Revista de Matemática: Teoría y Aplicaciones 2016 **23**(1) : 199–220
cimpa – ucr   issn: 1409-2433 (Print), 2215-3373 (Online)

An alternative to classical latent class models selection methods for sparse binary data: an illustration with simulated data

Un método alternativo para la selección de modelos de clases latentes en datos binarios escasos: una ilustración con datos simulados

Carlomagno Araya Alpízar[*]

*Received: 25 Jun 2014; Revised: 28 Aug 2015;
Accepted: 19 Oct 2015*

[*] Sede de Occidente, Universidad de Costa Rica, San Ramón, Costa Rica. E-Mail: carlo.araya@ucr.ac.cr

Abstract

Within the context of a latent class model with manifest binary variables, we propose an alternative method that solves the problem of estimating empirical distribution with sparse contingency tables and the chi-square approximation for goodness-of-fit will not be valid. We analyze sparse binary data, where there are many response patterns with very small expected frequencies in several data sets varying in degree of sparseness from 1 to 5 defined $d = n/2^p = n/R$ is a factor that is mentioned in almost all prior literature as being an important determinant of how well the distribution is represented by the chi-squared.The proposed approach produced results that were valid and reliable under the mentioned problematic data conditions. Results from the proposal presented compare the rates of Type I for traditional goodness-of-fit tests. We also show that with data density $d \leq 5$, Pearson's statistic $\left(\chi^2\right)$ should not be used to select latent class models using the Patterns Method, given that this has the probability of Type I error being greater than 5%. By comparing the Patterns Method and the Parametric Bootstrap for data density $d = 2$, we show that the Patterns Method has more accurate Type I error probabilities since the likelihood ratio, Read-Cressie and Freeman-Tukey statistics afford values of $\alpha < 0.05$. In contrast, the Parametric Bootstrap provides values in these statistics that surpass 5%.

Keywords: sparse data; latent class; goodness-of-fit; binary data.

Resumen

En el contexto de modelos de clases latentes con variables manifiestas binarias, se propone un método alternativo para resolver el problema de la estimación de la distribución empírica con tablas de contingencias escasas, donde la aproximación de los estadísticos de bondad de ajuste por la distribución Chi-Cuadrada no es valida. Se analiza datos binarios escasos, donde muchos patrones de respuesta que tienen frecuencias esperadas pequeñas, en conjuntos de datos con grados de datos escasos de 1 a 5, donde $d = n/2^p = n/R$ es un factor es mencionado en la literatura como determinante de la bondad de ajuste a la distribución Chi-Cuadrada. La propuesta presenta resultados validos y confiables en las condiciones de los datos mencionadas. Para los resultados se presenta tasas de error tipo I para las pruebas tradiciones de bondad de ajuste. También se muestra que para niveles de densidad de datos $d \leq 5$, el estadístico Pearson $\left(\chi^2\right)$ no es el apropiado para seleccionar modelos de clases latentes utilizando el Método de Patrones, dado que presenta probabilidad de error de tipo I mas grandes que 5%. Al comparar el Método de Patrones y el Bootstrap Paramétrico para la densidad $d = 2$, se muestra que el Método de Patrones tiene probabilidades de error de tipo I menores de 5% en los estadísticos de razón de verosimilitud, Read-Cressie y Freeman-Tukey. En contraste, el Bootstrap Paramétrico produce valores en estos estadísticos que superan un 5%.

Palabras clave: datos escasos; clases latentes; bondad de ajuste; datos binarios.

Mathematics Subject Classification: 62H30, 62H17.

1 Introduction

Latent class analysis is a statistical method for analyzing and understanding multivariate categorical data. These methods have been used extensively in the social sciences to model the heterogeneity of manifest variables in a multivariate sense; they can be used to identify unobserved subgroups within a population from multivariate categorical and/or continuous observed variables by estimating the characteristics of these latent clusters, returning the probability that each subject belongs to each group and identifying the variables that best serve to distinguish among classes [3].

