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

Mark Reiser [*] Silvia Cagnone [†] Junfei Zhu[‡]

Abstract

The Pearson and likelihood ratio statistics are commonly used to test goodness of fit for models applied to data from a multinomial distribution. When data are from a table formed by the cross-classification of a large number of variables, the common statistics may have low power and inaccurate Type I error level due to sparseness in the cells of the table. For the cross-classification of a large number of ordinal manifest variables, it has been proposed to assess model fit by using the *GFfit* statistic as a diagnostic to examine the fit on two-way subtables, and the asymptotic distribution of the *GFfit* statistic has been previously obtained. In this paper, a new version of the *GFfit* statistic is proposed by decomposing the Pearson statistic from the full table into orthogonal components defined on lower-order marginal distributions and then defining the *GFfit* statistic as a sum of a subset of these components. The new version of the *GFfit* statistic also extends the diagnostic to higher-order tables so that the *GFfit* statistics sum to the Pearson statistic. Simulation results and an application of the new *GFfit* statistic as a diagnostic for a latent variable model are presented.

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

1. Introduction

A test of fit for a multinomial model commonly challenges the null hypothesis $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$, where $\boldsymbol{\pi}$ is a vector of multinomial probabilities, and $\boldsymbol{\pi}(\boldsymbol{\beta})$ is a vector of the multinomial probabilities as a function of parameters in the vector $\boldsymbol{\beta}$. When the model parameters $\boldsymbol{\beta}$ are unknown and estimated, the null hypothesis $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$ is often tested with the Pearson-Fisher statistic:

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

where

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

and where

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

[*] School of Mathematical and Statistical Sciences, Arizona State University, USA

[†] Department of Statistical Sciences, University of Bologna, Italy

[‡] School of Mathematical and Statistical Sciences, Arizona State University, USA

The goodness-of-fit test based on Pearson's chi-squared statistic is sometimes considered to be an omnibus test that gives little guidance to the source of poor fit when the null hypothesis is rejected. It has also been recognized that the omnibus test can often be outperformed by focused or directional tests of lower order.

When data are from a table formed by the cross-classification of a large number of variables, the Pearson's chi-square and the likelihood ratio statistic may have low power and inaccurate Type I error level due to sparseness. Several statistics have been proposed that marginal distributions of the joint variables rather than the joint distribution. These statistics have mostly been applied to item response models or factor analysis of categorical variables and have very good performance for Type I error rate and power when the data table is formed from a moderate number of variables. The motivation for these focused tests has been that the statistics formed on lower-order marginals usually overcome the deleterious effects of sparseness, and the tests may often have higher power under commonly encountered circumstances.

The *GFfit* statistic was proposed by Joreskog and Moustaki (2001) as a diagnostic to aid in finding the source of model lack of fit. In this paper, a new version of the *GFfit* statistic is proposed by decomposing the Pearson statistic from the full table into orthogonal components defined on lower-order marginal distributions and then defining the *GFfit* statistic as a sum of a subset of these components. The new version of the *GFfit* statistic also extends the diagnostic to higher-order tables so that the *GFfit* statistics sum to the Pearson statistic. Simulation results and an application of the new *GFfit* statistic as a diagnostic for a latent variable model are presented.

2. Marginal Proportions

A traditional method such as Pearson's statistic uses the joint frequencies to calculate goodness of fit for a model that has been fit to a cross-classified table. This section presents a transformation from joint proportions or frequencies to marginal proportions. Marginal proportions are used to develop test statistics presented in Section 3.

2.1 First- and Second-Order Marginals

The relationship between joint proportions and first- and second-order marginals can be shown by using zeros and 1's to code the levels of categorical response random variables, $Y_i, i = 1, 2, \ldots, q$, where Y_i has $c \geq 2$ response categories. If $c = 2$, then a q-dimensional vector of zeros and 1's, sometimes called a response pattern, will indicate a specific cell from the contingency table formed by the cross-classification of q response variables. For dichotomous response variables, a response pattern is a sequence of zeros and 1's with length q. 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 matrix with response patterns as rows.

For $q = 3$ and $c = 2$,

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

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

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

and the true first-order marginal proportion is given by

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

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 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 $\mathbf{H} \equiv \mathbf{H}_{[t:u]}$.

