H}}_{[2]}^{'*(ij)}\widehat{\mathbf{H}}_{[2]}^{*(ij)}\mathbf{r}, \tag{4.6}$$

where $\mathbf{H}_{[2]}^*$ is calculated from $\boldsymbol{\Omega}_{\boldsymbol{e}[2]} = \mathbf{H}_{[2]}\boldsymbol{\Omega}_{\boldsymbol{r}}\mathbf{H}_{[2]}'$. So, deleting rows of $\mathbf{H}_{[1:q]}$ is not an issue for $GFfit_\perp^{(ij)}$ because the definition uses only $\mathbf{H}_{[2]}$, which means that more than g rows have already been deleted from $\mathbf{H}_{[1:q]}$. No additional rows are deleted unless a bivariate table happens to be exactly fit by the model.

$X_{[t:u]}^2$ statistics from Section 3.3 are sums of $GFfit_\perp$ statistics. For second-order marginals,

$$X_{[2]}^2 = \sum_{\ell=1}^{\ell=\binom{q}{2}(c-1)^2} \hat{\gamma}_\ell^2 = \sum_{i=1}^{i=q-1}\sum_{j=i+1}^{j=q} GFfit_\perp^{(ij)}, \tag{4.7}$$

and the $GFfit_\perp^{(ij)}$ statistics are orthogonal components of $X_{[2]}^2$. The extension to higher-order dimensions is straightforward. Define

$$GFfit_\perp^{(ijk)} = \sum_{\ell=d+(c-1)^2+1}^{\ell=d+(c-1)^3} \hat{\gamma}_\ell^2, \tag{4.8}$$

where $d = (i-1)(c-1)^2 + (j-2)(c-1)^2$, with similar definitions for $GFfit_\perp^{(ijkm)}$ to $GFfit_\perp^{(ijk,\cdots,v,w)}$.

The relationship between X_{PF}^2 and $GFfit_\perp$ in a cross-classified table is given in Result 4.1: **Result 4.1** Assuming the model $\boldsymbol{\pi}(\boldsymbol{\beta})$ is identified for $q \geq 3$ and $g \geq q(c-1)$, with at least $(c-1)$ unknown intercept or first-order parameters for each variable in the cross-classified table, and that $q(c-1)$ of the orthogonal components for first-order marginals are fixed at zero by eliminating the $\mathbf{H}_{[1]}$ partition from $\mathbf{H}$, then

$$\begin{aligned} X_{PF}^2 = &\sum_i\sum_j GFfit_\perp^{(ij)} + \sum_i\sum_j\sum_k GFfit_\perp^{(ijk)} + \cdots \\ &+ \sum_i\sum_j\cdots\sum_v GFfit_\perp^{(i,j,\ldots,v)} + GFfit_\perp^{(i,j,\ldots,v,w)}, \end{aligned} \tag{4.9}$$

where $(i,j,\ldots,v)$ is the $(q-1)$ - order of cross-classification.

The result follows because the $GFfit_\perp$ are partial sums of the non-zero $T-g-1$ components of X_{PF}^2. Since expression 4.9 decomposes the entire X_{PF}^2 statistic into components, rows of $\mathbf{H}$ that correspond to estimated parameters of the model $\boldsymbol{\pi}(\boldsymbol{\beta})$ should be considered, as mentioned in Section 3.1. The joint proportions of a c^q table can be transformed to $c^q - 1$

marginal proportions, but g of these marginal proportions will be exactly fit under a model with g estimated parameters, and the orthogonal components for the Pearson-Fisher statistic on these marginals will be identically equal to zero. So for Result 4.1, while $\mathbf{H}_{[1]}$ has been deleted from $\mathbf{H}_{[1:q]}$, the $\hat{\gamma}_\ell^2$ may contain additional components that are identically equal to zero, but the components that are identically equal to zero can be eliminated from $\hat{\boldsymbol{\gamma}}$ be deleting a total of g linearly independent rows from $\mathbf{H}_{[1:q]}$, as discussed further below. Then,

$$\begin{aligned} X_{PF}^2 = \hat{\boldsymbol{\gamma}}'\hat{\boldsymbol{\gamma}} = \sum_{\ell=1}^{\ell=T-(g-q(c-1))-1} \hat{\gamma}_\ell^2 &= \sum_{\ell=1}^{\ell=\binom{q}{2}(c-1)^2} \hat{\gamma}_\ell^2 + \sum_{\ell=\binom{q}{2}(c-1)^2+1}^{\ell=\binom{q}{2}(c-1)^2+\binom{q}{3}(c-1)^3} \hat{\gamma}_\ell^2 + \cdots \\ &+ \sum_{\ell=\sum_{k=2}^{q-2}\binom{q}{k}(c-1)^k+1}^{\ell=\sum_{k=2}^{q-1}\binom{q}{k}(c-1)^k} \hat{\gamma}_\ell^2 + \sum_{\ell=\sum_{k=2}^{q-1}\binom{q}{k}(c-1)^k+1}^{\ell=T-(g-q(c-1))-1} \hat{\gamma}_\ell^2 \end{aligned} \tag{4.10}$$

Each $GFfit^{(ij)}$ statistic is comprised of $(c-1)^2$ orthogonal components, and there are $\binom{q}{2}$ $GFfit^{(ij)}$ statistics, hence

$$\sum_i \sum_j GFfit_\perp^{(ij)} = \sum_{\ell=1}^{\ell=\binom{q}{2}(c-1)^2} \hat{\gamma}_\ell^2; \tag{4.11}$$

next, each $GFfit^{(ijk)}$ statistic is composed of $(c-1)^3$ orthogonal components, and there are $\binom{q}{3}$ $GFfit^{(ijk)}$ statistics, hence

$$\sum_i \sum_j \sum_k GFfit_\perp^{(ijk)} = \sum_{\ell=\binom{q}{2}(c-1)^2+1}^{\ell=\binom{q}{2}(c-1)^2+\binom{q}{3}(c-1)^3} \hat{\gamma}_\ell^2; \tag{4.12}$$

etc., until $T-g-1$ non-zero components have been accounted for. If $\mathbf{H}$ includes more than $T-g-1$ linearly independent rows, any components beyond the first $T-g-1$ non-zero components would be identically equal to zero. Then the $GFfit_\perp$ statistics are components of X_{PF}^2. Result 4.1 also means that components of X_{PF}^2 can be used to also form $GFfit_\perp$ diagnostics that are asymptotic independent chi-square statistics for higher-order marginal tables from cross-classified multi-category variables.

As mentioned above, expression 4.9 decomposes the entire X_{PF}^2 statistic, so rows of $\mathbf{H}_{[1:q]}$ corresponding to estimated parameters need further discussion. If a model has no parameters to be estimated, then no rows would be eliminated from $\mathbf{H}$. In the theory presented above, the assumption is stated that the partition for $\mathbf{H}_{[1]}$ has been eliminated from $\mathbf{H}$ as part of g rows to be deleted under a composite null hypothesis for the purpose of rendering $\boldsymbol{\Sigma_e}$ full rank. If the $X_{[1]}^2$ components are fixed at zero in this model and thus excluded from $\hat{\boldsymbol{\gamma}}$, $g-q(c-1)$ additional rows would need to be deleted from $\mathbf{H}$ in order to maintain full rank $\boldsymbol{\Sigma_e}$ in the complete decomposition of X^2_{PF}. By default the last $(g - q(c-1))$ components would be identically equal to zero and excluded. Then the upper limit of the sum in expression 4.9 would be $\ell = T - g - 1$. It is not necessary to delete any rows from $\mathbf{H}$ under a composite null hypothesis, but then a generalized inverse would be needed for $\boldsymbol{\Sigma}_{\boldsymbol{e}}$ and some components would be identically equal to zero. If rows are to be deleted from $\mathbf{H}$ to eliminate components identically equal to zero, the preferred choice of rows will be self-evident from the model $\boldsymbol{\pi}(\boldsymbol{\beta})$. In some models for a composite null hypothesis, fixing the components of $X^2_{[1]}$ at zero by deleting the partition for $\mathbf{H}_{[1]}$ from $\mathbf{H}$ is necessary if $\boldsymbol{\Sigma}_{\boldsymbol{e}}$ is to be full rank. For example, under a simple log-linear independence model for cross-classified variables, the first-order marginals would be exactly fit, so $X^2_{[1]} \equiv 0$ in that case, and deleting $\mathbf{H}_{[1]}$ from $\mathbf{H}$ produces a full rank $\boldsymbol{\Sigma}_{\boldsymbol{e}}$. A condition that could be checked for log-linear models is that the rows of matrix $\mathbf{H}$ must be linearly independent of the columns of the model matrix for the model $\boldsymbol{\pi}(\boldsymbol{\beta})$. On theory, in other models, any g rows could be deleted. For example, in the 2 PL item response (IRT) model using the marginal likelihood, no marginals for response variables are exactly fit and any g rows can be deleted from $\mathbf{H}$ in order to obtain full rank $\boldsymbol{\Sigma}_{\boldsymbol{e}}$.

In an application with a composite null hypothesis, the $\mathbf{H}_{[1]}$ partition would virtually always be deleted from $\mathbf{H}$ when constructing $X^2_{[t:u]}$ because $X^2_{[1]}$ contributes very little to the power for detecting poor fit with a composite null hypothesis since it consists of components that do not reflect variable associations. Also in an application, the complete decomposition of X^2_{PF} into components would rarely be done because the focus would be on lower-order marginals, and as discussed previously, $X^2_{[2]}$ and $GFfit^{(ij)}_{\perp}$ are calculated from the $\mathbf{H}_{[2]}$ partition of $\mathbf{H}_{[1:q]}$ so more than g rows have already been deleted, and no additional rows would be deleted unless a bivariate table happens to be exactly fit by the model.

A lack-of-fit diagnostic such as $GFfit_{\perp}$ is very focused and would be employed after a more omnibus test result indicates that the hypothesized model does not fit. The testing procedure would be similar to testing a global null hypothesis of no difference among groups in an analysis of variance and then partitioning the sum of squares into orthogonal components for testing contrasts across groups. Each $GFfit^{(ij)}_{\perp}$ is composed of $(c-1)^2$ components, and if a particular $GFfit^{(ij)}_{\perp}$ is unduly large, the individual 1 d.f. components of that $GFfit^{(ij)}_{\perp}$ may be examined for the purpose of obtaining more detail on lack of fit within the marginal table. A large number of variables produces a large number of *GFfit* statistics, and a multiple decision rule should be used to assess significance levels. Since the $GFfit^{(ij)}_{\perp}$ are asymptotically independent, control of the false discovery rate or familywise error rate is facilitated in comparison to other diagnostics. $GFfit^{(ij)}_{\perp}$ is an orthogonal version of the original $GFfit^{(ij)}$ statistic with the added features that the $GFfit^{(ij)}_{\perp}$ are independent and sum to $X^2_{[2]}$. As with the original $GFfit^{(ij)}$ statistic, it is proposed as a lack-of-fit diagnostic for variable pairs. Because of the sequential nature of the orthogonal components, the $GFfit_{\perp}$ in expression 4.9 have a sequential property and are order dependent. In the literature on orthogonal components of Pearson's X^2 statistic (Lancaster, 1969; Raynor and Best, 1989; Miralev, 1987; Eubank, 1997), components are sequential by nature because they are obtained by an application of the Cholesky factor or Gram-Schmidt orthogonalization. In many applications, such as educational testing, the order of the response variables, and hence the order of the $GFfit_{\perp}^{(ij)}$ will be arbitrary. Simulation results to be presented in Section 6 will show that an arbitrary order may have modest, if any, effect on power of $GFfit_{\perp}^{(ij)}$ to detect sources of lack of fit. Occasionally, it is possible to use an ordering among the variables to an advantage, as will be demonstrated in application to symptoms of psychological depression. In any case, $GFfit_{\perp}^{(ij)}$ will reliably have higher probability to detect existing lack of fit than other diagnostics after correction for multiple testing, as demonstrated in Section 7.

