A Power study of the GFfit statistic as a Lack-of-fit Diagnostic for sparse two-way subtables

Junfei Zhu[1], Mark Reiser[2], Maduranga Dassanayake[3], Silvia Cagnone[4]
[1]School of Mathematical and Statistical Sciences, Arizona State University, USA
[2]School of Mathematical and Statistical Sciences, Arizona State University, USA
[3]School of Mathematical and Statistical Sciences, Arizona State University, USA
[4]Department of Statistical Sciences, University of Bologna, Italy

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 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. It has been proposed to assess model fit by using a new version of GFfit statistic based on orthogonal components of Pearson chi-square as a diagnostic to examine the fit on two-way subtables. However, due to variables with a large number of categories and small sample size, even the GFfit statistic may have low power and inaccurate Type I error level due to sparseness in the two-way subtable. In this paper, a method based on choosing different orthogonal components for the GFfit statistic on the subtables is developed to improve the performance of the GFfit statistic. Simulation results for power and type I error rate for several different cases along with comparisons to other diagnostics are presented.

Key words: sparseness, GFfit statistic, orthogonal components, chi-square test, goodness-of-fit,

1.INTRODUCTION

Traditionally we use the likelihood ratio (LR and the Pearson chi-square (GF to test goodness of fit for a model fit on cross-classified variables

$$LR = 2n\sum_{r=1}^{k} f_r \ln(\frac{f_r}{\hat{\pi}_r})$$

$$GF = n\sum_{r=1}^{k} \frac{(f_r - \hat{\pi}_r)^2}{\hat{\pi}_r}$$

Suppose we have p categorical variables and the i-th variable has c_i categories. Thus there are $k = \prod_{\mathrm{i}=1}^{\mathrm{p}} c_i$ cells in the cross-classified table. Each cell corresponds to a response pattern. Then f_r is the sample proportion of the r-th response pattern and $\hat{\pi}_r$ is the estimated probability of the r-th response pattern. If the number of observations in each response pattern is large enough and under the conditions (Koehler and Larntz, 1980 that i) $H_0: \pi = \pi(\theta)$, ii) k is fixed and iii) $\min_{1\le r\le k} n\pi_r \to \infty$ for $n \to \infty$, both LR and GF are approximately distributed χ^2 with degree of freedom equal to $\mathrm{k} - 1 -$ number of estimated parameters. However, when there is a problem of sparseness, these two statistics may not have an approximate chi-square distribution. Several statistics have been proposed using marginal distributions of the joint variables rather than the joint distribution.

Joreskog and Moustaki (2001 proposed the GFfit statistic as a diagnostic to help in finding the source of model lack of fit. A new version of the GFfit statistic is proposed by Reiser, Cagnone & Zhu (2014 by decomposing the Pearson statistic from the full table into orthogonal components defined on lower-order marginal distributions. Then the GFfit statistic is defined as a sum of a subset of these components. However, due to variables with a large number of categories and small sample size, even this GFfit statistic may have low power and inaccurate Type I error level due to sparseness in the two-way subtable. In this paper, a method based on choosing different orthogonal components for the GFfit statistic on the subtables is developed to improve the performance of the GFfit statistic.

The paper is organized as follows: In Section 2 we introduce the marginal proportion and the GFfit orthogonal components. In Section 3 we introduce the method to choose several orthogonal components for the GFfit statistic. In Section 4 we give a discussion of the GLLVM model. In Section 5 simulation results for power of the GFfit statistic to detect lack of fit along with comparisons to other diagnostics 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.

2.1 First- and Second-order Marginals

Consider the three variables, two categories case. An 8 by 3 matrix $\boldsymbol{V}$ can be used to denote the response patterns as the rows:

$$V = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Let v_{is} represent element i of response pattern s, $i = 1, \dots, p$ and $s = 1, \dots, k$. In this example, $p = 3$ and $k = 8$. Then, under some specific model, which we will introduce later, the first-order marginal proportion for variable y_i can be defined as

$$P_i(\theta) = \text{Prob}(y_i = 1|\theta) = \sum_s v_{is}\pi_s(\theta)$$

and the true first-order marginal proportion is given by

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

Thus the marginal proportions are linear combination of joint proportions:

$$\boldsymbol{P} = \mathbf{H}\boldsymbol{\pi}$$

The H matrix can be defined from the V matrix. For first-order marginal, $\mathbf{H}_{[1]} = \boldsymbol{V}'$.