Matrix $\mathbf{H}_{[1]}$ can be defined from matrix $\boldsymbol{V}$ such that

$$\mathbf{H}_{[1]} = \boldsymbol{V}' \ .$$

If $c = 2$, the second-order marginal proportion under the model for variables Y_i and Y_j can be defined as

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

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

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

For second-order marginals, where $\ell = i - j + \sum_{0<r<j}(q-r)$, element ℓs of $\mathbf{H}_{[2]}$ is given by

$$\left[\mathbf{H}_{[2]}\right]_{\ell s} = \begin{cases} 1 & if\ v_{is} = v_{js} = 1 \\ 0 & otherwise. \end{cases}$$

Alternatively, matrix $\mathbf{H}_{[2]}$ can be defined by forming Hadamard products among the columns of the matrix $\boldsymbol{V}$:

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

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

Generally, $\boldsymbol{V}$ has $q(c-1)$ columns. For 3 variables with 3 categories $\mathbf{H}_{[1]} = \boldsymbol{V}'$, where

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

If $q = 3$ and $c = 3$ categories, $\mathbf{H}_{[2]}$ is an 18 by 27 matrix:

$$\mathbf{H}_{[2]} = \begin{pmatrix} (\boldsymbol{v}_1 \circ \boldsymbol{v}_3)' \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_4)' \\ \vdots \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_5)' \\ (\boldsymbol{v}_1 \circ \boldsymbol{v}_6)' \\ \vdots \\ (\boldsymbol{v}_3 \circ \boldsymbol{v}_5)' \\ \vdots \\ (\boldsymbol{v}_{i(c-1)} \circ \boldsymbol{v}'_{j(c-1)} \end{pmatrix}$$

2.2 Higher-Order Marginals

For higher-order marginal proportions, the columns of $\mathbf{H}$ are Hadamard products among the columns of $\boldsymbol{V}$. 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,

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

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 one-to-one mapping from joint proportions to the set of $(2^q - 1)$ marginal proportions:

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

where

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

is the vector of marginal proportions (Bartholomew, 1987).

2.3 Residuals

Define the unstandardized residual $\mathrm{r}_s = \hat{\mathrm{p}}_s - \pi_s(\hat{\boldsymbol{\beta}})$, and denote the vector of unstandardized residuals as $\mathbf{r}$ with element r_s .

A vector of simple residuals for marginals of any order may be defined such that

$$\boldsymbol{e} = \mathbf{H}(\hat{\mathbf{p}} - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})) = \mathbf{Hr},$$

and a vector, $\boldsymbol{\xi}$, of differences between the marginals specified by the relevant model and the true population marginals may be defined for marginals of any order such that

$$\boldsymbol{\xi} = \mathbf{H}(\boldsymbol{\pi} - \boldsymbol{\pi}(\boldsymbol{\beta})).$$

3. Testing Fit on Marginal Distributions

3.1 Linear Combinations of Joint Frequencies

A traditional composite null hypothesis for a test of fit on a multinomial model is $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$. Linear combinations of $\boldsymbol{\pi}$ may be tested under the null hypothesis $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$, or equivalently $H_o\colon \boldsymbol{\xi}_{[t:u]} = \mathbf{0}$. $\mathbf{H}$ may specify linear combinations that form marginal proportions as defined in the previous section. For q variables each with c categories, if $\mathbf{H}$ has rank $R = c^q - g - 1$, where g is the number of unknown model parameters to be estimated, then $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$ is equivalent to $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$, given some additional conditions. If $\mathbf{H}$ has rank less than R, then $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$ specifies a test that is sometimes known as limited-information, but focused test would be a better description. When rank($\mathbf{H}$) $< R$, null hypotheses $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$ may represent components of the null hypothesis $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$, with corresponding partition of degrees of freedom. If the null hypothesis $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$ is a component of the null hypothesis $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$, then "Reject $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$" is a sufficient but not necessary condition for "Reject

$H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$"; "Do not reject $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$" is a necessary but not sufficient condition for "Do not reject $H_o\colon \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$". "Do not reject $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$" is not a sufficient condition because it is possible that lack of fit may be manifest only in a direction not represented in $\mathbf{H}_{[t:u]}$. For q variables each with c categories, there are $T - g - 1$ independent orthogonal components, which may be summed, for testing null hypotheses $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$.