Calculation of $GFfit_{\perp}^{(ij)}$ is computationally intensive when the number of response variables is large. The computational issues for a large number of binary variables are discussed in detail by Reiser (2019). The dimensions of $\mathbf{H}$ and $\mathbf{r}$ increase exponentially with the number of variables, i.e, c^q, but these are sparse matrices in the sense that most of the entries are zeros. Computations done using the *R* software environment can take advantage of the *sparsematrix* function to substantially reduce the memory required to hold and perform calculations with these matrices. It is also possible to take advantage of multicore processing in the *R* environment. The *R* code available in the online supplement uses the *sparsematrix* function and multiple cores. Each of the four lack-of-fit statistics examined in the simulations require an exhaustive computation across all possible response patterns to obtain the expected values which are then collapsed to obtain the cells of the marginal tables. For a composite null hypothesis, $GFfit_{\perp}^{(ij)}$ requires calculation of the matrix $\mathbf{G} = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, a large matrix with dimension T by g. For a simple null hypothesis, $GFfit_{\perp}^{(ij)}$ does not include the $\mathbf{G}$ matrix, and the amount of random access memory needed for calculation is substantially reduced. Calculations for $GFfit_{\perp}^{(ij)}$ with a composite null hypothesis and 20 binary variables can be performed on a laptop with 16 GB of memory using vectorized code in about three minutes of CPU time. With more categories and more than 20 response variables, calculations to store and manipulate these matrices using vectorized coding require a server with large amounts of random access memory. With a typical workstation, the limit for binary variables is about 30 response variables. With more categories, the limit on the number of response variables is reduced. Although less efficient in terms of cpu time, calculations for first- and second-order marginals can be done with loops rather than vectorized code which reduces the demand for random access memory.

When the number of response categories increases beyond two, even the frequencies in second-order marginals may exhibit sparseness. The asymptotic chi-square approximation for $GFfit_{\perp}^{(ij)}$ and other diagnostics reviewed below may not be valid if sparseness is present in the frequencies. Cai and Hansen (2013) found that cells in corner of larger two-way tables tended to have low expected frequencies, although other cells may have low expected values depending on the model parameters. The $GFfit_{\perp}^{(ij)}$ statistic can be decomposed into components that would correspond to some cells that would be expected to be sparse and components for other cells that would not be sparse. The asymptotic chi-square approximation would be valid for a statistic obtained as a sum of components that correspond to cells that are not sparse. A method for decomposing the components of $GFfit_{\perp}^{(ij)}$ for this purpose is given in Appendix B.

5. Other Diagnostics for Marginal Tables with Multiple Categories

$GFfit_{\perp}^{(ij)}$ is a general lack-of-fit diagnostic because it can be applied to a wide variety of models for cross-classified tables. In addition to $GFfit^{(ij)}$, a few other general lack-of-fit statistics for marginal tables that can be applied to a wide variety of models for cross-classified variables have been proposed in the literature. Although general, applications have been mostly, although not exclusively, to IRT models with the aim of detecting the item associations that are not fit well by the model because the these models often involve high-dimensional cross-classified tables. In addition to $GFfit^{(ij)}$, the class of widely applicable statistics includes M_{ij} (Maydeu-Olivares & Joe, 2006) and $\bar{\bar{X}}_{ij}^2$, a statistic based on a mean and variance correction given by Asparouhov and Muthén (2010). M_{ij} and $\bar{\bar{X}}_{ij}^2$ will be reviewed in this section because they will be compared to $GFfit_{\perp}^{(ij)}$ using simulations in Section 6. The Lagrange multiplier approach of Glas (1999) will also be reviewed in this section, although it is less general in that it is specific to the IRT model.

M_{ij} is a bivariate version of the M_r statistic (Maydeu-Olivares and Joe, 2005, 2006). M_r is an omnibus lower-order test statistic for cross-classified data, where marginals up to order r are included in the statistic. If M_r detects a model misfit, Maydeu-Olivares and Joe (2006) suggest to identify the item pair associations that are not fit well by applying M_r to marginal tables. In the case of bivariate residuals, the statistic becomes

$$M_{ij} = \boldsymbol{e}_{ij}'\widehat{\boldsymbol{C}}_{ij}\boldsymbol{e}_{ij} \tag{5.1}$$

with $\boldsymbol{e}_{ij} = \boldsymbol{H}_{[2]}^{(ij)}(\widehat{\boldsymbol{p}} - \boldsymbol{\pi}(\widehat{\boldsymbol{\beta}}))$ and

$$\widehat{\boldsymbol{C}}_{ij} = (\widehat{\boldsymbol{D}}_{[2]}^{(ij)})^{-1} - (\widehat{\boldsymbol{D}}_{[2]}^{(ij)})^{-1}\widehat{\boldsymbol{G}}_{[2]}^{(ij)}((\widehat{\boldsymbol{G}}_{[2]}^{(ij)})'(\widehat{\boldsymbol{D}}_{[2]}^{(ij)})^{-1}\widehat{\boldsymbol{G}}_{[2]}^{(ij)})^{-1}(\widehat{\boldsymbol{G}}_{[2]}^{(ij)})'(\widehat{\boldsymbol{D}}_{[2]}^{(ij)})^{-1} \tag{5.2}$$

where $\widehat{\boldsymbol{D}}_{[2]}^{(ij)} = \boldsymbol{H}_{[2]}^{(ij)}\boldsymbol{D}(\widehat{\boldsymbol{\pi}})(H_{[2]}^{(ij)})'$ and $\widehat{\boldsymbol{G}}_{[2]}^{(ij)} = \boldsymbol{H}_{[2]}^{(ij)}\widehat{\boldsymbol{G}}$. M_{ij} is asymptotically distributed as chi-square with degrees of freedom equal to $c^2 - g_{ij} - 1$ where g_{ij} is defined above, although the joint distribution of a set of M_{ij} for a set of bivariate tables has not been given. M_{ij} can be easily applied to higher-order marginal tables (Maydeu-Olivares and Liu, 2012). Liu and Maydeu-Olivares (2014) compared the performance of M_{ij} with several statistics based on bivariate marginal tables for pairs of items.

Another statistic is the Pearson X^2 applied to bivariate marginal tables:

$$X_{ij}^2 = n(\boldsymbol{e}_{ij}'(\widehat{\boldsymbol{D}}_{[2]}^{(ij)})^{-1}\boldsymbol{e}_{ij}) \tag{5.3}$$

X_{ij}^2 is the same statistic as the original $GFfit^{(ij)}$. Since, in general, X_{ij}^2 does not follow an asymptotic chi-square distribution, it can be adjusted with its asymptotic mean and variance so that the asymptotic distribution is closer to chi-square. The estimated asymptotic moments are given by

$$\widehat{\mu}_1 = tr\left((\widehat{\boldsymbol{D}}^{(ij)}_{[2]})^{-1}\widehat{\boldsymbol{\Sigma}}^{(ij)}_{\boldsymbol{e}}\right) \qquad \widehat{\mu}_2 = 2tr\left((\widehat{\boldsymbol{D}}^{(ij)}_{[2]})^{-1}\widehat{\boldsymbol{\Sigma}}^{(ij)}_{\boldsymbol{e}}\right)^2. \tag{5.4}$$

Using a mean and variance correction given by Asparouhov and Muthén (2010), $\bar{\bar{X}}^2_{ij} = a^* + b^* X^2_{ij}$, where a^* and b^* are chosen so that the mean and variance of $\bar{\bar{X}}^2_{ij}$ are df_{ij} and $2\mathrm{df}_{ij}$, where $\mathrm{df}_{ij} = c^2 - g_{ij} - 1$, and g_{ij} is the number of parameters for the bivariate table. Then,

$$\bar{\bar{X}}^2_{ij} = X^2_{ij}\sqrt{\frac{2\mathrm{df}_{ij}}{\hat{\mu}_2}} + \mathrm{df}_{ij} - \sqrt{\frac{2\mathrm{df}_{ij}\hat{\mu}^2_1}{\hat{\mu}_2}}. \tag{5.5}$$

While $GFfit^{(ij)}_{\perp}$ can be applied to a model for binary response variables, neither M_{ij} nor $\bar{\bar{X}}^2_{ij}$ can be applied to binary cross-classified variables with estimated parameters because the degrees of freedom would become negative

Liu and Maydeu-Olivares (2014) proposed a similar statistic, $\bar{X}^2_{ij}$, which extends the statistic derived by Bartholomew and Leung (2002) and Cai et. al. (2004) to composite null hypothesis for bivariate marginal tables. The latter authors also proposed a correction to improve the power of the test. Another general approach for marginal tables is the standardized residual for ordered data in the form suggested by Reiser (1996) and Maydeu-Olivares and Joe (2005).

While $GFfit^{(ij)}_{\perp}$, $\bar{\bar{X}}^2_{ij}$, and M_{ij} can be applied to a variety of models for cross-classified variables, there is a sizeable literature on lack-of-fit diagnostics specific to IRT models. Among that literature, the entry most relevant to the approach taken in this paper is a Lagrange multiplier statistic proposed by Glas (1999) to assess fit on pair-wise associations in IRT models. For manifest variables with multiple categories, Glas worked with the nominal response model (Bock, 1972). Introducing lack-of-fit parameters $(\boldsymbol{\gamma}_{jk} \vdots \boldsymbol{\delta}_{jk})'$ into the model, under the composite null hypothesis $H_0 : (\boldsymbol{\gamma}_{jk} \vdots \boldsymbol{\delta}_{jk})' = \mathbf{0}$, the Lagrange multiplier statistic $LM(\boldsymbol{\gamma}_{jk}, \boldsymbol{\delta}_{jk})$ is distributed asymptotically as a central chi-square statistic on $2(c_j - 1)(c_k - 1)$ degrees of freedom. Simulations and an example that employ the GLLVM for the graded response model will be presented in Sections 4.1 and 7. It is not clear how to extend the Lagrange multiplier method for a lack-of-fit diagnostic to graded response model, and further work is needed to develop this approach for the multi-category graded response model. So, a comparison of the performance of $GFfit^{(ij)}_{\perp}$ and $LM(\boldsymbol{\gamma}_{jk}, \boldsymbol{\delta}_{jk})$ in terms of power and Type I error for the IRT model are not included in the remainder of this paper and are left for future research.

Also in the area of IRT models, Glas & Verhelst (1989, 1995), Glas & Suarez-Falcon (2003) have developed lack-of-fit diagnostics specific to the Rasch model based on the underlying multinomial distribution.

The diagnostics reviewed above are based on assessing the fit of a model to pair-wise associations among a set of manifest variables. A different approach was taken by Yen (1981), who developed a statistic Q_1 to assess the fit of item characteristic curves for IRT models. Since

Q_1 is designed to determine if an IRT model does not fit well to a particular manifest variable, it is fundamentally different from the other diagnostics reviewed above that assess the fit of a model to the *associations* among manifest variables. Because of this fundamental difference in the Q_1 approach, it is not included in the simulation studies that are presented below. Bock (1972) developed a similar statistic for the item characteristic curve.

Among all of the diagnostics discussed above for marginal tables with multiple categories, $GFfit_{\perp}^{(ij)}$ has two unique properties: a set of $GFfit_{\perp}^{(ij)}$ are asymptotically independent with known joint distribution, and the set of $GFfit_{\perp}^{(ij)}$ are components of X_{PF}^2.

6. Simulation Studies and Asymptotic Power Calculations

The statistics developed above can be applied to cross-classified tables from a variety of models including categorical variable factor analysis, latent class analysis, and manifest variable log-linear models. Simulation studies to assess the performance of $X_{[2]}^2$ and $GFfit_{\perp}^{(ij)}$ applied to multi-category variables were conducted using the Generalized Linear Latent Variable Model (GLLVM).