For 3 variables with 3 categories, $\mathbf{H}_{[1]} = \boldsymbol{V}'$, where

$$V_{27*6} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Under the model, for two categories, the second-order marginal proportion for variable y_i and y_j can be defined as

$$P_{ij}(\boldsymbol{\theta}) = \text{Prob}\big(y_i = 1, y_j = 1\big|\boldsymbol{\theta}\big) = \sum_s v_{is}v_{js}\pi_s(\theta) \ ,$$

and the true second-order marginal proportion is given by

$$P_{ij} = \text{Prob}\big(y_i = 1,\ y_j = 1\big) = \sum_s v_{is}v_{js}\pi_s \ .$$

If the number of categories c is greater than 2, the second-order marginal proportions for y_i and y_j can be represented as a c by c table with $(c-1)^2$ independent proportions. Thus for second-order marginal proportions, the rows of H are Hadamard products among the columns of V. For 3 variables with 3 categories, $\mathbf{H}_{[2]}$ is an 12 by 27 matrix:

$$\mathbf{H}_{[2]} = \begin{bmatrix} (v_1 \circ v_3)' \\ (v_1 \circ v_4)' \\ \vdots \\ (v_1 \circ v_5)' \\ (v_1 \circ v_6)' \\ \vdots \\ (v_3 \circ v_5)' \\ \vdots \\ (v_{i(c-1)} \circ v_{j(c-1)})' \end{bmatrix}$$

where v_i is the column i of matrix V, and $v_i \circ v_j$ is the Hadamard product of columns i and j.

2.2 Test statistic

Linear combinations of $\boldsymbol{\pi}$ may be tested under the null hypothesis H_0: $\mathbf{H}\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\pi}(\boldsymbol{\theta})$ and the test statistic is

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

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

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

$D(\boldsymbol{\pi})$ =diagnal matrix with (s, s) element equal to $\pi_s(\boldsymbol{\theta})$

$$\mathbf{A} = D(\pi)^{-1/2}\frac{\partial\boldsymbol{\pi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$$

$$\mathbf{G} = \frac{\partial\boldsymbol{\pi}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$$

$\boldsymbol{e} = \mathbf{H}(\boldsymbol{f} - \boldsymbol{\pi})$ is the matrix form of the marginal residuals.

$\mathbf{H} = \mathbf{H}_{[1:2]}$ produces $X^2_{[1:2]}$ and $\mathbf{H} = \mathbf{H}_{[2]}$ produces $X^2_{[2]}$. It has been proven that for two categories, the distributions of $X^2_{[1:2]}$ and $X^2_{[2]}$ are chi-square distributions with degrees of freedom equal to $q(q+1)/2$ and $q(q-1)/2$ respectively. $X^2_{[1:q]} = GF$. $X^2_{[t:u]}$ is a score statistic, Reiser (1996 , Reiser and Lin (1999 , Cagnone and Mignani (2007 , Rayner and Best (1989 .

2.3 Orthogonal components

Consider the $k - g - 1$ by c^q matrix $\mathbf{H}^* = \boldsymbol{F}'\mathbf{H}_{[1:q;-g]}$, where g is the number of unknown model parameters to be estimated and $\mathbf{H}_{[1:q;-g]}$ is matrix $\mathbf{H}_{[1:q]}$ deleting g rows. $\mathbf{H}^*$ 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 $D(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-\mathbf{1}}\mathbf{G}'$.

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

where $\widehat{\mathbf{H}}^* = \mathbf{H}^*(\widehat{\boldsymbol{\theta}})$, and $\mathbf{r} = (\widehat{\boldsymbol{p}} - \boldsymbol{\pi}(\widehat{\boldsymbol{\theta}}))$.