3.2 Test Statistic

Linear combinations of $\boldsymbol{\pi}$ may be tested under the null hypothesis $H_o\colon \mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\beta})$ by using the quadratic form statistic

$$X^2_{[t:\,u]} = \boldsymbol{e}'\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}^{-1}\boldsymbol{e}$$

where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}} = n^{-1}\boldsymbol{\Omega}_{\boldsymbol{e}}$ with $\boldsymbol{\Omega}_{\boldsymbol{e}}$ evaluated at the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$, and where

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

$\mathbf{H} = \mathbf{H}_{[1:2]}$ produces $X^2_{[1:2]}$, and $\mathbf{H} = \mathbf{H}_{[2]}$ produces $X^2_{[2]}$. The limiting distribution is chi-square with degrees of freedom determined by the number linearly independent columns in $\mathbf{H}$. If $c = 2$, $X^2_{[1:2]}$ is the statistic from Reiser (1996), and $X^2_{[2]}$ is the statistic from Reiser and Lin (1999). Also when $c = 2$, it is generally the case that $df = q(q+1)/2$ for $X^2_{[1:2]}$, and $df = q(q-1)/2$ for $X^2_{[2]}$, although some models may involve certain theoretically known linear dependencies among the marginals so that these expressions for degrees of freedom would be adjusted accordingly. Cagnone and Mignani (2007) extended the statistic to manifest variables with two or more categories. Furthermore, if $\mathbf{H}_{[1:q;-g]}$ represents $\mathbf{H}_{[1:q]}$ with g columns deleted so that $\mathbf{H}_{[1:q;-g]}$ has rank $R = T - g - 1$, then $X^2_{PF} = X^2_{[1:q;-g]}$, where $X^2_{[1:q;-g]}$ is defined with $\mathbf{H} = \mathbf{H}_{[1:q;-g]}$ (Reiser, 2008). $X^2_{[t:u]}$ is essentially a version of the score statistic from Rayner and Best (1989).

4. Extended $GFfit^{ij}$ Statistic Using Orthogonal Components

4.1 Orthogonal Components

Consider the $T - g - 1$ by c^q matrix $\mathbf{H}^* = \boldsymbol{F}'\mathbf{H}_{[1:q;-g]}$. $\mathbf{H}^*$ has full row rank. $\boldsymbol{F}$ 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 $\boldsymbol{\Omega}_{\boldsymbol{e}}$. Premultiplication by $(\boldsymbol{C}')^{-1}$ orthonormalises the matrix $\mathbf{H}_{[1:q;-g]}$ in the matrix $D(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{G}'$.

$$X^2_{PF} = X^2_{[1:q;-g]} = n^{-1}\mathbf{r}'(\widehat{\mathbf{H}}^*)'\widehat{\mathbf{H}}^*\mathbf{r}$$

where $\widehat{\mathbf{H}}^* = \mathbf{H}^*(\hat{\boldsymbol{\beta}})$.

Define

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

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

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

$\widehat{\mathbf{H}}^*\mathbf{r}$ has asymptotic covariance matrix $\boldsymbol{F}'\boldsymbol{\Omega}_{\boldsymbol{e}}\boldsymbol{F} = \boldsymbol{I}_{T-g-1}$ The elements $\hat{\gamma}_j^2$ are asymptotically independent χ^2_1 random variables (Reiser, 2008).

Using sequential sum of squares, redefine:

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

Perform the regression of $\boldsymbol{z}$ on the columns of $\boldsymbol{H}'$:

$$\boldsymbol{z} = \mathrm{H}\boldsymbol{\theta}$$

Then,

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}\widehat{\boldsymbol{W}}\mathrm{H}')^{-1}\mathbf{H}\widehat{\boldsymbol{W}}\mathbf{u}$$

where $\mathbf{u} = \sqrt{n}\mathbf{r}$, $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{D}}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{D}}^{-\frac{1}{2}} = \widehat{\boldsymbol{D}}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{D}}^{-\frac{1}{2}}$, and $\boldsymbol{D} = diag(\boldsymbol{\pi}(\boldsymbol{\beta}))$. $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\beta}) = (\boldsymbol{I} - \boldsymbol{\pi}^{\frac{1}{2}}(\boldsymbol{\pi}^{\frac{1}{2}})' - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}')$ is idempotent.