In theory, a p-value value for the goodness-of-fit statistic (GFS) can be obtained by comparing the statistics to the reference chi-square theoretical probability distribution corresponding to the degrees of freedom in the model. The assumption is valid based on the Integral Theorem of De Moivre-Laplace, when both the observed frequencies of the different response patterns and the sample size are large. ($f_{\mathbf{x}_h} \to \infty, n \to \infty$).

This is a special case of the central limit theorem. It states that the binomial distribution of the number of successes in n independent Bernoulli trials with a probability p of success in each trial approximates a normal distribution with mean np and standard deviation $\sqrt{npq}$ if n is very large and some conditions are satisfied [26].

Statistical inference problems caused by sparsity of contingency tables are widely discussed in the literature. The problem arises because the possible increase in Type I error rates of goodness-of-fit statistics do not match their expected rates under the chi-square approximation [6]. In particular, sparseness is a function of the sample size and the size of the contingency table (or the total of the response patterns, $R = 2^p$). The ratio $d = n/R$ is used to measure the amount of spareness present in a table. In this sense, Larntz [21] showed that when d is less than 5, the distribution of the G^2 test statistic is not well approximated by the chi-square distribution, and instead is unknown, making it difficult to test the absolute fit of a model. Accordingly, small expected frequencies will provide high values in the GFS and will be more likely to lead us to reject the model even though it is appropriate for describing the data set. Furthermore, sparse data have an adverse effect on goodness-of-fit tests as they may invalidate using the χ^2 distribution. Many suggestions have been given on how to measure sparseness in a multi-way contingency table. But, to date, no universal definition of sparseness has been adopted. The most widely used rules of thumb are to consider the percentage of expected cell frequencies smaller than or equal to 1, 5 or 10 [2, 7, 13, 15, 31, 19, 9] and the percentage of observed zero frequencies.

Other contributions to the study of the sparse contingency tables are: Bayesian modeling of temporal dependence in large sparse contingency tables [18], nonparametric criteria for sparse contingency tables [30], goodness-of-fit tests for sparse nominal data based on grouping [28], accurate directional inference for vector parameters [10], chi-square orthogonal components for assessing goodness-of-fit [24], profile statistics for sparse contingency tables under Poisson sampling [27], the measurement of model fit for sparse categorical data [17] and modeling and measuring association for ordinal data [14].

We propose an alternative method to calculate the empirical probability distribution of the GFS when sparseness is extreme and here we refer to this as the Patterns Method. The structure of this article is as follows. In Section 2, we present the basic concepts related to Latent Class Models. In Section 3, the sparse data problem and the statistical tests that assess the goodness-of-fit of the latent class model are presented. In Section 4, we present the Patterns Method. The design of the study is described in Section 5. Section 6 includes the results of the Pattern Method regarding probability of Type I error. Section 7 describes the construction of the tables of critical values for the Pattern Method. Finally, Section 8 presents the key results of this study.

2 Latent class models

The latent class model (LCM) is a multivariate statistical technique that allows study of the existence of one (or several)latent class(es) by means of a set of manifest variables observed, and makes it possible to define, from their classes, a classification or typology of the individuals analyzed. LCM was introduced by Lazarsfeld and Henry [22], who used the technique as a tool for building typologies (or clustering) based on dichotomous observed variables.

In Latent Class analysis, the measurement levels of both the manifest variables and the latent variable are categorical. Each latent class is characterized by a pattern of probabilities of response for the manifest variables. A particular case of LCM occurs when the manifest variables are binary; that is, there are only two levels of response: 0 and 1. Formally we have a collection $\mathbf{X}' = (X_1, \cdots, X_p)$ of binary indicators for each individual, these being the presence or absence of particular events. Let $\mathbf{X}'$ be a vector of p binary manifest variables which form a $p-$dimensional contingency table. Let us assume that these p variables are considered to be indicators of a latent variable Y with C categories or latent classes. The LCM describing this situation is given by $\pi_{\mathbf{X}}\left(\mathbf{x}\right) = \sum_{c=1}^{C} \pi_{\mathbf{X},Y}\left(\mathbf{x}, c\right)$ where $\pi_{\mathbf{X},Y}\left(\mathbf{x}, c\right) = P(\mathbf{X} = \mathbf{x}, Y = c)$ is the overall likelihood that a randomly selected individual will have a response $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ and is in the latent class c.