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 $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. Then

$$X^2_{PF} = \hat{\boldsymbol{\gamma}}'\hat{\boldsymbol{\gamma}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{j=k-g-1}} \hat{\gamma}_j^2$$

$\widehat{\mathbf{H}}^*\mathbf{r}$ has asymptotic covariance matrix $\boldsymbol{F}'\boldsymbol{\Omega}_{\boldsymbol{e}}\boldsymbol{F} = \boldsymbol{I}_{\boldsymbol{k-g-1}}$. The elements $\hat{\gamma}_j^2$ are asymptotically independent chi-square random variables with $df = 1$ (Reiser, 2008 .

Using Sequential Sum of Squares: Redefine

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

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

$$\boldsymbol{z} = \mathbf{H}'\boldsymbol{\beta}$$

Then,

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

where $\mathbf{u} = \sqrt{n}\mathbf{r}$, $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{D}}^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{D}}^{\frac{1}{2}} = \widehat{\boldsymbol{D}}^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{D}}^{\frac{1}{2}}$, and $\boldsymbol{D} = diag(\boldsymbol{\pi}(\boldsymbol{\theta}))$.

$\Sigma = \Sigma(\theta) = (\boldsymbol{I} - \boldsymbol{\pi}^{\frac{1}{2}}\left(\boldsymbol{\pi}^{\frac{1}{2}}\right)' - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-\mathbf{1}}\boldsymbol{A}')$ is idempotent.

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

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

$\hat{\gamma}_j^2, j = 1, k-g-1$ are the sequential SS from this regression. $\boldsymbol{\gamma} = \boldsymbol{C}'\boldsymbol{\beta}$ are the orthogonal coefficients.

Now define an orthogonal components version of $GFfit$:

$$GFfit_{\perp}^{(ij)} = \sum_{l=m+1}^{l=m+(c-1)^2} \hat{\gamma}_l^2$$

where $m = ((i-1)p - \frac{i(i-1)}{2})(c-1)^2 + (j-2)(c-1)^2$, assuming $\mathbf{H} = \mathbf{H}_{[2]}$. The extended $GFfit_{\perp}^{(ij)}$ are independent chi-square statistics with $df = (c-1)^2$ because of the definition on orthogonal components. The original $GFfit^{(ij)}$ statistics are not necessarily independent and do not necessarily sum to $X_{[2]}^2$. $GFfit_{\perp}^{(ij)}$ statistics are order dependent since they are defined on orthogonal components.

3. Improve $GFfit_{\perp}^{(ij)}$ by choosing appropriate orthogonal components.

Although using $GFfit_{\perp}^{(ij)}$ is a good remedy to problem of sparseness, sometimes even $GFfit_{\perp}^{(ij)}$ may have low power and inaccurate Type I error level due to sparseness in the two-way subtable.

We want to improve $GFfit_{\perp}^{(ij)}$ by choosing appropriate orthogonal components. Since the problem is due to sparseness, one way to improve $GFfit_{\perp}^{(ij)}$ is just using the orthogonal components corresponding to several cells with the largest frequencies. I denote this statistic by $GFfit_{\perp(t)}^{(ij)}$, where t means computing the statistic with the t cells having the largest frequencies. In this case we only use t orthogonal components, so the degrees-of-freedom is t for $GFfit_{\perp(t)}^{(ij)}$. Since we are selecting the t orthogonal components corresponding to t cells with the largest frequencies, not the t largest orthogonal components, $GFfit_{\perp(t)}^{(ij)}$ is not an order statistic.

In the following table, I labeled the cells that used to compute the $GFfit_{\perp}^{(ij)}$ for the four categories case. For example, cell 2 is the cell with category 3 in variable i and category 4 in variable j.