Let $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{D}}^{-\frac{1}{2}}\mathbf{H}'$. Then

$$\hat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{M}}'\widehat{\boldsymbol{M}})^{-1}\widehat{\boldsymbol{M}}'\boldsymbol{z}$$

$\hat{\boldsymbol{\gamma}}_j^2$, $j = 1, T$ are the sequential SS from this regression. $\boldsymbol{\gamma} = \boldsymbol{C}'\boldsymbol{\theta}$ are the orthogonal coefficients. Sequential SS from Goodnight's Sweep operator are very accurate numerically (Goodnight, 1978; SAS PROC REG).

4.2 Extended $GFfit$ Statistic

Joreskog and Moustaki (2001) proposed the following *GFfit* statistic:

$$GFfit^{(ij)} = n\Sigma_{ab}\frac{(\hat{p}^{(ij)}_{ab} - \hat{\pi}^{(ij)}_{ab})^2}{\hat{\pi}^{(ij)}_{ab}}$$

where $i = 1, \ldots, q-1$; $j = i+1, \ldots, q$; $a = 1, \ldots, c$; $b = 1, \ldots, c$.

Define

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

where $m = (i-1)(c-1)^2 + (j-2)(c-1)^2$. Then $GFfit^{(ij)}$ is a special case of $X^2_{[t:u]}$ (Cagnone and Mignani, 2007):

$$GFfit^{(ij)} = \boldsymbol{e}'\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}^{-1}\boldsymbol{e}$$

$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}} = n^{-1}\boldsymbol{\Omega}_{\boldsymbol{e}}$ with $\boldsymbol{\Omega}_{\boldsymbol{e}}$ evaluated at the MLE $\hat{\boldsymbol{\beta}}$. Now

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

$\mathbf{H}^{(ij)}_{[2]}$ is a partition of the general matrix $\mathbf{H}_{[1:q]}$ The extension to higher-order statistics is straightforward: Define

$$\mathbf{H}^{(ijk)}_{[3]} = \begin{pmatrix} \boldsymbol{h}'_{m+1} \\ \boldsymbol{h}'_{m+2} \\ \vdots \\ \boldsymbol{h}'_{m+(k-1)^3} \end{pmatrix}_{[3]}$$

where $m = (i-1)(c-1)^3 + (j-2)(c-1)^3 + (k-3)(c-1)^3$

$$GFfit^{(ijk)} = \boldsymbol{e}'\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}^{-1}\boldsymbol{e}$$

$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}} = n^{-1}\boldsymbol{\Omega}_{\boldsymbol{e}}$ with $\boldsymbol{\Omega}_{\boldsymbol{e}}$ evaluated at the MLE $\hat{\boldsymbol{\beta}}$. Now

$$\boldsymbol{\Omega}_{\boldsymbol{e}} = \ \mathbf{H}^{(ijk)}_{[3]}(D(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{G}')(\mathbf{H}^{(ijk)}_{[3]})'$$

$\mathbf{H}^{(ijk)}_{[3]}$ is a partition of the general matrix $\mathbf{H}_{[1:q]}$

Now define an orthogonal components version of $GFfit$:

$$GFfit^{(ij)}_{\perp} = \sum_{\ell=m+1}^{\ell=m+(c-1)^2} \hat{\gamma}^2_{\ell}$$

where $m = q + (i-1)(c-1)^2 + (j-2)(c-1)^2$.

Then

$$X^2_{[2]} = \sum_{i=1}^{i=q-1}\sum_{j=i+1}^{j=q} GFfit^{(ij)}_{\perp}$$

More general,

$$X^2_{PF} = \sum_{\ell=1}^{\ell=q(c-1)} \hat{\gamma}^2_{\ell} + \sum_{\ell=q(c-1)+1}^{\ell=\binom{q}{2}(c-1)^2} \hat{\gamma}^2_{\ell} + \sum_{\ell=\binom{q}{2}(c-1)^2+1}^{\ell=\binom{q}{3}(c-1)^3} \hat{\gamma}^2_{\ell} + \cdots + \hat{\gamma}^2_{T-g-1}$$