6.1. GLLVM

GLLVM is a latent variable response model for categorical variables with two or more graded categories and has features of a proportional odds model. Let $\mathbf{y} = (y_1,\ y_2, \cdots, y_q)'$ be the vector of q ordinal observed variables, each of them having c_i categories. Thus there are $\prod_{i=1}^{q} c_i$ cells, also called response patterns in the cross-classified table. Response pattern s is indicated as $\mathbf{y}_s = (y_1 = a_1,\ y_2 = a_2, \cdots, y_q = a_q)'$, where a_i is the value of the i^{th} observed variable ($a_i = 1, \ldots, c_i$ and $i = 1, \ldots, q$). Let $\boldsymbol{X} = (X_1,\ X_2, \cdots, X_p)'$ be the vector of p continuous latent variables. Then the probability of response pattern s is given by

$$\pi_s(\boldsymbol{\beta}) = \mathbb{P}(\boldsymbol{Y} = \mathbf{y}_s \mid \boldsymbol{\beta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{P}(\boldsymbol{Y} = \mathbf{y}_s \mid \boldsymbol{\beta}, \boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x},\ s = 1, 2, \ldots, T \tag{6.1}$$

where $f(\boldsymbol{x})$ is the density function of $\boldsymbol{X}$, and the multiple integral is over dimension p.

The conditional probability of $\boldsymbol{Y}$ given $\boldsymbol{x}$ expresses conditional independence:

$$\mathbb{P}(\boldsymbol{Y} = \mathbf{y}_s | \boldsymbol{x}) = \prod_{i=1}^{q} \pi_{a_{is}}^{(i)}(\boldsymbol{x}) = \prod_{i=1}^{q} (\eta_{a_{is}}^{(i)} - \eta_{a_{i-1,s}}^{(i)}) \tag{6.2}$$

where $\eta_{a_i}^{(i)} = \pi_1^{(i)}(\boldsymbol{x}) + \pi_2^{(i)}(\boldsymbol{x}) + \cdots + \pi_{a_i}^{(i)}(\boldsymbol{x})$ is the probability of a response in category a_i or lower on the variable i, and $\pi_{a_i}^{(i)}(\boldsymbol{x})$ is the probability of a response in category a_i on the variable i. Logistic regression is used to model the interrelationship between $\eta_{a_i}^{(i)}$ and the latent variables:

$$\log\left(\frac{\eta_r^{(i)}}{1 - \eta_r^{(i)}}\right) = \beta_{i0}(r) - \sum_{j=1}^{p} \beta_{ij} x_j,$$

$r = 1, \ldots, c_{i-1}$, where c_i is the number of categories for variable i, $\beta_{i0}(r)$ and β_{ij} are the parameters of the model. $\beta_{i0}(r)$ is an intercept and β_{ij} is the j^{th} slope for variable i. The intercepts should satisfy the condition $\beta_{i0}(1) \leq \beta_{i0}(2) \leq \cdots \leq \beta_{i0}(c_i)$. The integrals are approximated through the Gauss-Hermite quadrature method.

6.2. Monte Carlo Simulation for $GFfit_{\perp}^{(ij)}$ Type I Error

We now present results for simulations of the diagnostic statistics. A simulation study was conducted using GLLVM to assess the accuracy of the Type I error rates for $GFfit_{\perp}^{(ij)}$, M_{ij}, X_{ij}^2, and $\bar{\bar{X}}_{ij}^2$. M_{ij} is the Maydeu-Olivares and Joe (2005) statistic applied to the (i, j) table. X_{ij}^2 is similar to the Pearson's statistic, except instead of using the joint frequencies, it is calculated by using the marginal frequencies. $\bar{\bar{X}}_{ij}^2$ is X_{ij}^2 with adjustment using the first two moments. Although the full Pearson statistic is distributed chi-square, X_{ij}^2 is not, as mentioned previously. Simulation results for X_{ij}^2 were calculated using the central chi-square distribution on $(c_i - 1)(c_j - 1)$ degrees of freedom as the reference distribution, as suggested by Joreskog and Moustaki (2001), where $c_i = c_j = c$. The central chi-square distribution was also used as the reference distribution for $\bar{\bar{X}}_{ij}^2$. Standardized residuals (Reiser, 1996; Maydeu-Olivares & Joe, 2005) and $\bar{X}_{ij}^2$ were not included in the comparison because results reported elsewhere (Dassanayake, Reiser, & Zhu, 2016) show that $GFfit_{\perp}^{ij}$ has higher power than these two diagnostics.

The design of this Monte Carlo study was as follows. Pseudo data was generated and fit with a one-factor GLLVM. The number of manifest variables was $q =$ 4, 6, 8, and 10. The number of categories was $c = 3$ and 4 at four manifest variable, and $c = 4$ for six, eight, and ten manifest variables. The number of pseudo samples for each simulation was 1000, and sample size was both 300 and 500 for each set of manifest variables. In addition, simulations were also run using sample size 1000 and 5000 with six and eight variables. The joint frequencies in the simulated data tables for six, eight and ten variables are very sparse: $4^8 = 65,536$, so even with sample size 5000, a high proportion of the joint frequencies are zeros.

In the GLLVM model, the maximum likelihood estimator is consistent and efficient, but the estimator has large mean square error in finite samples when model parameters have large magnitudes, i.e, large intercepts, either positive or negative, and large slopes. Since this study is concerned with the performance of the test statistics, only modest values for GLLVM parameters were used in order to avoid confounding of the performance of the test statistic with bias in parameter estimation. To generate the pseudo data, intercept values were specified in as -1.5 and 1.5 for three categories and -1.5, 0, and 1.5 with four categories. Slope parameters were specified as follows: With four variables, slopes were 0.2, 0.5, 1.0, 2.0; for six variables, slopes were 0.2, 0.5, 0.75, 1.0, 1.5, 2.0; for eight variables slope parameter values were 0.2, 0.5, 0.65, 0.75, 1.0, 1.25, 1.5, and 2.0; and for ten variables 0.2, 0.5, 0.65, 0.75, 1.0, 1.15, 1.25, 1.5, 1.75, and 2.0. At four and six manifest variables, parameter values for intercept and slope were the same as the values used in the simulations for the omnibus statistics described in Section **??** of

Appendix B. All intercept and slope parameters were estimated using the *mirt* package and the *grm* function from the *ltm* package in R. Test statistics were calculated using a custom *R* script. $GFfit_{\perp}^{ij}$ were calculated using the orthogonal regression presented in Reiser (2008). Integrals to approximate the probability of response patterns were calculated with 21 quadrature points. A higher number of quadrature points, and the use of adaptive quadrature would be useful for more accurate estimation of GLLVM parameters when the intercepts and slopes are more extreme, and more accurate estimation of model parameters could improve the performance of the lack-of-fit diagnostics. The intercept and slope parameters used for these simulations were not extreme, and using more quadrature points does not change the results.

The null hypothesis for each marginal table is $H_o : \mathbf{H}_{[2]}^{(ij)}\boldsymbol{\pi} = \mathbf{H}_{[2]}^{(ij)}\boldsymbol{\pi}(\boldsymbol{\beta})$. Simulation results for Type I error using sample size 300 and 500 with four, six and eight variables are shown in the Tables 1 to 3, which show empirical Type I error for nominal $\alpha = 0.05$. While Table 1 has four variables with three categories, Table **??** in Appendix C has Type I error simulation results using four variables with four categories, and results for ten variables are shown in Table **??** of Appendix C.

Table 1: Type I Error Results for $GFfit_{\perp}^{(ij)}$, M_{ij}, X_{ij}^2, and $\bar{\bar{X}}_{ij}^{2}$
Four variables, Three categories

	$n = 300$				$n = 500$			
(ij)	$GFfit_{\perp}^{(ij)}$	M_{ij}	X_{ij}^2	$\bar{\bar{X}}_{ij}^{2}$	$GFfit_{\perp}^{(ij)}$	M_{ij}	X_{ij}^2	$\bar{\bar{X}}_{ij}^{2}$
(12)	0.059	0.053	0.040	0.046	0.047	0.049	0.041	0.043
(13)	0.052	0.046	**0.030**	0.043	**0.035**	**0.035**	**0.020**	**0.036**
(14)	0.063	0.050	**0.032**	0.043	0.055	0.057	**0.033**	0.054
(23)	0.051	0.045	**0.026**	0.044	**0.036**	0.039	**0.021**	0.047
(24)	0.057	0.041	**0.028**	0.057	0.046	0.042	**0.025**	0.040
(34)	0.052	0.037	0.044	**0.070**	0.047	0.055	**0.022**	0.041

1000 samples; 991 ($n = 300$), 999 ($n = 500$) convergence

Tables 1 to 3 contain empirical Type I error rates for 98 bivariate marginal tables. Using a large sample approximation for the binomial distribution, a margin of error interval for these error rates is $0.05 \pm 1.96\sqrt{(0.05)(0.95)/1000} = (0.0365, 0.0635)$. All Type I error results that are outside of this interval are shown in bold in the Type I error tables. From these tables it can be seen that $GFfit_{\perp}^{(ij)}$ has four error rates outside this interval, M_{ij} has four error rates outside this interval, $\bar{\bar{X}}_{ij}^{2}$ has seven error rates outside of this interval, but X_{ij}^2 has 38 error rates below this interval.

Table 2: Type I Error Results for $GFfit_{\perp}^{(ij)}$, M_{ij}, X^2_{ij}, and $\bar{\bar{X}}^2_{ij}$
Six variables, Four categories

	$n = 300$				$n = 500$			
(ij)	$GFfit_{\perp}^{(ij)}$	M_{ij}	X^2_{ij}	$\bar{\bar{X}}^2_{ij}$	$GFfit_{\perp}^{(ij)}$	M_{ij}	X^2_{ij}	$\bar{\bar{X}}^2_{ij}$
(12)	0.037	0.044	0.037	0.040	0.039	0.041	0.040	0.042
(13)	0.045	0.043	0.037	0.041	0.042	0.045	0.040	0.043
(14)	0.039	0.041	**0.024**	**0.031**	0.052	0.044	0.041	0.049
(15)	0.054	0.053	**0.036**	0.049	0.058	0.055	0.044	0.059
(16)	0.051	0.046	0.045	0.049	0.058	0.061	0.043	0.061
(23)	0.051	0.046	0.045	0.049	0.057	0.053	0.051	0.056
(24)	0.041	0.037	**0.034**	0.041	0.045	0.045	0.043	0.050
(25)	0.050	0.052	0.041	0.051	0.055	0.053	0.048	0.056
(26)	0.053	0.053	**0.031**	0.054	0.045	0.052	0.037	0.051
(34)	0.062	0.052	0.049	0.059	0.055	0.050	0.038	0.047
(35)	0.048	0.053	0.037	0.056	0.048	0.050	**0.034**	0.046
(36)	0.057	0.058	0.045	**0.064**	0.052	0.047	**0.035**	0.054
(45)	0.054	0.044	0.037	0.050	0.052	0.053	**0.029**	0.047
(46)	0.053	0.054	0.040	0.051	0.049	0.047	**0.030**	0.050
(56)	0.059	**0.068**	0.045	**0.067**	0.043	0.052	**0.030**	0.046

1000 samples, 100% convergence

Table 3: Type I Error Results for $GFfit_{\perp}^{(ij)}$, M_{ij}, X^2_{ij}, and $\bar{\bar{X}}^2_{ij}$
Eight variables, Four categories