TABLE 1: label of cells for variables 4 variables 4 categories case

Label of the cells		Category of variable j			
		1	2	3	4
Category of variable i	1	16	12	8	4
	2	15	11	7	3
	3	14	10	6	2
	4	13	9	5	1

Theoretically, we can choose any 9 cells that can produce the full table to compute $GFfit_{\perp}^{(ij)}$. By default, we will use the cells in the bottom right corner to compute $GFfit_{\perp}^{(ij)}$. For the four categories case, these cells are 1, 2, 3, 5, 6, 7, 9, 10 and 11. When computing $GFfit_{\perp(t)}^{(ij)}$, we only use t orthogonal components corresponding to t cells with the largest frequencies. Here "largest" means the largest frequencies among the $(c-1)^2$ cells we choose to compute $GFfit_{\perp}^{(ij)}$.

4. THE GENERALIZED LINEAR LATENT VARIABLE MODEL

Let $\boldsymbol{y} = (y_1, y_2, \dots, y_p)$ be the vector of p ordinal observed variables, each of them having ci categories. Thus there are $\prod_{i=1}^{p} c_i$ cells, also called response patterns in the cross-classified table. The r-th response pattern is indicated as $\boldsymbol{y_r} = (y_1 = a_1, y_2 = a_2, \dots, y_p = a_p)$, where a_i is the value of the i-th observed variable($a_i = 1, \dots, c_i \; and \; i = 1, \dots, p$.

Let $\boldsymbol{z} = (z_1, z_2, \dots, z_p)$ be the vector of q continuous latent variables. Then the probability of the r-th response pattern $\boldsymbol{y_r}$ is given by

$$\pi_r(\theta) = \int \pi_r(\boldsymbol{z})\, h(\boldsymbol{z})\, d\boldsymbol{z}\,,$$

where θ is a vector of parameters. $h(\boldsymbol{z})$ is the density function of z, and we assume every latent variable to be distributed standard normal independently. $\pi_r(z)$ is the conditional probability of y_r given z and it is a multinormial probability function

$$\pi_r(\boldsymbol{z}) = \prod_{i=1}^{p} \pi_{a_i}^{(i)}(z) = \prod_{i=1}^{p} (\tau_{a_i}^{(i)} - \tau_{a_{i-1}}^{(i)})$$

where $\tau_{a_i}^{(i)} = \pi_1^{(i)}(z) + \pi_2^{(i)}(z) + \cdots + \pi_{a_i}^{(i)}(z)$ is the probability of a response in category a_i or lower on the variable i and $\pi_{a_i}^{(i)}(z)$ is the probability of a response in category a_i on the variable i.

We use logistic regression to model the interrelationship between $\tau_{a_i}^{(i)}$ and the latent variables.

$$log\left[\frac{\tau_s^{(i)}}{1 - \tau_s^{(i)}}\right] = \alpha_{i0}(s) - \sum_{j=1}^{q} \alpha_{ij} z_j\,, \;\; s = 1, \dots, c_{i-1}$$

$\alpha_{i0}(s)$ and α_{ij} are the parameters of the model. $\alpha_{i0}(s)$ is the intercept and α_{ij} is the j-th slope for variable i. The intercepts should satisfy the condition $\alpha_{i0}(1) \leq \alpha_{i0}(2) \leq \cdots \leq \alpha_{i0}(c_i)$.

We use the E-M algorithm to calculate the maximum likelihood estimator for the parameters in the model. The integrals are approximated through the Gauss-Hermite quadrature method (Cagnone & Mignani, 2007 .

5. Monte Carlo Simulations

A simulation study was conducted using GLLVM to assess the power of $GFfit_{\perp}^{(ij)}$, M_{ij} and $GFfit_{\perp(\mathrm{t})}^{(ij)}$. M_{ij} is the individual Joe & Maydeu-Olivares chi-square statistic.