Then

$$X^2_{PF} = \sum_{i} GFfit^{(i)}_{\perp} + \sum_{i}\sum_{j} GFfit^{(ij)}_{\perp} + \sum_{i}\sum_{j}\sum_{k} GFfit^{(ijk)}_{\perp} + \cdots + GFfit^{(1,2,\ldots,q)}$$

because

$$X^2_{PF} = \hat{\boldsymbol{\gamma}}'\hat{\boldsymbol{\gamma}} = \sum_{\ell=1}^{\ell=T-g-1} \hat{\gamma}^2_{\ell}$$

The extended $GFfit^{(ij)}_{\perp}$ are independent chi-squared statistics on $(c-1)^2$ degrees of freedom due to the definition on orthogonal components. The original $GFfit^{(ij)}$ statistics are not necessarily independent and do not necessarily sum to $X^2_{[2]}$. Because the extended $GFfit^{(ij)}_{\perp}$ are defined on orthogonal components, they are order dependent.

5. Simulation

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. A simulation study to assess the performance of the $GFfit_{\perp}^{(ij)}$ was performed using the Generalized Linear Latent Variable Model (GLLVM).

5.1 GLLVM

GLLVM is a latent variable response model for categorical variables with 2 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

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

where $f(\boldsymbol{x})$ is the density function of $\boldsymbol{X}$. Simulations were conducted under the assumption $\boldsymbol{X} \sim N_p(0, \boldsymbol{\Sigma}_{\boldsymbol{X}})$.

The conditional probability of $\boldsymbol{Y}$ given $\boldsymbol{x}$ is a multinomial probability function:

$$\pi(\boldsymbol{Y} = \mathbf{y}_s | \boldsymbol{x}) = \prod_{i=1}^{q} \pi_{a_i}^{i}(\boldsymbol{x}) = \prod_{i=1}^{q} (\eta_{a_i}^{i} - \eta_{a_{i-1}}^{i})$$

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.

5.2 Simulation Design and Results

A simulation study was conducted using GLLVM to assess the accuracy of the Type I error rates for X_{PF}^2, $X_{[2]inv}^2$, and $X_{[2]ss}^2$. $X_{[2]inv}^2$ is $X_{[2]}^2$ calculated directly using a generalized inverse, and $X_{[2]}^2 ss$ is $X_{[2]}^2$ calculated by using orthogonal components obtained from sequential sum of squares.The simulation had several conditions, with $q = 4$, $q = 5$, and $q = 6$ variables, and $c = 3$ and $c = 4$ categories. Each condition used 500 pseudo samples of size $n = 500$ each. Pseudo samples were generated from a model with one latent factor, and the GLLVM was fit with the specification of one latent factor. To generate the pseudo data, intercept values were specified in the range of -3 to 3. Slope parameters were specified as follows: For $q = 4$, $\boldsymbol{\beta}_1 = (0.0,\ 0.1,\ 0.2,\ 0.6)$; for $q = 5$, $\boldsymbol{\beta}_1 = (0.0,\ 0.3,\ 0.2,\ 0.1,\ 0.2)$; and for $q = 6$, $\boldsymbol{\beta}_1 = (0.0,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5)$

Simulation results for Type I error are shown in Table 1. The table shows empirical Type I error for nominal $\alpha = 0.05$. The results demonstrate that the chi-square approximation for the Pearson statistic is not valid when the cross-classified table becomes very sparse. The empirical Type I error rate for $X^2_{[2]inv}$ was somewhat inflated due to the difficulty of accurately computing an inverse matrix. $X^2_{[2]ss}$ gave more reliable results due to more stable numeric calculations, although the Type I error rate was somewhat conservative for $q = 6$ and $c = 4$. Some sparseness in the 4 by 4 two-way tables may have produced this conservative result.

Table 1: Monte Carlo Simulation Type I Error

q	4	4	5	6
c	3	4	4	4
X^2_{PF}	0.060	0.086	0.166	0.25
$X^2_{[2]inv}$	0.098	0.050	0.052	0.078
$X^2_{[2]ss}$	0.04	0.066	0.052	0.034

Table 2 shows the mean $GFfit^{(ij)}$ statistics for four variables when $c = 3$ and $c = 4$. Under H_0, the $GFfit^{(ij)}$ statistics are distributed as central chi-square on 4 degrees of freedom when $c = 3$ and on 9 degrees of freedom when $c = 4$. None of the mean values appear large relative to the degrees of freedom, as would be expected if the chi-square approximation is valid.