	$n = 300$				$n = 500$			
(ij)	$GFfit_{\perp}^{(ij)}$	Mij	X^2_{ij}	$\bar{\bar{X}}^2_{ij}$	$GFfit_{\perp}^{(ij)}$	M_{ij}	X^2_{ij}	$\bar{\bar{X}}^2_{ij}$
(12)	0.047	0.057	0.042	0.044	0.046	0.043	0.044	0.046
(13)	0.054	0.056	0.047	0.051	0.049	0.052	0.049	0.050
(14)	0.058	0.048	0.050	0.056	0.061	0.051	0.056	0.060
(15)	0.045	0.051	0.034	0.040	0.056	0.048	0.040	0.046
(16)	0.054	0.046	0.039	0.045	0.048	0.049	0.040	0.052
(17)	0.061	**0.065**	0.056	**0.065**	**0.066**	0.062	0.048	0.060
(18)	0.049	0.047	**0.033**	0.045	0.044	0.049	**0.033**	0.046
(23)	0.048	0.048	0.043	0.046	0.048	0.041	0.039	0.044
(24)	0.042	0.049	0.041	0.045	0.052	0.054	0.048	0.050
(25)	0.048	0.050	0.034	0.042	0.047	0.043	0.039	0.044
(26)	0.051	0.043	**0.033**	0.047	0.041	0.041	**0.029**	0.039
(27)	0.045	0.043	**0.033**	0.047	0.044	0.044	**0.031**	0.039
(28)	0.042	0.050	**0.031**	0.041	0.044	0.050	**0.034**	0.045
(34)	0.046	0.048	0.041	0.044	0.049	0.042	0.043	0.050
(35)	0.045	0.052	0.039	0.047	0.063	0.062	0.046	0.055
(36)	0.056	0.060	0.046	0.055	0.046	0.048	**0.034**	0.043
(37)	0.057	0.051	0.040	0.054	0.050	0.058	0.046	0.058
(38)	0.045	0.046	**0.028**	0.043	0.047	0.049	0.039	0.050
(45)	0.052	0.056	0.046	0.058	0.058	0.054	0.049	0.055
(46)	0.050	0.046	0.034	0.045	0.036	0.048	**0.033**	0.040
(47)	0.043	0.052	0.039	0.053	0.044	0.055	**0.036**	0.046
(48)	0.054	0.046	0.038	0.050	0.048	0.046	**0.030**	0.046
(56)	0.055	0.056	**0.034**	0.048	0.060	0.045	0.043	0.051
(57)	0.056	0.045	**0.033**	0.045	**0.068**	**0.066**	0.049	**0.065**
(58)	0.056	0.050	0.040	0.059	0.040	0.056	**0.035**	0.055
(67)	0.049	0.042	0.037	0.048	0.064	0.051	**0.033**	0.052
(68)	0.052	0.048	**0.032**	0.047	0.045	0.051	0.039	0.047
(78)	0.051	0.052	**0.036**	0.055	0.053	0.048	**0.034**	0.055

1000 samples, 100% convergence

To further investigate Type I error, simulations were conducted for six variables and eight variables using pseudo samples with size 1000 and 5000. Results for these simulations are shown in Tables **??** and **??** of Appendix C. Results for these simulations with larger sample sizes show a similar pattern as seen in simulations for sample sizes 300 and 500. $\mathit{GFfit}_{\perp}^{(ij)}$, M_{ij}, and $\bar{\bar{X}}_{ij}^{2}$ have error rates outside the interval described above at a level that would be expected by chance, but X_{ij}^{2} still has too many error rates below the interval described earlier.

A Kolomogorov-Smirnov test was also employed to assess the fit of the chi-square distribution for simulation results on four, five and six variables with sample size 300 and 500. Type I error simulation results for the lack-of-fit diagnostics with five variables are shown in Appendix C. The Kolomogorov-Smirnov test results are consistent with the empirical Type I error rates: out of 74 bivariate marginal tables, $\mathit{GFfit}_{\perp}^{(ij)}$ had two p-values below 0.05, M_{ij} also had two p-values below 0.05, $\bar{\bar{X}}_{ij}^{2}$ had 12 p-values below 0.05, and X_{ij}^{2} had 33 p-values below 0.05. These results for the diagnostic statistics under the null hypothesis indicate that for sample sizes from 300 and 500, the distribution of $\mathit{GFfit}_{\perp}^{(ij)}$ is well approximated by a chi-square distribution with $(c-1)^2$ degrees of freedom, the distribution of M_{ij} is well approximated by a chi-square distribution with degrees of freedom $c^2 - g^{(ij)} - 1$, and the distribution of $\bar{\bar{X}}_{ij}^{2}$ is well approximated by a chi-square distribution with $c^2 - g_{ij} - 1$ degrees of freedom. While the chi-square distribution can be rejected as the asymptotic distribution of X_{ij}^{2}, the approximation may still be close enough to be useful.

6.3. Power Study for $GFfit_{\perp}^{(ij)}$

Power for $GFfit_{\perp}^{(ij)}$ as a diagnostic statistic was studied by first calculating asymptotic power for $GFfit_{\perp}^{(ij)}$ and then comparing to empirical power for $GFfit_{\perp}^{(ij)}$, M_{ij}, X_{ij}^{2}, and $\bar{\bar{X}}_{ij}^{2}$ obtained from Monte Carlo simulations. The power study used GLLVM with four categories for the four, six, eight and ten variable cases. Pseudo data for 1000 samples were generated from a confirmatory two-factor model with all parameters fixed and then fit with a one factor model. Sample size was 300 and 500. Intercept parameters for the manifest variables in all generating models were -1.5, 0, and 1.5. Slope parameters for the data generating model were as follows: with four variables slopes for factor 1, $\boldsymbol{\beta}_1 = (0.2,\ 0.5,\ 1.0,\ 2.0)'$, and slopes for factor 2, $\boldsymbol{\beta}_2 = (0.8,\ 0.8,\ 0.0,\ 0.0)'$; For six variables, $\boldsymbol{\beta}_1 = (0.2, 0.5, 0.75, 1.0, 1.5, 2.0)^{'}$, and $\boldsymbol{\beta}_2 = (0.8, 0.8, 0.8, 0.0, 0.0, 0.0)^{'}$. For eight variables, $\boldsymbol{\beta}_1 = (0.1, 0.1, 0.1, 1.2, 1.2, 1.2, 0.2, 0.2)'$, and $\boldsymbol{\beta}_2 = (1.0, 1.0, 1.0, 0, 0, 0, 1.0, 1.0)'$; for ten variables, $\boldsymbol{\beta}_1 = (0.1, 0.1, 0.1, 1.2, 1.2, 1.2, 0.2, 0.2, 1.2, 0.2)'$, and $\boldsymbol{\beta}_2 = (1.0, 1.0, 1.0, 0, 0, 0, 1.0, 1.0, 0, 1.0)'$. With four variables, the slope values were chosen so that two variables cross-load on both factors. For six variables, slopes were chosen so that three variables cross-load. There is more heterogeneity in the cross-loading patterns for eight and ten variables. Anther power study was done using six variables with non-zero cross-loading for all variables: $\boldsymbol{\beta}_1 = (0.2, 0.5, 0.75, 1.0, 1.5, 2.0)^{'}$, and $\boldsymbol{\beta}_2 = (1.0, 1.0, 1.0, 1.0, 1.0, 1.0)^{'}$. With four and six variables, slope parameters are the same as the values used for power simulations of the omnibus statistic in Section **??** of Appendix B. The two latent variables were specified as uncorrelated, each with variance equal to 1.0, so $\mathbf{\Sigma}_{\boldsymbol{X}} = \boldsymbol{I}$, an identity matrix. Estimation of the one-factor GLLVM converged for 1000 cases for both the four variable and six variable simulation, and the convergence rate was very high for the eight and ten variable simulations. The power simulation results with four, six and eight variables are shown in Tables 4 to 6. Power simulation results for ten variables are shown in Tables **??** and **??** of Appendix C. Power simulation results for six variables with non-zero cross loadings for all variables are shown in Table **??** of Appendix C. The column headed "A.Power" in these tables shows asymptotic power. Asymptotic power was calculated by the method given in Reiser (2008). In tables showing results from simulations for power, entries in all rows with asymptotic power higher than 0.05 are highlighted in bold.

Table 4: Power Results for $GFfit_{\perp}^{(ij)}$, M_{ij}, X_{ij}^2, and $\bar{\bar{X}}_{ij}^2$
Four Variables, Four Categories

	A.Power	Empirical Power			
(ij)	$GFfit_{\perp}^{(ij)}$	$GFfit_{\perp}^{(ij)}$	M_{ij}	X_{ij}^2	$\bar{\bar{X}}_{ij}^2$
(12)	**0.7059**	**0.6480**	0.0430	**0.5410**	**0.549**
(13)	0.0502	0.0550	0.0490	0.0490	0.0570
(14)	0.0506	0.0600	0.0520	0.0460	0.0590
(23)	0.0516	0.0570	0.0480	0.0440	0.0620
(24)	0.0525	0.0550	0.0510	0.0280	0.0540
(34)	0.0500	0.0670	0.0540	0.0550	0.0720

n=500, 1000 samples, 100% convergence

From these tables, M_{ij} rarely has power larger than the size of the test. This type of result for M_{ij} has been found in previous studies (Liu & Maydeu-Olivares, 2014). From the four variable case, $GFfit_{\perp}^{(12)}$ has empirical power of 0.6480, which shows that primarily the association between variables 1 and 2 was not adequately explained by the one-factor model, which is intuitive because variables 1 and 2 are primarily associated with only factor 2. The empirical power for $GFfit_{\perp}^{(12)}$ does not quite reach the asymptotic power level. X_{12}^2 appears to have even larger power of 0.5410, but as demonstrated above, it does not distribute chi-square. $\bar{\bar{X}}_{12}^2$ has lower power than $GFfit_{\perp}^{(12)}$ for the lack of fit on variables 1 and 2.

For the results from six variables, there is a similar conclusion: Empirical power for $GFfit_{\perp}^{(ij)}$ is very close to asymptotic power especially for $\{i, j\} = (1, 2), (1, 3), (2, 3)$, where the one-factor

Table 5: Power Results for $GFfit_{\perp}^{(ij)}$, M_{ij}, X_{ij}^2, and $\bar{\bar{X}}_{ij}^2$
Six variables, Four categories

	$n = 300$					$n = 500$				
(ij)	A.Power	$GFfit_{\perp}^{(ij)}$	M_{ij}	X_{ij}^2	$\bar{\bar{X}}_{ij}^2$	A.Power	$GFfit_{\perp}^{(ij)}$	M_{ij}	X_{ij}^2	$\bar{\bar{X}}_{ij}^2$
(12)	**0.344**	**0.348**	0.047	**0.274**	**0.288**	**0.572**	**0.572**	0.033	**0.479**	**0.492**
(13)	**0.368**	**0.322**	0.050	**0.203**	**0.217**	**0.606**	**0.582**	0.043	**0.386**	**0.410**
(14)	0.051	0.057	0.058	0.055	0.060	0.052	0.051	0.051	0.052	0.060
(15)	0.051	0.044	0.040	0.036	0.047	0.052	0.039	0.048	0.036	0.053
(16)	0.051	0.045	0.038	0.034	0.047	0.052	0.053	0.042	0.040	0.050
(23)	**0.354**	**0.372**	0.049	**0.209**	**0.238**	**0.587**	**0.591**	0.038	**0.356**	**0.386**
(24)	0.052	0.046	0.047	0.040	0.046	0.054	0.045	0.046	0.035	0.046
(25)	0.053	0.041	0.043	0.030	0.047	0.056	0.059	0.053	0.054	0.068
(26)	0.053	0.057	0.047	0.044	0.058	0.055	0.058	0.058	0.050	0.065
(34)	0.054	0.049	0.053	0.032	0.045	0.057	0.061	0.063	0.049	0.062
(35)	0.056	0.048	0.050	0.041	0.056	0.060	0.055	0.049	0.040	0.052
(36)	0.055	0.055	0.050	0.034	0.059	0.059	0.060	0.050	0.039	0.055
(45)	0.050	0.050	0.049	0.042	0.057	0.050	0.058	0.056	0.029	0.048
(46)	0.050	0.057	0.057	0.033	0.051	0.051	0.051	0.043	0.028	0.051
(56)	0.051	0.055	0.036	0.029	0.044	0.052	0.051	0.048	0.028	0.047

1000 samples, 100% convergence

Table 6: Power Results for $GFfit_{\perp}^{(ij)}$, M_{ij}, X^2_{ij}, and $\bar{\bar{X}}^2_{ij}$
Eight variables, Four categories