We shall assume conditional independence; therefore, the overall likelihoods follow a Bernoulli distribution $\pi_{X_i/Y(c)}(x_i) = \pi_{ic}^{x_i}\left(1 - \pi_{ic}\right)^{1-x_i}$ where π_{ic} is the conditional probability of obtaining a positive response in the X_i variable for an individual of the latent class c. In practice, for each response pattern this set of probabilities is inspected and the individual is assigned to the latent class in which this probability is greatest (modal assignment).

The estimation for item parameters and sizes of latent classes are estimated in the expectation-maximization (EM) algorithm. The EM is a general method for maximum likelihood estimation in a missing data setting and convergence is checked by determining the relative change in the log-likelihood of subsequent iterations.The usual procedure to decide on the number of classes begins with a small number of classes and then checks whether an additional class could improve the fit significantly.

3 Sparse contingency tables

The statistical tests that assess the goodness-of-fit of the model to the data are based on null hypotheses derived from the theoretical models. For categorical data, significance tests normally entail a comparison between the observed and expected frequencies that are derived by substituting maximum-likelihood estimates for parameters in the theoretical model. The three most commonly used GFS for goodness-of-fit testing of a latent class model are: the Pearson chi-squared statistic (χ^2), the likelihood ratio statistic (G^2) and the Freeman-Tukey statistic (FT). All of the above statistics are embedded in a family of power divergence statistics thoroughly discussed by Cressie and Read (1984) for multinomial sampling they are obtained using the following formula

$$RC(\lambda) = \frac{2}{\lambda\left(\lambda + 1\right)} \sum_{h=1}^{R} f_{\mathbf{x}_h} \left[\left(\frac{f_{\mathbf{x}_h}}{\hat{f}_{\mathbf{x}_h}} \right)^{\lambda} - 1 \right] \tag{1}$$

where the $f_{\mathbf{x}_h}$ are the observed cell frequencies, $\hat{f}_{\mathbf{x}_h}$ are the expected cell frequencies and R is the number of cells. The special cases of power divergence statistics are Pearson chi-squared statistic (put $\lambda = 1$) and the likelihood ratio statistic (the limit as $\lambda \to 0$).

These statistical tests are asymptotically χ^2 distributed under the null hypothesis in large samples, with a degree of freedom specific for the model. The reference chi-square theoretical probability distribution for statistical tests is based on an asymptotic result under the assumption that each expected cell count is large. In fact, the χ^2 approximation can also break down when the table is small but contains very large as well as small cell counts. Nevertheless, if the number of manifest variables and/or the number of categories of each variable are large, with a small sample size, the multi$-$way contingency table of the observed variables yields sparse data. Suppose that data are available on $p = 12$ dichotomous variables (each variable can take only the values 0 and 1) and with a sample size the $n = 500$ all $2^{12} = 4096$ number of possible response patterns. On average, the expected frequency will be too small (0.122) for the χ^2 approximation to sampling distribution to be valid.

Sparseness is not restricted to the tables with smaller sample sizes alone, but could also occur with large sample sizes; this is due to the high concentration of frequencies in certain cells, with poor frequencies or none at all in others. It is clear that for such a sparse table an approximation with the asymptotic result is not appropriate.

The distorting effect of sparseness on the Chi-square test is well known; see for example, Mielker and Berry [23]. Among the order statistics, the likelihood ratio (G^2) appears to be the most susceptible to the effects of sparseness for the one-factor model with dichotomous variables [29]. In this sense too, moreover, Dayton [11] provides computational details for Pearson's χ^2, the likelihood ratio (G^2) and the Read-Cressie statistic (RC), concluding that the RC is the best option when there are small expected frequencies. Bartholomew and Tzamourani [5] proposed alternative ways for assessing the goodness-of-fit of the latent trait model for binary responses based on Monte Carlo methods and residual analysis.