In this power study, I tried one 4 variables 4 categories case and one 5 variables 5 categories case. For the 4 variables case, two sample sizes are used, 150 and 500. For the 5 variables case, the sample size is 150. In both simulations, data were generated from a two-factor model and fitted with a one-factor model. The parameters for the data generating models are the following: for 4 variables case $\alpha_{0(1)} = (-1, -1, -1, -1)', \alpha_{0(2)} = (0.5,0.5,0.5,0.5)', \alpha_{0(3)} = (2,2,2,2)', \alpha_1 = (0.0, 1.0, 1.0, 0.0)', \alpha_2 = (2.0,0.1,0.2,2.0)'$ for 5 variables case, $\alpha_{0(1)} = (-1.59, -2.30, -1.43, -3.02, -1.26)', \alpha_{0(2)} = (-0.84, -0.38, -0.32, -1.50, -0.21)', \alpha_{0(3)} = (0.71,0.16,0.15,0.57,0.78)', \alpha_{0(3)} = (1.48,1.80,1.66,2.13,1.65)', \alpha_1 = (1.5,1.7,1.9,2.1,2.3)', \alpha_2 = (0.8,0.8,0,0,0)'$.

For the four variables 500 sample size case, the power of $GFfit_{\perp}^{(ij)}$ and M_{ij} are listed below.

TABLE 2: power of $GFfit_{\perp}^{(ij)}$ and M_{ij}, 4 variables 4 categories 500 sample size

	Power	
	$GFfit_{\perp}^{(ij)}$	M_{ij}
(12	0.067	0.057
(13	0.070	0.041
(14	0.364	0.036
(23	0.816	0.065
(24	0.057	0.064
(34	0.052	0.050

The $GFfit_{\perp}^{(23)}$ has a power of 0.816, which is already a pretty good power since this case is not very sparse for $GFfit_{\perp}^{(ij)}$. All the M_{ij}'s have very low power. Then the power of $GFfit_{\perp(t)}^{(23)}$ are also computed and listed in Table 3.

TABLE 3: Power of $GFfit_{\perp(t)}^{(23)}$ for 4 variables 500 sample size case

t=number of cells	Power of $GFfit_{\perp(t)}^{(23)}$
1	0.284
2	0.441
3	0.565
4	0.647
5	0.721
6	0.752
7	0.796
8	0.816
9	0.816

We can see that the original $GFfit_{\perp}^{(23)}$, which is equivalent to $GFfit_{\perp(9)}^{(23)}$, already has a pretty good power. Then in this case, using only several cells with the largest frequencies to compute $GFfit_{\perp(t)}^{(23)}$ won't be able to improve the power. For example, $GFfit_{\perp(2)}^{(23)}$ is computed by summing up the two orthogonal components corresponding to the cells with largest and second largest frequencies in the (2,3 subtable. And in this case, it only has a power of 0.441, which is much lower than 0.816, the power of the original $GFfit_{\perp}^{(23)}$

Then for the 4 variables 150 sample size case, the table is sparse for $GFfit_{\perp}^{(ij)}$. The power of $GFfit_{\perp}^{(ij)}$ and M_{ij} are listed below.

TABLE 4: power of $GFfit_{\perp}^{(ij)}$ and M_{ij}, 4 variables 4 categories 150 sample size

	Power	
	$GFfit_{\perp}^{(ij)}$	M_{ij}
(12	0.057	0.049
(13	0.065	0.052
(14	0.133	0.049
(23	0.276	0.061
(24	0.041	0.054
(34	0.064	0.053

For the 4 variables 150 sample size case, the table is sparse for $GFfit_{\perp}^{(ij)}$. The original $GFfit_{\perp}^{(23)}$ has a power of 0.276, which is not very good. The power of $GFfit_{\perp(t)}^{(23)}$ are

computed and listed in Table 5. We can see that $GFfit_{\perp(4)}^{(23)}$ has the largest power of 0.415. The reduction in power after t=4 is due to sparseness in the 4 by 4 table for variable 2 and 3.

TABLE 5: Power of $GFfit_{\perp(t)}^{(23)}$ for 4 variables 150 sample size case

t=number of cells	Power of $GFfit_{\perp(t)}^{(23)}$
1	0.303
2	0.360
3	0.394
4	0.415
5	0.410
6	0.413
7	0.371
8	0.317
9	0.276

For the 5 variables 150 sample size case, the table is quite sparse even for $GFfit_{\perp}^{(ij)}$. The power of $GFfit_{\perp}^{(ij)}$ and M_{ij} are listed below.