		$n = 300$					$n = 500$			
(ij)	A.Power	$GFfit_{\perp}^{(ij)}$	M_{ij}	X^2_{ij}	$\bar{\bar{X}}^2_{ij}$	A.Power	$GFfit_{\perp}^{(ij)}$	M_{ij}	X^2_{ij}	$\bar{\bar{X}}^2_{ij}$
(12)	0.051	0.054	0.048	0.034	0.053	0.052	0.061	0.061	0.041	0.065
(13)	0.052	0.064	0.040	0.041	0.058	0.054	0.061	0.041	0.031	0.042
(14)	0.064	0.064	0.052	0.045	0.055	0.074	0.063	0.040	0.042	0.052
(15)	0.064	0.067	0.058	0.052	0.063	0.074	0.074	0.041	0.059	0.064
(16)	0.064	0.089	0.056	0.056	0.058	0.075	0.091	0.043	0.062	0.075
(17)	0.050	0.054	0.053	0.038	0.052	0.051	0.056	0.050	0.035	0.051
(18)	0.050	0.052	0.041	0.042	0.052	0.051	0.064	0.042	0.035	0.052
(23)	0.055	0.086	0.058	0.060	0.076	0.059	0.074	0.053	0.045	0.057
(24)	**0.073**	**0.076**	**0.055**	**0.050**	**0.058**	**0.090**	**0.089**	**0.042**	**0.044**	**0.051**
(25)	**0.073**	**0.076**	**0.058**	**0.055**	**0.063**	**0.091**	**0.114**	**0.063**	**0.067**	**0.081**
(26)	**0.074**	**0.101**	**0.043**	**0.046**	**0.057**	**0.092**	**0.115**	**0.052**	**0.045**	**0.060**
(27)	0.050	0.050	0.052	0.053	0.062	0.051	0.062	0.058	0.042	0.058
(28)	0.050	0.057	0.046	0.049	0.061	0.051	0.039	0.045	0.035	0.043
(34)	**0.097**	**0.097**	**0.049**	**0.054**	**0.061**	**0.135**	**0.145**	**0.047**	**0.053**	**0.062**
(35)	**0.098**	**0.118**	**0.048**	**0.050**	**0.057**	**0.137**	**0.175**	**0.046**	**0.061**	**0.068**
(36)	**0.100**	**0.130**	**0.040**	**0.042**	**0.052**	**0.142**	**0.158**	**0.045**	**0.052**	**0.059**
(37)	0.052	0.063	0.047	0.048	0.063	0.054	0.062	0.056	0.042	0.051
(38)	0.052	0.062	0.051	0.048	0.062	0.053	0.060	0.059	0.045	0.059
(45)	**0.925**	**0.847**	**0.083**	**0.895**	**0.897**	**0.996**	**0.977**	**0.093**	**0.990**	**0.991**
(46)	**0.938**	**0.893**	**0.074**	**0.911**	**0.912**	**0.997**	**0.988**	**0.090**	**0.995**	**0.995**
(47)	0.051	0.044	0.041	0.036	0.045	0.051	0.040	0.045	0.035	0.046
(48)	0.051	0.050	0.043	0.040	0.049	0.051	0.057	0.050	0.037	0.049
(56)	**0.925**	**0.910**	**0.068**	**0.891**	**0.893**	**0.998**	**0.994**	**0.076**	**0.990**	**0.990**
(57)	0.051	0.053	0.046	0.045	0.051	0.051	0.053	0.057	0.049	0.063
(58)	0.051	0.042	0.048	0.032	0.040	0.051	0.044	0.049	0.038	0.044
(67)	0.051	0.049	0.040	0.043	0.054	0.051	0.037	0.045	0.034	0.044
(68)	0.051	0.050	0.052	0.046	0.055	0.053	0.058	0.050	0.052	0.055
(78)	0.050	0.051	0.033	0.034	0.045	0.050	0.049	0.037	0.033	0.040

n=300, 1000 samples, 996 converged; n=500, 1000 samples, 994 converged model has the most severe misspecification. The X^2_{ij} appear to be substantially inflated, and $\bar{\bar{X}}^2_{ij}$ has considerably lower power than $GFfit^{(ij)}_{\perp}$ where the model is most severely misspecified. Results shown in Table **??** show that all four lack-of-fit diagnostics have low power under the design where all variables have non-zero cross-loadings on both factors. The omnibus statistics have low power in this simulation design also.

The simulation results for eight variables are somewhat more informative when the effect size is small. For the (45), (46), and (56) associations, the lack-of-fit effect size is very large, so there is little difference when comparing the empirical power for $GFfit^{(ij)}_{\perp}$ to the other lack-of-fit statistics. But when the lack-of-fit effect size is smaller, as in the (24), (25), (26), (34), (35), and (36) associations, higher power for $GFfit^{(ij)}_{\perp}$ to detect lack of fit is clearly apparent in the simulation results. For these associations with smaller lack-of-fit effect size, the power for X^2_{ij} and $\bar{\bar{X}}^2_{ij}$ appear to be, as with M_{ij}, not much higher than the size of the test. A similar pattern of higher power for $GFfit^{(ij)}_{\perp}$ is found in the ten variable power simulation results shown in Appendix C.

The $GFfit^{(ij)}_{\perp}$ are dependent on the order of the variables in the residual vector **r**, so it is informative to compare the simulation results with four and six variables, where lack of fit is present among the first few elements of **r**, to the results for eight and ten variables, where order of variables, and hence order of $GFfit^{(ij)}_{\perp}$, is arbitrary and lack of fit is spread out among the elements of **r**. For eight variables, there are nine bivariate tables where asymptotic power for $GFfit^{(ij)}_{\perp}$ is above the 0.05 level. The rows for these tables are highlighted in bold in Table 6. In each of these rows, the asymptotic power for $GFfit^{(ij)}_{\perp}$ is higher than the empirical power for M_{ij} X^2_{ij} and $\bar{\bar{X}}^2_{ij}$, and in eight of the nine rows, the empirical power for $GFfit^{(ij)}_{\perp}$ is equal to or higher than the empirical power for M_{ij}, X^2_{ij} and $\bar{\bar{X}}^2_{ij}$. Even among the association for variables five and six, which is the 23^{rd} $GFfit^{(ij)}_{\perp}$ statistic, $GFfit^{(ij)}_{\perp}$ still has higher power than the other three lack of fit diagnostics displayed in the table. A very similar pattern can be seen in the simulation results for ten variables in Appendix C.

Although there are 8! = 40,320 permutations of the order of eight variables, for purposes of investigating the effect of variable order, Table **??** in Appendix C shows the asymptotic and empirical power for $GFfit^{(ij)}_{\perp}$ when the order of the eight variables is reversed. (Reversing the order of the variables changes the order of the $GFfit^{(ij)}_{\perp}$ but does not reverse the order of the $GFfit^{(ij)}_{\perp}$.) By reversing the variable order, the $GFfit^{(ij)}_{\perp}$ statistics for marginals (4,5), (4,6) and (5,6) are earlier in the sequence of orthonormalization, similar to the six variable simulation. Table **??** shows that asymptotic power increases slightly when the variables are reordered this way, 0.925 to 0.967 for marginal (4,5), 0.938 to 0.962 for marginal (4,6), and 0.925 to 0.956 for marginal (5,6). There is a somewhat larger change in empirical power for these three marginals when the variables have this order. The simulation results in Table **??** look similar to the results for six variables in that the empirical power for $GFfit^{(ij)}_{\perp}$ is higher than the power for the other diagnostics for all three of these marginals. On the other hand, empirical power for $GFfit_{\perp}^{(ij)}$ on marginal (5,6) is slightly reduced when it is earlier in the sequence. Although the effect of variable order on the power of $GFfit_{\perp}^{(ij)}$ is modest, in some circumstances, it may be possible to employ a strategy to increase power of $GFfit_{\perp}^{(ij)}$ by placing variables that may not be well fit by a model earlier in the orthonormalization. It is important to realize that although small power differences in one-at-a-time tests for $GFfit_{\perp}^{(ij)}$ relative to other diagnostics may result from order dependence as seen in marginal (4,5), $GFfit_{\perp}^{(ij)}$ would nevertheless be substantially more reliable to detect existing lack of fit after a correction for multiple testing because the $GFfit_{\perp}^{(ij)}$ are asymptotic independent while the $\bar{\bar{X}}_{ij}^2$ are not independent due to the origin from X_{ij}^2 and have unknown joint distribution. The correction to maintain Type I error level is much more conservative for the statistics that have unknown joint distribution. Strategies for variable order and correction for multiple testing are demonstrated and discussed further in the application in the next section.

7. Application to Depression Symptoms

The $GFfit_{\perp}^{(ij)}$ statistic was used to evaluate the fit of a one factor model to responses given to questions related to the psychiatric condition depression. In this example, the $GFfit_{\perp}^{(ij)}$ will be compared to $\bar{\bar{X}}_{ij}^2$. M_{ij} and X_{ij}^2 are not included in the example because of inadequate performance in terms of Type I error and power properties as demonstrated in the simulations presented in earlier sections. The data used in this example consist of the responses from 294 adults to the following seven depression symptom questions: (1) “I felt that I could not shake off the blues even with the help of my family and friends,”(2) “I felt hopeful about the future,” (3) “I felt that everything was an effort,” (4) ”I felt lonely,” (5) “I felt fearful,” (6) I thought my life had become a failure,” and (7) “I felt that people disliked me.” The four response categories, ordered from most to least severe, were (1) “most or all of the time,” (2) “occasionally or a moderate amount of time,” (3) “some or little of the time,” and (4) “rarely or none of the time.” The source of the data is Afifi & Clark (1984), and the data is also reproduced in Sharma (1995). Fifty of the 294 sample members were classified as depressed based the Center for Epidemiologic Studies Depression Scale (CESD), which would give a rate that is higher than the prevalence of about 7% for depression in the general U.S. population of adults (NIMH, 2019). The marginal proportions for the severe level of all of the symptoms in the sample are very low, at five percent or less, so the 4^7 cross-classified table is very sparse with a large number of cells that have count equal to 0.

The seven items used for this example were chosen to demonstrate two features of the $GFfit_{\perp}^{(ij)}$ statistics. First, seven item questions were chosen because it is possible to calculate higher-order $GFfit_{\perp}$ in addition to the $GFfit_{\perp}^{(ij)}$ statistics with a 4^7 cross-classified table. Second, as mentioned in the previous section, occasionally, when response variables have a substantive meaning such as in an attitude survey or clinical psychological symptoms, it may be possible to employ a strategy for selecting a variable order based on substantive theory to obtain a moderate increase in power to detect lack of fit in some bivariate tables among the first few variables. We demonstrate this strategy in the clinical psychology example below. However, in other applications, such as educational testing, it may not be possible to devise a variable ordering on substantive considerations to increase power and order of variables will be arbitrary. In the example, we will demonstrate that the correction for multiple testing among correlated statistics with unknown joint distribution is so conservative that $GFfit_{\perp}^{(ij)}$ statistics can be expected to have higher power than the other available diagnostics to detect existing source of lack of fit regardless of variable order.

Depression symptoms are heterogeneous in view of a one-dimensional IRT model (Reiser, 1989), so most of the symptom questions chosen for this example are more homogeneous in that they ask respondents about how they "felt." Six of the symptom questions are "negatively" worded in that "most or all of the time" response would indicate possible depression, but the second symptom question is "positively" worded in that the "most or all of the time" response would indicate a healthy psychological state. This positively worded symptom question may have associations with some of the other symptom questions that are not well fit by a one-dimensional IRT model. Since $GFfit_{\perp}^{(ij)}$ that are extracted early may tend to be larger, based on substantive theory this positively worded symptom question was placed near the beginning of the variables that are cross-classified in the 4^7 table.