This problem can be overcome using parametric bootstrap procedures to generate an empirical distribution of the model fit statistic and use this distribution to test the fit statistic from the original data [8, 1, 16]. In this sense, Tollenaar and Mooijaart [32] reported that the validity of the bootstrap is associated with the statistic used in the hypothesis test, because there are problems when estimating Type I error through the χ^2 and G^2 statistics. Based on a Monte Carlo study, von Davier (1997) concluded that bootstrap procedures work adequately for the χ^2 statistic and the RC fit statistic [34].

4 The patterns method

The parametric bootstrap is the most commonly used method for categorical data whenever the frequency table to be analyzed is sparse [35]. However, this requires both knowledge of advanced statistics and computationally intense methods because to obtain a stable p-value several hundred bootstrap resamples are needed for each model the researcher is interested in comparing. In this sense, there could be problems when certain parameter estimates in the parametric bootstrap are on the boundary of the parameter space, because the sampling is taken from the empirical distribution; a data pattern that is not observed in the sample has probability zero of being selected into the bootstrap samples and, consequently the estimated distribution may be too far from the true distribution [33].

We propose a new method, namely the Patterns Method, which solves the problems of the parametric bootstrap. The Patterns Method is an alternative for the latent class model diagnostic when the data are sparse. The basic idea behind this method is to take the total number of patterns possible $(R = 2^p)$ as the population and apply simple random sampling with replacement in order to simulate samples of similar size to the original sample for constructing the empirical probability distribution of the GFS.

This focus of the Patterns Method differs from previous ones, such as the non-parametric bootstrap, which builds the unknown probability distribution of the statistic by resampling of the original sample. Likewise, the Patterns Method differs from the parametric bootstrap, which uses the parameters of the latent class model to reproduce new data sets. A basic feature of the Patterns Method is the substitution of the underlying function of the unknown probability distribution $F(\mathbf{X})$ by an estimator. Sampling with replacement of the response patterns is used to obtain a large number of random samples to perform the estimation. The empirical probability distribution $\hat{F}(\mathbf{X}^*)$, obtained from the response patterns, assigns a probability of $1/R$ to each response pattern $\mathbf{x}_r$, for $r = 1, 2, \cdots, R$, where R is the total number of patterns $(R = 2^p)$ for p binary manifest variables.

The number of all the possible response patterns (R) is used as a starting point for simulation. Thus, the probability of random selection of a pattern is $1/R$. By iterating this process n times, we obtain a set of data that form the so-called random pattern sample. Thus, $\mathbf{x}_i^* = (x_1^*, x_2^*, \cdots, x_p^*)$ represents a response pattern that will be given by the matrix: $\mathbf{X}^* = [\mathbf{x}_1^*, \mathbf{x}_2^* \cdots \mathbf{x}_n^*]$. Following this, we obtain A random pattern samples (p.j. A=500), until they are considered acceptable to estimate the empirical probability distribution.

Accordingly, although each random pattern sample will have the same number of elements as the original sample and by random sampling with replacement, that sample may contain most of the patterns of the original sample, together with other new ones that are part of the population of patterns that were not observed in data acquisition. Thus, the proportion of response patterns in each resample is increased above the levels observed in the original sample, with the same amount of information to estimate the empirical distribution of GFS.

For each of these random samples of patterns, the GFS can be calculated, because the latent class model assumption is accepted as appropriate for the original sample data. In order to differentiate the goodness-of-fit statistics calculated on the values of the original sample $\hat{\theta}$ and the goodness-of-fit statistics for sample a, the latter will be denoted as $\hat{\theta}^*_a$. After selecting random samples of responses patterns, it is possible to estimate the empirical probability distribution of $\hat{\theta}^*_a$, assigning a probability of $1/A$ to each value of the statistic calculated: $\hat{\theta}^*_1, \hat{\theta}^*_2, \cdots, \hat{\theta}^*_A$.

This distribution is thus converted into an immediate estimator for the distribution function θ and can be used to test the validity of the latent class model hypothesized to describe the original data. In summary, the Patterns Method is executed in the following procedure for assessing the goodness-of-fit of a latent class model for binary data:

1. Fit model to the observed data and calculate the goodness-of-fit statistics $(\hat{\theta})$. This is called the original sample.