Table 2: Mean of $GFfit$, 4 variables

	$c = 3$	$c = 4$
$GFfit^{(ij)}_{\perp}$	Mean	Mean
(4,3)	3.82	8.89
(4,2)	3.72	9.15
(4,1)	3.94	8.93
(3,2)	4.12	9.29
(3,4)	4.13	9.29
(2,1)	3.93	9.45

A simulation to examine power of the statistics was also conducted using GLLVM. Pseudo data for 1000 samples each of size 500 were generated from a confirmatory two-factor model with all parameters fixed and then fit with a model specifying one factor. Each sample had four variables with three categories. Parameters for the data generating model included intercepts, $\boldsymbol{\beta}_{0(1)} = (-1.5,\ -0.6,\ 0.3,\ 1)'$ and $\boldsymbol{\beta}_{0(2)} = (-1.0,\ -0.3,\ 0.6,\ 1.5)'$, slopes for factor 1, $\boldsymbol{\beta}_1 = (1.0,\ 1.0,\ 1.0,\ 1.0)'$, and slopes for factor 2, $\boldsymbol{\beta}_2 = (0.0,\ 0.1,\ 0.2,\ 0.6)'$.

The two latent variables were specified as uncorrelated, each with variance equal to 1.0. Estimation of the one-factor GLLVM converged for 970 of the 1000 samples, so simulation results reported in Tables 3 and 4 are based on 970 samples.

Table 3 shows empirical power for X^2_{PF}, $X^2_{[2]inv}$, and $X^2_{[2]ss}$. It is difficult to assess the empirical power of X^2_{PF} because it has an inflated Type I error rate in sparse tables, as demonstrated above. $X^2_{[2]inv}$ appears to have the highest empirical power, but direct calculation of $X^2_{[2]}$ using a matrix inverse is unreliable as demonstrated by the very high standard deviation of the values calculated in the simulation. $X^2_{[2]ss}$, on the other hand, gives numerically stable results.

Table 3: Power Simulation Results

	Mean	SD	Power
X^2_{PF}	84.53	25.47	0.268
$X^2_{[2]inv}$	72.98	100.28	0.800
$X^2_{[2]ss}$	39.74	10.23	0.591

Table 4 shows means for the $GFfit^{(ij)}_{\perp}$ statistics calculated in the simulation. Under the null hypothesis, these statistics are distributed chi-square on $(3-1)(3-1)=4$ degrees of freedom. These $GFfit^{(ij)}_{\perp}$ statistics show that primarily the association between variables 2 and 3 was not adequately explained by the one-factor model.

Table 4: $GFfit$ from Power Simulation

$GFfit^{(ij)}_{\perp}$	Mean
(1,2)	4.11
(1,3)	5.95
(1,4)	6.68
(2,3)	14.93
(2,4)	4.04
(3,4)	4.03

6. Application

The extended *GFfit* statistic was used to evaluate the fit of a single factor model to responses given to five questions about the psychiatric condition known as agoraphobia. Agoraphobia is described as an anxiety disorder where a person suffers from a fear and avoidance of situations that might cause panic, or a feeling of being trapped, helpless or embarrassed (Wittchen, Gloster, Beesdo-Baum, Fava, and Craske; 2010). The questions asked about (1) fear of tunnels or bridges, (2) fear of being in a crowd, (3) fear of transportation, (4) fear of going out of the house alone, and (5) fear of being alone. The responses to the questions were collected as part of the Epidemiological Catchment Area Study of 1980-1985 (U.S. Dept. of Health and Human Services, 1985). The data used in this example consists of the responses from 3,305 adults sampled from the Baltimore catchment area. The responses were coded into three ordered categories: (1) fear present at a clinical level, (2) fear present but not at a clinical level, and (3) fear not present. The marginal proportion of any of the fears in the sample is very low, at five percent or less, so the 3^5 cross-classified table is very sparse with a large number of cells that have count equal to 0.