The GLLVM with one factor was fit to these data using the R package *mirt*, and fit statistics were calculated using a custom R program. $X_{PF}^2 = 11,515.55$ on $16,355$ degrees of freedom, but the chi-square approximation for the full Pearson statistic should not be considered valid because of the high degree of sparseness in the data table. $X_{[2]}^2 = 222.62$ on 189 degrees of freedom, with p-value < 0.047, indicating that the model should be rejected, although the p-value result might be considered marginal.

$GFfit_{\perp}^{(ij)}$ and $\bar{\bar{X}}_{ij}^2$ goodness-of-fit statistics for the two-way associations are shown in Table 7. Since each survey question had four response categories, the $GFfit^{(ij)}$ statistics follow an asymptotic chi-square distribution on $(4-1)^2 = 9$ degrees of freedom in this application, and the $\bar{\bar{X}}_{ij}^2$ follow an asymptotic chi-square distribution with $4^2 - 8 - 1 = 7$ degrees of freedom. Lack-of-fit results from $GFfit_{\perp}^{(ij)}$ and $\bar{\bar{X}}_{ij}^2$ are fairly similar in this example. Looking at p-values obtained from the central chi-square distribution for one-at-a-time tests without correction for multiple testing, both statistics indicate that three of the two-way associations are not well fit by the one-dimensional GLLVM. Both statistics show large values for the association between "could not shake off the blues" and "felt hopeful about the future," and also for the association the between "everything was an effort" and "felt fearful." However, the two statistics differ on the fit for the association between a third pair of items: $GFfit_{\perp}^{(ij)}$ has a large value for the association between "everything was an effort" and "felt lonely," while $\bar{\bar{X}}_{ij}^2$ has a large value for the association between "felt hopeful" and "people disliked me." This type of difference can be explained by different properties of the statistics: whereas the $GFfit_{\perp}^{(ij)}$ are independent, each $\bar{\bar{X}}_{ij}^2$

is obtained in a manner that ignores the other $\bar{\bar{X}}^2_{ij}$. So, for example, $\bar{\bar{X}}^2_{2,7}$ may be larger than $GFfit_{\perp}^{(2,7)}$ because $\bar{\bar{X}}^2_{2,7}$ may include some overlap with other $\bar{\bar{X}}^2_{ij}$ due to lack of independence.

Since Table 7 has goodness-of-fit statistics for 21 item pairs, it is important to consider the inflation of Type I error rate due to multiple testing. As mentioned earlier, an important difference between $GFfit_{\perp}^{(ij)}$ and $\bar{\bar{X}}^2_{ij}$ is that $GFfit_{\perp}^{(ij)}$ are asymptotically independent while $\bar{\bar{X}}^2_{ij}$ are not independent. For the independent statistics, the false discovery rate (FDR) procedure of Benjamini and Hochberg (1995) can be used to control the inflation of Type I error among the $GFfit_{\perp}^{(ij)}$. The FDR p-values for the $GFfit_{\perp}^{(ij)}$ are shown in the fourth column of Table 7, where it can be seen that $GFfit_{\perp}^{(12)}$ for the association between "could not shake off the blues" and "felt hopeful" is still significant at the $\alpha = 0.05$ level indicating lack of fit. For the $\bar{\bar{X}}^2_{ij}$, which are not independent and have unknown joint distribution, the FDR method is not valid, and the method of Benjamini and Yekutieli (2001) for controlling the false discovery rate under dependency is more appropriate. The BY p-values for $\bar{\bar{X}}^2_{ij}$ are shown in the seventh column of Table 7, where it can be seen that the BY procedure is much more conservative and none of the $\bar{\bar{X}}^2_{ij}$ are significant at even the $\alpha = 0.50$ level after the correction. Strictly for purposes of comparison, FDR p-values for the $\bar{\bar{X}}^2_{ij}$ are also shown in the last column of Table 7, and although the FDR procedure is less conservative, there are still no $\bar{\bar{X}}^2_{ij}$ that are significant at the $\alpha = 0.05$ level after the correction. This example demonstrates that occasionally it may be possible to select a variable order based on substantive theory that would modestly increase the power for the first few $GFfit_{\perp}^{(ij)}$ statistics. More importantly, it demonstrates that the asymptotic independence property of $GFfit_{\perp}^{(ij)}$ allows for a much less conservative correction for multiple testing compared to other diagnostics considered in this study, which will result in higher probability of detecting existing lack of fit for $GFfit_{\perp}^{(ij)}$ regardless of variable order when controlling Type I error. Although the $\bar{\bar{X}}^2_{ij}$ are not order dependent, there are issues of dependency and inflation of the Type I error rate when using this diagnostic. Regardless of variable order, $GFfit_{\perp}^{(ij)}$ will always sum to the same value for the $X^2_{[2]}$ statistic.

With seven variables in the cross-classified table, it is possible to compute some of the higher-order $GFfit_{\perp}$ statistics. These higher-order $GFfit_{\perp}$ are also independent components of the X^2_{PF} statistic and will be used to demonstrate the power of a focused test and further issues of sparseness with an omnibus statistics. In this example with symptoms of depression, the $GFfit_{\perp}^{(ijk)}$ sum to 788.31, which is equal to the value of $X^2_{[3|2]}$, where $X^2_{[u|u-1,u-2,\ldots,t]}$ indicates a statistic calculated using only the columns corresponding $\mathbf{H}^*_{[u]}$ from the larger $\mathbf{H}^*_{[1:q]}$ matrix. A few studies (Mavridis, Moustaki & Knott, 2007; Dassanayake, Reiser, & Zhu, 2016) have indicated that third-order marginal frequencies may be substantial enough in magnitude so that an asymptotic chi-square distribution may be valid for a statistic on third-order marginals. If the asymptotic distribution is valid in this example, then $X^2_{[3|2]}$ would be distributed approximately chi-square on 945 degrees of freedom. Since $X^2_{[2]}$ and $X^2_{[3|2]}$ are sequential and independent, they may be pooled to form the statistic $X^2_{[2:3]}$ which would be equal to 1010.93 on 1134 degrees of freedom. Using this statistic, the null hypothesis $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$ would not be rejected, demonstrating that the $X^2_{[2:3]}$ statistic is *diluted* if the lack of fit is present in the second-order marginals. A study by Salomaa (1990) found that lack of fit for an IRT type model in social science applications is predominantly found in the second-order marginals, so a more *focused* test using the $X^2_{[2]}$ statistic is preferable because it would be expected to have higher power. The sum of the fourth-order $GFfit_\perp$ produce $X^2_{[4|3,2]}$ which is equal to 2284.08, the sum of the fifth-order $GFfit_\perp$ produce $X^2_{[5|4,3,2]}$ which is equal to 3824.28. $GFfit_\perp$ statistics from the sixth- and seventh-order marginals sum to 4395.86 and fill out the X^2_{PF} statistic, although some of the seventh-order $GFfit_\perp$ would be equal to zero because only 2180 degrees of freedom remain for the seventh-order $GFfit_\perp$ and $3^7 = 2187$. The marginal tables for higher-order $GFfit_\perp$ are undoubtedly too sparse to apply an asymptotic chi-square approximation, but the parametric bootstrap could be used for a test of $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$, where $\mathbf{H}$ is $\mathbf{H}_{[4:7]}$ to cover possible lack of fit in higher-order marginals.

Each of the $GFfit^{(ij)}$ statistics shown in Table 7 is a sum of nine orthogonal components of Pearson's statistic. The $GFfit^{(ij)}$ statistics shown in Table 7 sum to 222.62, which is equal to the value of the $X^2_{[2]}$ statistic, and the $X^2_{[2]}$ statistic is then the sum of 189 individual orthogonal components. In a similar way, $X^2_{[3]}$ is a sum of ten $GFfit^{(ijk)}$ statistics.

8. Conclusions

Components of Pearson's chi-square statistic have a long history in the statistical literature. The present work on $GFfit^{(ij)}_\perp$ as a component of Pearson's statistic places this lack-of-fit diagnostic for cross-classified tables in the tradition of Lancaster (1969), Mirvaliev (1987), Raynor & Best (1989) and Eubank(1997). The $GFfit^{(ij)}_\perp$ are asymptotic independent chi-square statistics and are a powerful diagnostic to detect the source of lack of fit in a cross-classified table when a more global test indicates that the hypothesized model does not fit; the more global test should be conducted first. As demonstrated, a more global test statistic such as $X^2_{[2]}$ based on second-order marginals can also be obtained as as a partial sum of $GFfit^{(ij)}_\perp$. Power calculations for an IRT model show that $GFfit^{(ij)}_\perp$ has substantially higher power for detecting the source of lack of fit compared to other general diagnostics on bivariate marginal tables. Simulation results using an IRT model show that $GFfit^{(ij)}_\perp$ has good Type I error performance even if the joint frequencies in the cross-classified table are very sparse, and $GFfit^{(ij)}_\perp$ generally has higher empirical power than the other diagnostics for detecting model lack of fit in bivariate tables. An application with a large number of variables will produce a large number of lack-of-fit statistics on bivariate tables, and it is important for applied researchers to use a multiple decision rule to maintain validity when identifying unusually large values.

A primary purpose of models for cross-classified tables is to explain the association among the variables. $GFfit^{(ij)}_\perp$ is a member of a class of general lack-of-fit statistics for marginal tables

Table 7: GFfit Statistics for Depression Example

(ij)	$GFfit_{\perp}^{(ij)}$	p-val	FDR	$\bar{\bar{X}}^2_{ij}$	p-val	BY	FDR
(12)	**25.948**	**0.002**	**0.044**	**16.583**	**0.020**	0.777	0.213
(13)	6.050	0.735	0.908	2.669	0.914	1.000	0.960
(14)	9.628	0.381	0.720	11.434	0.121	1.000	0.386
(15)	10.179	0.336	0.706	8.129	0.321	1.000	0.547
(16)	7.341	0.602	0.877	8.678	0.277	1.000	0.547
(17)	3.861	0.920	0.946	5.282	0.626	1.000	0.773
(23)	3.415	0.946	0.946	1.544	0.981	1.000	0.981
(24)	12.408	0.191	0.574	7.831	0.348	1.000	0.547
(25)	13.979	0.123	0.517	10.816	0.147	1.000	0.386
(26)	7.106	0.626	0.877	7.645	0.365	1.000	0.547
(27)	14.816	0.096	0.505	**14.850**	**0.038**	0.969	0.266
(34)	**20.161**	**0.017**	0.119	12.176	0.095	1.000	0.386
(35)	**20.397**	**0.016**	0.119	**18.525**	**0.010**	0.751	0.206
(36)	9.285	0.411	0.720	4.680	0.699	1.000	0.776
(37)	4.812	0.850	0.946	8.186	0.316	1.000	0.547
(45)	13.166	0.155	0.543	6.186	0.518	1.000	0.726
(46)	4.465	0.878	0.946	5.577	0.590	1.000	0.773
(47)	10.539	0.309	0.706	10.958	0.140	1.000	0.386
(56)	7.190	0.617	0.877	4.653	0.702	1.000	0.776
(57)	11.173	0.264	0.693	13.683	0.057	1.000	0.300
(67)	6.700	0.668	0.877	8.503	0.290	1.000	0.547

that can be applied broadly to models for cross-classified variables. Other members of the class include $\bar{\bar{X}}_{ij}^2$ and M_{ij}, although neither $\bar{\bar{X}}_{ij}^2$ nor M_{ij} can be applied to binary cross-classified variables with estimated parameters because the degrees of freedom would become negative. The $GFfit_{\perp}^{(ij)}$ also differ from $\bar{\bar{X}}_{ij}^2$ and M_{ij} in that they are obtained jointly, in a sequential manner with known asymptotic joint distribution function. Because of the sequential feature, using substantive theory to plan the order of variables may be fruitful in terms of power to detect lack of fit, similar to selecting an order for variables in a multiple regression. When fitting an IRT model, for example, items that might be associated with multiple factors or newly introduced items could be placed at the top in the order. However, in many applications, a substantive theory will not be available, and ordering of the variables will be arbitrary. Then conditional on (i, j) and a given false model, even though this particular $GFfit_{\perp}^{(ij)}$ might have modestly higher or lower one-at-a-time power to detect lack of fit with a different variable order, applied researchers can nevertheless still expect that after a correction for multiple testing that takes advantage of the asymptotic independence property for a set of $GFfit_{\perp}^{(ij)}$, there will be higher probability to detect existing lack of fit using $GFfit_{\perp}^{(ij)}$ than other diagnostics for cross-classified tables. Furthermore, when diagnostics have different properties, applied researchers may find it useful to examine multiple lack-of-fit diagnostics, some of which are not order dependent. SAS PROC MIXED, for example, provides three versions of model diagnostic residuals, including order-dependent Cholesky residual (scaled residual), marginal residual, and conditional residual (Schabenberger, 2005). For cross-classified tables, $GFfit_{\perp}^{(ij)}$, $GFfit^{(ij)}$, and $\bar{\bar{X}}_{ij}^2$ could be examined, keeping in mind the need to maintain Type I error level. Based on simulation results, M_{ij} is not recommended as a lack-of-fit diagnostic for the IRT model because it has very low power for detecting misspecification of variables associations, which is the most common misspecification in applications of IRT models.