2. Generate one sample of the same size as the original data by simulating from possible response patterns.

3. Fit a model with the same structure as in Step 1. and calculate the goodness-of-fit statistics.

4. Repeat Steps 2 and 3 a great many times (e.g. $A = 500$) to approximate the distribution of the GFS, assuming that the fitted model is correct.

5. Reject the model if the *p-value* (pv) is larger than the significance α.

From this research, calculation of the level of significance was based on estimating the sampling distribution of the GFS under the hypothetical latent class model. If $\hat{\theta}$ is the value obtained in the original sample, the significance test will be to calculate how unusual $\hat{\theta}$ is with respect to the sampling distribution of $\hat{\theta}^*$. Thus, the significance level of the test statistic is,

$$pv = P\left[\hat{\theta}^* \geq \hat{\theta}\right] = \frac{\text{Number of times that } \left[\hat{\theta}^* \geq \hat{\theta}\right] + 1}{A+1}. \tag{2}$$

The decision rule regarding the hypothetical model will be to reject the latent class model if $pv < \alpha$, where α is the significance level set *a priori*. Thus, the one-sided significance level is simply the proportion of simulated samples in which the value of $\hat{\theta}^*$ is greater than or equal to the estimated $\hat{\theta}$ in the original sample. Furthermore, a small pv implies that the data of the original sample are implausible (or have a small probability of occurring) under the null hypothesis

In sum, the proposed Patterns Method aims to estimate the empirical probability distribution with a view to drawing inferences about the appropriate latent class model for the data of the original sample, although the mode of action is different from the parametric and non-parametric bootstrap techniques. Furthermore, it is worth mentioning that the Patterns Method does not rely on the assumption that the sample data are drawn from a given probability distribution, or on the assumption that the GFS has an χ^2 theoretical probability distribution.

5 Design of the study

In this section, we apply the Patterns Method to several data sets varying in their degree of sparseness. The ratio $d = n/2^p = n/R$ is used to measure the amount of sparseness present in a table, where p is the number of manifest variables, n is the size of the original sample and $R = 2^p$ is the number of response patterns; random samples were simulated, with sizes given by $n_k = k * R$, for $k = 1, \cdots, 5$. These degrees of sparseness in the data are the major problems to justify the asymptotic approximation of the GFS to the χ^2 theoretical distribution and have been used in previous investigations similar to the present one [4, 8].

We then simulated binary data corresponding to different sample sizes and, therefore, with varying degrees of density (or sparseness), which is a factor determinant on how well the distribution is represented by the χ^2. Specifically, the number of binary manifest variables varied from 5 to 9 and models with two to five latent classes were examined (original sample).

The detail of the simulated data sets is shown in Table 1. For example, for 7 binary manifest variables we simulated samples that had 2, 3 and 4 latent classes. As may be seen, the size of the samples, $n_1, \cdots, n_5$, were 128, 256, 384, 512 and 640, respectively.

The main goal of the simulation study presented here is to establish whether the Patterns Method can be used under different sparseness conditions. In order to do so, four frequently used goodness-of-fit statistics were chosen for this study, namely the likelihood-ratio (G^2), Pearson's (χ^2), the Freeman-Tukey (FT) and the Read-Cressie (RC) statistic. The verification process of the simulation is intended to show that the simulated model is confirmed as valid by the method of diagnosis.

Table 1: Sizes of samples associated with the degrees of density of the data according to the number of manifest variables and latent classes.

Manifest variables	Latent Class	Sizes of samples simulated				
p	C	n_1	n_2	n_3	n_4	n_5
5	2	32	64	96	128	160
5	3					
6	2	64	128	192	256	320
6	3					
7	2	128	256	384	512	640
7	3					
7	4					
8	2	256	512	768	1024	1280
8	3					
8	4					
9	2	512	1024	1536	2048	2560
9	3					
9	4					

The set up of the simulation study was as follows. We present illustrations using several simulated data sets varying in their degree of sparseness with the **CMABOOT**–a computational program developed using **MATLAB** language designed by the author of this article. It was designed in three stages: in the first stage, the original sample is simulated; in the second one, the pattern samples are simulated, and in the third phase, the Type I Error is calculated (i.e., the probability of incorrectly rejecting a true model). This implementation of computational algorithms proves to be a complex task, especially when it is necessary to process mathematical operations with hefty computational requirements.