TABLE 6: power of $GFfit_{\perp}^{(ij)}$ and M_{ij}, 5 variables 5 categories 150 sample size

	Power	
	$GFfit_{\perp}^{(ij)}$	M_{ij}
(12	0.087	0.032
(13	0.059	0.051
(14	0.050	0.056
(15	0.054	0.057
(23	0.063	0.059
(24	0.060	0.063
(25	0.050	0.061
(34	0.053	0.053
(35	0.052	0.043
(45	0.049	0.041

The $GFfit_{\perp}^{(12)}$ has a very low power of 0.086, which is already the largest power among all the $GFfit_{\perp}^{(ij)}$'s. All the M_{ij}'s have very low power again. Then the power of $GFfit_{\perp(t)}^{(12)}$ are computed and listed below.

TABLE 7: Power of $GFfit_{\perp(t)}^{(12)}$ for 5 variables 150 sample size case

t	Power of $GFfit_{\perp(t)}^{(12)}$
1	0.139
2	0.190
3	0.187
4	0.190
5	0.184
6	0.192
7	0.190
8	0.185
9	0.178
10	0.155

11	0.141
12	0.135
13	0.116
14	0.103
15	0.089
16	0.086

We can see that by using only two cells with the largest frequencies, $GFfit_{\perp(2)}^{(12)}$ has a power of 0.19, which is more than twice of 0.086, the power of the original $GFfit_{\perp}^{(12)}$.

6. Conclusion

The $GFfit_{\perp(\mathrm{t})}^{(ij)}$ statistics can be calculated by choosing the t orthogonal components corresponding to t cells with the largest frequencies. This statistic is not an order statistic. Monte Carlo simulations demonstrated that the $GFfit_{\perp(\mathrm{t})}^{(ij)}$ statistics perform well when sparseness is present. It can be used as diagnostic to assist in detecting the source of poor fit when the model specified in the null hypothesis is rejected. However, when $GFfit_{\perp}^{(ij)}$ has a good power, using only several cells with the largest frequencies to compute $GFfit_{\perp(\mathrm{t})}^{(ij)}$ won't be able to improve the power. When the dataset is really sparse and $GFfit_{\perp}^{(ij)}$ has a very poor power, $GFfit_{\perp(\mathrm{t})}^{(ij)}$ can improve the power. Further research is needed to determine the level of sparseness in a two-way subtable when $GFfit_{\perp(\mathrm{t})}^{(ij)}$ would have higher power than $GFfit_{\perp}^{(ij)}$.

REFERENCES

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

Bartholomew, D. J. & Knott, M. (1999 . *Latent Variable Models and Factor Analysis*, Kendall's Library of statistics, London, second edition.

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

Cochran, W. G. (1954 . Some methods for strengthening the common chi-square tests. *Biomedical Journal*, **10**, 417-451.

Goodman, L. A. (1964 . Simple methods for analyzing three-factor interaction in contingency tables. *Journal of the American Statistical Association*, **59**, 319-385.

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

Holst, L. (1972 . Asymptotic normality and efficiency for certain goodness-of-fit test. *Biometrika*, **59**, 137-145

Koehler, K. J. (1986 . Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the American Statistical Association*, **81**,483-493

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

Morris, C. (1975 . Central limit theorems for multinomial sums. *Annals of Statistics*, **3**,

365-384.
Rayner, J. C. W. & Best, D, J. (1989 . *Smooth Tests of Goodness of Fit.* Oxford: New York.
Reiser, M. & Vandenberg, M. (1994 . Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, 47, 85-107.
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. (2012 . Limited-information statistics when the number of variables is large. Proceedings of 2012 Joint Statistical Meetings, San Diego, CA.
Reiser, M., Cagnone, S. & Zhu, J. (2014 . An Extended *GFfit* Statistic Defined on Orthogonal Components of Pearson's Chi-Square. *In JSM Proceedings, Biometrics Section, Alexandria, VA: American Statistical Association.*
Zhu, J., Reiser, M. & Cagnone, S. (2015 A Power Study of the GFfit Statistic as a Lack-of-Fit Diagnostic. *JSM Proceedings, Biometrics Section, Alexandria VA, American Statistical Association.*