Calculation of $GFfit_{\perp}^{(ij)}$ requires careful computation because there is high collinearity among the columns of matrix $\mathbf{H}$. Computing $GFfit_{\perp}^{(ij)}$ by using the sum of squares from an orthogonal regression as discussed in Section 3.2 has high numerical stability and reliability. Because calculation of $GFfit_{\perp}^{(ij)}$ requires a large amount of memory as the number of response variables increases, it is more suitable for applications such as attitude surveys, personality and clinical psychological assessments and medical applications (Breinegaard, Rabe-Hesketh & Skrondal (2017), rather than educational testing where the number of items may be sizable. For a composite null hypothesis, a lack-of-fit index somewhat similar to $GFfit_{\perp}^{(ij)}$ could be calculated by simply ignoring the matrix $\boldsymbol{G}$, which would reduce the demand for memory substantially, but the indices would not have the theoretical properties for probability distribution and degrees of freedom obtained in Section 4.1.

High-dimensional cross-classified tables are often found in social science applications, and in this paper, $GFfit_{\perp}^{(ij)}$ was employed to identify lack of fit among second-order marginals for the IRT model applied to symptoms of depression with four response categories. In this application, the performance of $GFfit_{\perp}^{(ij)}$ and $\bar{\bar{X}}_{ij}^{2}$ for identifying lack of fit were similar, but the asymptotic independent $GFfit_{\perp}^{(ij)}$ statistics had a substantial advantage when controlling the Type I error rate. This example also showed the advantage that an omnibus test formed on a partial sum of $GFfit_{\perp}^{(ij)}$ had higher power to reject a false null model than X_{PF}^{2}. Although the application was to the IRT model, $GFfit_{\perp}^{(ij)}$ can be applied to many other models. For example, a closely related approach has been used as a diagnostic for models applied to cross-classified longitudinal variables (Breinegaard, Rabe-Hesketh and Skrondal, 2017).

More simulation studies are needed to determine if an asymptotic chi-square approximation would be valid for $GFfit_{\perp}^{(ijk)}$ and other higher-order $GFfit_{\perp}$ that could be used as lack-of-fit diagnostics. The asymptotic approximation may be less reliable in higher-order marginal tables due to sparseness among the marginal frequencies. If the asymptotic approximation is not valid, the parametric bootstrap can be used to obtain a p-value for $GFfit_{\perp}^{(ijk)}$ and for other higher-order $GFfit_{\perp}$ as well. Another important area for future research in IRT applications is a comparison of the approach using components of Pearson's statistic to the Lagrange multiplier approach of Glas (1999). A study for this comparison using binary manifest variables is now underway and will also include $\bar{X}_{ij}^{2}$ and standardized bivariate residuals. More studies are also needed to extend application of $GFfit_{\perp}^{(ij)}$ to other models for cross-classified tables from longitudinal studies.

9. Appendix A

9.1. First- and Second-order Marginals

Define $\mathbf{H}_{[1]} = \boldsymbol{V}'$. Then, under the model $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$, the first-order marginal proportion for variable Y_i can be defined as

$$\begin{aligned} &\pi^{(i)}(a;\boldsymbol{\beta}) = \text{Prob}(Y_i = a|\boldsymbol{\beta}) = \sum_s h_{\ell s}\pi_s(\boldsymbol{\beta}) = \boldsymbol{h}_\ell^{'}\boldsymbol{\pi}(\boldsymbol{\beta}), \\ &a = 2,\ldots,c;\ \ell = (c-1)(i-1)+a-1;\ s = 1,\ldots,T, \end{aligned} \tag{9.1}$$

where $h_{\ell s}$ is an element of the $q(c-1)$ by T matrix $\mathbf{H}_{[1]}$, and where $\boldsymbol{h}'_\ell$ is row ℓ of matrix $\mathbf{H}_{[1]}$. The true first-order marginal proportion is given by

$$\pi^{(i)}(a) = \text{Prob}(Y_i = a) = \sum_s h_{\ell s}\pi_s = \boldsymbol{h}_\ell^{'}\boldsymbol{\pi}\ . \tag{9.2}$$

The second-order marginal proportion for variables Y_i and Y_j under the model can be defined as

$$\pi^{(ij)}(a,b;\boldsymbol{\beta}) = \text{Prob}(Y_i = a, Y_j = b|\boldsymbol{\beta}) = \sum_s h_{ms}h_{\ell s}\pi_s(\boldsymbol{\beta}) = (\boldsymbol{h}'_m \circ \boldsymbol{h}_\ell^{'})\boldsymbol{\pi}(\boldsymbol{\beta}), \tag{9.3}$$

where $i = 1,\cdots,q-1$; $j = i,\cdots,q$; $m = (c-1)(i-1)+a-1$; $\ell = (c-1)(j-1)+b-1$; $a = 2,\ldots,c$; $b = 2,\ldots,c$; and $\boldsymbol{h}'_m \circ \boldsymbol{h}_\ell^{'}$ represents the Hadamard product (Magnus & Neudecker, 1999) of rows m and ℓ from matrix $\mathbf{H}_{[1]}$. Then the true second-order marginal proportion is given by

$$\pi^{(ij)}(a,b) = \text{Prob}(Y_i = a, Y_j = b) = \sum_s h_{ms}h_{\ell s}\pi_s = (\boldsymbol{h}'_m \circ \boldsymbol{h}_\ell^{'})\boldsymbol{\pi}\ . \tag{9.4}$$

9.2. *V* Matrix

The matrix $\boldsymbol{V}$ has $(c-1)$ kernel patterns, each of dimension c. For $c = 2$, the kernel pattern is $\boldsymbol{f}_1 = (0,\ 1)^{'}$, and for $c = 3$, the kernel patterns are $\boldsymbol{f}_1 = (0,\ 0,\ 1)^{'}$ and $\boldsymbol{f}_2 = (0,\ 1,\ 0)^{'}$. In general, the kernel patterns, as columns, form a $(c-1)$ by $(c-1)$ matrix $\boldsymbol{J} - \boldsymbol{I}$ adjoined to a row of zeros. The matrix $\boldsymbol{V}$ can be generated by Kronecker products of the kernel patterns with the vector $\mathbf{1}_c$, which is a vector of length c where each element is 1. The pattern of columns is

$$\begin{aligned} \boldsymbol{V} = (&\boldsymbol{f}_1 \otimes (\mathbf{1}_c \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c),\ \boldsymbol{f}_2 \otimes (\mathbf{1}_c \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c) \ldots \boldsymbol{f}_{c-1} \otimes (\mathbf{1}_c \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c), \\ &\mathbf{1}_c \otimes (\boldsymbol{f}_1 \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c),\ \mathbf{1}_c \otimes (\boldsymbol{f}_2 \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c), \ldots \mathbf{1}_c \otimes (\boldsymbol{f}_{c-1} \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c), \ldots \\ &\mathbf{1}_c \otimes (\mathbf{1}_c \cdots \otimes \mathbf{1}_c \otimes \boldsymbol{f}_1),\ \mathbf{1}_c \otimes (\mathbf{1}_c \cdots \otimes \mathbf{1}_c \otimes \boldsymbol{f}_2), \ldots \mathbf{1}_c \otimes (\mathbf{1}_c \cdots \otimes \mathbf{1}_c \otimes \boldsymbol{f}_{c-1})) \end{aligned} \tag{9.5}$$

With $q = 3$ and $c = 2$, $\boldsymbol{V}$ is generated as

$$\boldsymbol{V} = \Big(\boldsymbol{f}_1 \otimes (\mathbf{1}_2 \otimes \mathbf{1}_2),\ \mathbf{1}_2 \otimes (\boldsymbol{f}_1 \otimes \mathbf{1}_2),\ (\mathbf{1}_2 \otimes \mathbf{1}_2) \otimes \boldsymbol{f}_1\ .\Big) \tag{9.6}$$

For $q = 3$ and $c = 3$, $\boldsymbol{V}$ is generated as

$$\begin{aligned} \boldsymbol{V} = (&\boldsymbol{f}_1 \otimes (\mathbf{1}_3 \otimes \mathbf{1}_3),\ \boldsymbol{f}_2 \otimes (\mathbf{1}_3 \otimes \mathbf{1}_3),\ \mathbf{1}_3 \otimes (\boldsymbol{f}_1 \otimes \mathbf{1}_3), \\ &\mathbf{1}_3 \otimes (\boldsymbol{f}_2 \otimes \mathbf{1}_3),\ (\mathbf{1}_3 \otimes \mathbf{1}_3) \otimes \boldsymbol{f}_1,\ (\mathbf{1}_3 \otimes \mathbf{1}_3) \otimes \boldsymbol{f}_2), \end{aligned} \tag{9.7}$$

and for $q = 4$ and $c = 4$, $\boldsymbol{V}$ is generated as

$$\begin{aligned} \boldsymbol{V} = (&\boldsymbol{f}_1 \otimes (\mathbf{1}_4 \otimes \mathbf{1}_4 \otimes \mathbf{1}_4),\ \boldsymbol{f}_2 \otimes (\mathbf{1}_4 \otimes \mathbf{1}_4 \otimes \mathbf{1}_4),\ \boldsymbol{f}_3 \otimes (\mathbf{1}_4 \otimes \mathbf{1}_4 \otimes \mathbf{1}_4), \\ &\mathbf{1}_4 \otimes (\boldsymbol{f}_1 \otimes \mathbf{1}_4 \otimes \mathbf{1}_4),\ \mathbf{1}_4 \otimes (\boldsymbol{f}_2 \otimes \mathbf{1}_4 \otimes \mathbf{1}_4),\ \mathbf{1}_4 \otimes (\boldsymbol{f}_3 \otimes \mathbf{1}_4 \otimes \mathbf{1}_4), \\ &\mathbf{1}_4 \otimes (\mathbf{1}_4 \otimes \boldsymbol{f}_1 \otimes \mathbf{1}_4),\ \mathbf{1}_4 \otimes (\mathbf{1}_4 \otimes \boldsymbol{f}_2 \otimes \mathbf{1}_4),\ \mathbf{1}_4 \otimes (\mathbf{1}_4 \otimes \boldsymbol{f}_3 \otimes \mathbf{1}_4), \\ &\mathbf{1}_4 \otimes (\mathbf{1}_4 \otimes \mathbf{1}_4 \otimes \boldsymbol{f}_1),\ \mathbf{1}_4 \otimes (\mathbf{1}_4 \otimes \mathbf{1}_4 \otimes \boldsymbol{f}_2),\ \mathbf{1}_4 \otimes (\mathbf{1}_4 \otimes \mathbf{1}_4 \otimes \boldsymbol{f}_3)) \end{aligned} \tag{9.8}$$