The research methodology consisted first of assuming a latent class model in which the overall and conditional probabilities are known, after which, by means of a **MATLAB** application, a random sample is built. We shall call this the original sample. For the second step, in order to analyze the effectiveness of the method as regards correct determination of the model with which the data of the original sample were generated, we simulated $A = 500$ pattern samples. For each, we used a decision criterion or cut-off value of 5%, to determine how many values of the GFS are lower than the cut-off point and determine non-rejection of the null hypothesis (the number of latent classes of the original sample). In the third step the simulation experiment of the $A = 500$ pattern samples was repeated 100 times, calculating the probability of Type I Error (α), which is represented by the proportion of repeats in which the decision is the incorrect one, rejection of the null hypothesis being correct.

In sum, to carry out the research, we needed 1,500,000 replications of the experiment, which involved more than $2,500$ processing hours on 10 computers carried out simultaneously, with the following features: Intel (R) Core(TM) i5 2.40 GHz 4 GB RAM PC.

6 Results

This section presents summaries from the simulation study. We then analyzed the results of the simulations in terms of the probability of Type I error. The validation process consisted of ascertaining that the Patterns Method had low α probabilities for the four goodness-of-fit statistics considered under different sparseness conditions, $d = 1, 2, \cdots, 5$. We need consider only small values of d associated with distribution of the GFS which do not follow a probability chi-square distribution. Furthermore, these represent the degrees of density frequently studied by researchers to evaluate latent class models.

The most critical situation arises when the density of the data is very low ($d = 1$), meaning that most of the response patterns are not observed. As can be seen Table 2 shows that the most appropriate GFS proves to be the likelihood ratio statistic (G^2), followed by Read-Cressie (RC). Similarly, the Freeman-Tukey statistic(FT)has low probabilities of α in the models examined, except when there are 6 manifest variables and three latent classes, for which $\alpha = 0.15$ and the model for 5 manifest variables and two latent classes, where $\alpha = 0.07$. Hence, the Pearson statistic $\left(\chi^2\right)$ provides more unfavorable α values, for some of the models examined were greater than 5%. For example, for 9 variables and two latent classes the probability of Type I error is $\alpha = 0.23$.

Also, for a density of the original samples ($d = 2$) the results are very similar in contrast to the above analysis. The likelihood ratio statistic gives probabilities of α mostly of 0. The RC and FT statistics have α values lower than 0.05, with the exception of the model for 5 manifest variables and two latent classes, where with RC we have $\alpha = 0.06$. The Pearson statistic $\left(\chi^2\right)$ continued to give unsatisfactory results since the α probabilities were greater than expected.

In this same setting, for the density of data $d = 3$ the results show a similar trend to the two previous cases (Table 3). The G^2 and RC statistics have α values lower than 0.05, except for the model with 6 manifest variables and two latent classes, for which, with the RC statistic we have $\alpha = 0.06$. Freeman-Tukey (FT) gave acceptable α values $(\alpha < 0.05)$, except for the model with 7 variables and 4 latent classes, where we have $\alpha = 0.09$. Also, for 5 and 6 variables, both

Table 2: Probabilities of Type I error according to the number of manifest variables and latent classes for a data density of 1 $(n_1 = R = 2^p)$.

Manifest variables	Latent Class	G^2	χ^2	FT	RC
5	2	0.00	0.02	0.07	0.00
5	3	0.01	0.03	0.01	0.02
6	2	0.00	0.03	0.00	0.01
6	3	0.00	0.06	0.15	0.01
7	2	0.00	0.12	0.00	0.02
7	3	0.00	0.06	0.01	0.00
7	4	0.00	0.01	0.00	0.00
8	2	0.00	0.17	0.00	0.01
8	3	0.00	0.08	0.00	0.01
8	4	0.00	0.04	0.00	0.00
9	2	0.00	0.23	0.00	0.00
9	3	0.00	0.10	0.00	0.01
9	4	0.00	0.10	0.00	0.00
9	5	0.00	0.11	0.00	0.00

with 3 classes, the α values are 0.06. The Pearson statistic $\left(\chi^2\right)$ has type I error probabilities greater than 0.05 only for the models that have 5 manifest variables, with two and three latent classes $\alpha < 0.05$.