The GLLVM with one factor was fit to these data, and fit statistics were calculated, using R software. Goodness-of-fit test results are shown in Table 5. The results indicate that the model of one underlying factor does not fit well for the agoraphobia symptoms. 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. The stable statistic $X^2_{[2]ss}$ indicates that the model should be rejected. $GFfit^{(ij)}$ statistics for the two-way associations are shown in Table 6. Since each survey question had three response categories, the $GFfit^{(ij)}$ statistics follow a chi-square distribution on $(3-1)^2 = 4$ degrees of freedom. Relative to the central chi-square distribution, several of the $GFfit^{(ij)}$ statistics are large, but $GFfit^{(1,3)}$ and $GFfit^{(2,3)}$ are particularly large. Question 3 asks about fear of transportation, while Question 1 asks about fear of tunnels or bridges and Question 2 asks about fear of being in a crowd. Transportation often involves tunnels and bridges, and public transportation such as a bus or a train involves crowds as well, so these symptoms overlap more than can be accounted for by the model of a single latent factor.

Table 5: Goodness-of-Fit Tests

	Value	DF	p-value
X^2_{PF}	382.97	227	< 0.0001
$X^2_{[2]inv}$	180.46	40	< 0.0001
$X^2_{[2]ss}$	185.48	40	< 0.0001

Each of the $GFfit^{(ij)}$ statistics shown in Table 6 are the sum of four orthogonal components of Pearson's statistic. The $GFfit^{(ij)}$ statistics shown in Table 6 sum to 185.48, which is equivalent to the value of the $X^2_{[2]}$ statistic. The $X^2_{[2]}$ statistic is then the sum of 40 orthogonal components. In a similar way, $X^2_{[3]}$ is a sum of 80 orthogonal components, or 10 $GFfit^{(ijk)}$ statistics. Table 7 shows partitioning of Pearson's full statistic into blocks of components associated with marginals of order 2 to 5. The full Pearson statistic has 227 degrees of freedom in this case, and the value taken on can be obtained by summing 227 orthogonal components. The first-order marginals usually provide little information on lack of fit for this type of model, so components from marginals of order 2 to 5 are used to obtain the value of the full Pearson statistic. Using 40 components from $X^2_{[2]}$, 80 components from $X^2_{[3]}$, 80 components from $X^2_{[4]}$, and 27 (out of 32 possible) components from $X^2_{[5]}$, it can be seen that $X^2_{[2]} + X^2_{[3]} + X^2_{[4]} + X^2_{[5]} = 382.29$, which compares, with round-off error, to $X^2_{PF} = 382.97$. Other statistics, such as $X^2_{[2:3]}$ and $GFfit^{(ijk)}$ can be easily obtained from the orthogonal components of X^2_{PF}, but the third-order marginals may be too sparse for the chi-square approximation to hold. Including third-order marginals may also dilute the test. Further simulations would be required to evaluate their use.

Table 6: $GFfit_{\perp}^{(ij)}$ Agoraphobia Symptom Items

$GFfit_{\perp}^{(ij)}$	Value
(1,2)	16.41
(1,3)	55.32
(1,4)	5.70
(1,5)	6.82
(2,3)	24.96
(2,4)	11.00
(2,5)	11.79
(3,4)	9.43
(3,5)	14.97
(4,5)	29.06

Table 7: Partitions of Pearson's Statistic

	Value	DF
$X_{[2]}^2$	185.48	40
$X_{[3]}^2$	101.8	80
$X_{[4]}^2$	68.24	80
$X_{[5]}^2$	26.76	27

7. Conclusions

Pearson's statistic can be decomposed into components that are defined as an extended version of the $GFfit$ statistic. The $GFfit_{\perp}^{(ij)}$ statistics can be calculated reliably using sum of squares from an orthogonal regression. A more global test statistic such as $X_{[2]}^2$ based on second-order marginals can be obtained as as a sum of $GFfit_{\perp}^{(ij)}$ statistics. When applied to the GLLVM, the $GFfit_{\perp}^{(ij)}$ are useful as item diagnostics to detect the source of lack of fit. The global test should be conducted first. A study with a large number of manifest variables will produce a large number of $GFfit_{\perp}^{(ij)}$ statistics, and a multiple decision rule should be used to identify unusually large values of $GFfit_{\perp}^{(ij)}$. An application to agoraphobia symptoms showed that fear of public transportation has overlap with fear of tunnels/bridges and fear of crowds that cannot be explained by a model of a single underlying factor for the five agoraphobia symptoms.

References

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

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

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

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

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

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

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

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

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

Wittchen, H., Gloster, A. T., Beesdo-Baum, K., Fava, G., and Craske, M. (2010). Agoraphobia: A review of the diagnostic classificatory position and criteria. *Depression and Anxiety*, 27:113-133.