9.3. $\boldsymbol{H}$ Matrix

For second-order marginals, a $(c-1)^2 q(q-1)/2$ by c^q matrix $\mathbf{H}_{[2]}$ can be defined by forming Hadamard products among the columns $\boldsymbol{V}$:

$$\boldsymbol{H}_{[2]} = \begin{pmatrix} (\boldsymbol{v}_1 \circ \boldsymbol{v}_c)' \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_{c+1})' \\ \vdots \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_{q(c-1)})' \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_c)' \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_{c+1})' \\ \vdots \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_{q(c-1)})' \\ \vdots \\ (\boldsymbol{v}_{c-1} \circ \boldsymbol{v}_c)' \\ (\boldsymbol{v}_{c-1} \circ \boldsymbol{v}_{c+1})' \\ \vdots \\ (\boldsymbol{v}_{c-1} \circ \boldsymbol{v}_{q(c-1)})' \\ \vdots \\ (\boldsymbol{v}_c \circ \boldsymbol{v}_{(q-1)(c-1)})' \\ \vdots \\ (\boldsymbol{v}_c \circ \boldsymbol{v}_{q(c-1)})' \\ \vdots \\ (\boldsymbol{v}_{(q-1)(c-1)} \circ \boldsymbol{v}_{(q-1)(c-1)+1})' \\ \vdots \\ (\boldsymbol{v}_{(q-1)(c-1)} \circ \boldsymbol{v}_{q(c-1)})' \end{pmatrix} \tag{9.9}$$

where $\boldsymbol{v}_\ell$ represents column ℓ of matrix $\boldsymbol{V}$. To place the marginals in a convenient order, the columns of $\mathbf{H}$ from the products $(\boldsymbol{v'}_m \circ \boldsymbol{v}_\ell^{\prime})$ are arranged in lexicographical order. If $c = 2$,

$$\mathbf{H}_{[2]} = \begin{pmatrix} (\boldsymbol{v}_1 \circ \boldsymbol{v}_2)' \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_3)' \\ \vdots \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_q)' \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_3)' \\ \vdots \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_q)' \\ \vdots \\ (\boldsymbol{v}_{q-1} \circ \boldsymbol{v}_q)' \end{pmatrix}, \tag{9.10}$$

If $q = 3$ and $c = 4$ categories, $\mathbf{H}_{[2]}$ is a 27 by 64 matrix:

$$\mathbf{H}_{[2]} = \begin{pmatrix} (\boldsymbol{v}_1 \circ \boldsymbol{v}_4)' \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_5)' \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_6)' \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_4)' \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_5)' \\ (\boldsymbol{v}_2 \circ \boldsymbol{v}_6)' \\ (\boldsymbol{v}_3 \circ \boldsymbol{v}_4)' \\ (\boldsymbol{v}_3 \circ \boldsymbol{v}_5)' \\ (\boldsymbol{v}_3 \circ \boldsymbol{v}_6)' \\ \vdots \\ (\boldsymbol{v}_4 \circ \boldsymbol{v}_7)' \\ (\boldsymbol{v}_4 \circ \boldsymbol{v}_8)' \\ (\boldsymbol{v}_4 \circ \boldsymbol{v}_9)' \\ (\boldsymbol{v}_5 \circ \boldsymbol{v}_7)' \\ (\boldsymbol{v}_5 \circ \boldsymbol{v}_8)' \\ (\boldsymbol{v}_5 \circ \boldsymbol{v}_9)' \\ (\boldsymbol{v}_6 \circ \boldsymbol{v}_7)' \\ (\boldsymbol{v}_6 \circ \boldsymbol{v}_8)' \\ (\boldsymbol{v}_6 \circ \boldsymbol{v}_9)' \end{pmatrix} \tag{9.11}$$

9.4. $\boldsymbol{M}$ *Matrix*

Consider c kernel patterns $\boldsymbol{f}_\ell$, $\ell = 1, 2, \ldots, c$ that form, as columns, a c by c matrix $\boldsymbol{J} - \boldsymbol{I}$, and consider the cq by T matrix $\boldsymbol{U}$ given by

$$\begin{aligned} \boldsymbol{U} = (&\boldsymbol{f}_1 \otimes (\mathbf{1}_c \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c),\ \boldsymbol{f}_2 \otimes (\mathbf{1}_c \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c) \ldots \boldsymbol{f}_c \otimes (\mathbf{1}_c \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c), \\ &\mathbf{1}_c \otimes (\boldsymbol{t}_1 \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c),\ \mathbf{1}_c \otimes (\boldsymbol{f}_2 \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c), \ldots \mathbf{1}_c \otimes (\boldsymbol{f}_c \otimes \mathbf{1}_c \cdots \otimes \mathbf{1}_c), \ldots \\ &\mathbf{1}_c \otimes (\mathbf{1}_c \cdots \otimes \mathbf{1}_c \otimes \boldsymbol{f}_1),\ \mathbf{1}_c \otimes (\mathbf{1}_c \cdots \otimes \mathbf{1}_c \otimes \boldsymbol{f}_2), \ldots \mathbf{1}_c \otimes (\mathbf{1}_c \cdots \otimes \mathbf{1}_c \otimes \boldsymbol{f}_c)) \end{aligned} \tag{9.12}$$

Then a $c^2 q(q-1)/2$ by T matrix $\boldsymbol{M}$ is defined using Hadamard products among the columns of $\boldsymbol{U}$:

$$
\boldsymbol{M}_{[2]} = \begin{pmatrix}
(\boldsymbol{u}_1 \circ \boldsymbol{u}_{c+1})' \\
(\boldsymbol{u}_1 \circ \boldsymbol{u}_{c+2})' \\
\vdots \\
(\boldsymbol{u}_1 \circ \boldsymbol{u}_{qc})' \\
(\boldsymbol{u}_2 \circ \boldsymbol{u}_{c+1})' \\
(\boldsymbol{u}_2 \circ \boldsymbol{u}_{c+2})' \\
\vdots \\
(\boldsymbol{u}_2 \circ \boldsymbol{u}_{qc})' \\
\vdots \\
(\boldsymbol{u}_c \circ \boldsymbol{u}_{c+1})' \\
(\boldsymbol{u}_c \circ \boldsymbol{u}_{c+2})' \\
\vdots \\
(\boldsymbol{u}_c \circ \boldsymbol{u}_{qc})' \\
\vdots \\
(\boldsymbol{u}_{c+1} \circ \boldsymbol{u}_{2c+1})' \\
\vdots \\
(\boldsymbol{u}_{c+1} \circ \boldsymbol{u}_{qc})' \\
\vdots \\
(\boldsymbol{u}_{(q-2)c+1} \circ \boldsymbol{u}_{(q-1)c+1})' \\
\vdots \\
(\boldsymbol{u}_{(q-1)c} \circ \boldsymbol{u}_{qc)})'
\end{pmatrix} \tag{9.13}
$$

Linear dependencies exist among the columns of $\boldsymbol{U}$; $\boldsymbol{V}$ from Section 2 consists of the linear independent columns of $\boldsymbol{U}$ such that $\boldsymbol{V} = \boldsymbol{U}\boldsymbol{A}$, where $\boldsymbol{A} = \boldsymbol{I} \otimes (\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_c)$.

The $(c-1)^2 q(q-1)/2$ by $c^2 q(q-1)/2$ matrix $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{I}_\ell \otimes \boldsymbol{A}^{(1)} & \mathbf{0}_{q(c-1)\text{ x }qc} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_\ell \otimes \boldsymbol{A}^{(1)} & \mathbf{0}_{g(c-1)\text{ x}(q-1)c} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{I}_{2\text{x}2} \otimes \boldsymbol{A}^{(1)} & \mathbf{0}_{2(c-1)\text{x}c} \end{pmatrix} \tag{9.14}$$

where $\ell = (q-d)(c-1)$ for column d of $\boldsymbol{A}$, and $\boldsymbol{A}^{(1)} = (\boldsymbol{I}_{(c-1)} \vdots \mathbf{0})$.

References

Afifi, A.A., and Clark, V. (1984). Computer-Aided Multivariate Analysis, Lifetime Learning Publications, Belmont: CA

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

Asparouhov, T., and Muthén, B. (2010). Simple Second Order Chi-Square Correction. Mplus Technical Report. (https://www.statmodel.com/download/WLSMV_new_chi21.pdf)

Bartholomew (1987). *Latent Variable Models an Factor Analysis.* New York: Oxford University Press.

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.

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

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

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

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

Dassanayake, M., Reiser, M., Zhu, J. (2016) Power calculations for statistics based on orthogonal components of Pearson's chi-square. In *JSM Proceedings*, Biometrics Section, Alexandria VA, American Statistical Association, pp. 1079-1093.

Eubank, R. L. (1997). Testing goodness of fit with multinomial data. *Journal of the American Statistical Association, 92*, no 439, 1084-1093.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525-546.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*, no. 3, 273-294.

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

Houseman, E. A., Ryan, L. M. and Coull, B. A. (2004). Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. *Journal of the American Statistical Association*, *99*, no 486, 383-394.

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

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M., and Thiebaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*, *51*, no. 10, 5142-5154.

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

Kelderman, H. (1984). Log-linear Rasch model tests. *Psychometrika*, *49*, 223-245.

Koehler, K. J. (1986). Goodness-of-fit tests for log-linear models is sparse contingency tables. *Journal of the American Statistical Association,* June, 336-344.

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

Lancaster, H. O. (1969). *The chi-squared distribution.* Wiley, New York.

Liu, Y. and Maydeu-Olivares, O. (2014). Identifying the Source of Misfit in Item response Theory Models. *Multivariate Behavioral Research* 49, 354-371.

Maydeu-Olivares, A. and Joe, H. (2005). Limited- and Full-Information Estimation and Goodness-of-Fit Testing in 2^n Contingency Tables: A Unified Framework. *Journal of the American Statistical Association*, 100(471), 1009-1020.

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

Maydeu-Olivares, A. and Liu, Y. (2012). Local dependence Diagnostics in IRT Modeling of Binary data. *Educational and Psychological Measurement* 73(2), 254-274.

Maydeu-Olivares, A., & Montaño, R. (2013). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika, 78,* 116-133.

Mavridis, D., and Moustaki, I. and Knott, M. (2007). Goodness-of-fit measures for latent variable models for binary data. In S.-Y- Lee (Ed.) *Handbook of latent variable and related models* 135-161, Amsterdam, The Netherlands, Elsevier.

Mirvaliev, M. (1987). The components of chi-squared statistics for goodness-of-fit tests. *Journal of Soviet Mathematics, 38*, 2357-2363. https://doi.org/10.1007/BF01095078.

National Institute of Mental Health (NIMH) (2019). *Results from the 2017 National Survey on Drug Use and Mental Health.* https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2017/NSDUHDetailedTabs2017.htm#tab8-56A

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

Reiser, M. (1989). An application of the item response model to psychiatric epidemiology. *Sociological Methods & Research*, *18*, 66-103.

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

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

Reiser, M (2019). Goodness-of-fit testing in sparse contingency tables when the number of variables is large. *WIRES Computational Statistics,* 11(6), e1470.

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

Salomaa, H. (1990). Factor analysis of dichotomous data. Helsinki, Findland: Statistical Society.

Schabenberger, O. (2005). Mixed model influence diagnostics. *SAS Users Group International Conference (SUGI)*, 189-29.

Sharma, S. (1995). *Applied Multivariate Techniques*. Wiley, New York.

United States Department of Health and Human Services, National Institute of Mental Health. *Epidemiological Catchment Area (ECA) Survey of Mental Disorders, Wave I (Household), 1980-1985: [United States].* Rockville, MD: U.S. Dept of Health and Human Services, National Institute of Mental Health [producer], 1985. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1991. doi:10.3886/ICPSR08993.v1

Verbeke, G., and Molenberghs, G. (2009). *Linear mixed models for longitudinal data.* Springer Science & Business Media, New York.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5(2)*, 245-262