It was found that the the probabilities of Type I (α) error tend to vary when the value of the degree of density is $d = 4$. The Read-Cressie statistic (RC) and the likelihood ratio statistic (G^2) are those providing the most acceptable results, that is, α values less than 5%, except for the model with 6 variables and 3 latent classes, where for both statistics we have $\alpha = 0.08$. In nine out of fourteen models studied, the Freeman-Tukey statistic is no longer effective since it has probabilities greater than 0.05, the most critical being the models that have 3 and 4 latent classes. Pearson's statistic $\left(\chi^2\right)$ only provides acceptable values of $\alpha < 0.05$ in two simulated models, for 5 variables and 2 latent classes, as well as that composed of 9 variables and 3 latent classes.

The probability of Type I error when the degree of sparseness is $d = 5$ ($n/2^p = 5$) (Table 4) shows the most acceptable statistics that give the probabilities of α in Read-Cressie (RC) and in the likelihood ratio $\left(G^2\right)$ because the most frequent values of α are lower than 0.05. The Pearson statistic $\left(\chi^2\right)$ provides values that are not appropriate for use in the latent class models diagnosis for sparse tables by means of the Patterns Method; only with the model for 5 variables and 2 latent classes do we have a probability $\alpha = 0.05$. The results for

Table 3: Probabilities of Type I error according to the number of manifest variables and latent classes for a data density of 3 $(n_3 = 3 * 2^p)$.

Manifest variables	Latent Class	G^2	χ^2	FT	RC
5	2	0.00	0.05	0.00	0.02
5	3	0.04	0.04	0.06	0.04
6	2	0.02	0.11	0.02	0.06
6	3	0.04	0.08	0.06	0.05
7	2	0.00	0.13	0.00	0.02
7	3	0.00	0.10	0.00	0.03
7	4	0.02	0.09	0.09	0.05
8	2	0.00	0.14	0.00	0.01
8	3	0.00	0.20	0.00	0.01
8	4	0.00	0.10	0.05	0.02
9	2	0.00	0.24	0.00	0.02
9	3	0.00	0.13	0.04	0.00
9	4	0.00	0.11	0.00	0.00
9	5	0.00	0.10	0.00	0.01

the Freeman-Tukey statistic, however, show that this test is less effective, since in most models analyzed $\alpha > 0.05$, only in the model with 9 variables and 2 latent classes do we have $\alpha = 0.00$.

Finally, we compared the Parametric Bootstrap with our proposal as regards the behavior of the magnitude of Type I error in order to determine whether there were differences between them in the models analyzed. For example, as illustrated in (Table 5),for $d = 2$ density data degree, the magnitude of α for the Patterns Method is smaller in all the models than those obtained using the Parametric Bootstrap with respect to the G^2, FT and RC statistics. We also see that the α values are lower than 5%, a situation not found with the Parametric Bootstrap, where it is observed that many values of α are higher than the expected value ($\alpha = 0.05$), since it is the quota of type 1 error fixed on performing the significance tests upon each of the 100 replicates. However, using the Pearson statistic in all models the Parametric Bootstrap has Type I error values lower than those obtained with the Patterns Method. For some models the α obtained with the Bootstrap are greater than expected ($\alpha > 0.05$). For example, for the model with 6 variables and 2 latent classes we obtain the value $\alpha = 0.14$.

Table 4: Probabilities of type I error according to the number of manifest variables and latent classes for a data density of 5 $(n_5 = 5 * 2^p)$.

Manifest variables	Latent Class	G^2	χ^2	FT	RC
5	2	0.04	0.05	0.14	0.04
5	3	0.10	0.08	0.18	0.08
6	2	0.05	0.11	0.23	0.05
6	3	0.11	0.08	0.26	0.07
7	2	0.00	0.21	0.11	0.04
7	3	0.03	0.15	0.23	0.04
7	4	0.17	0.10	0.65	0.10
8	2	0.00	0.18	0.08	0.04
8	3	0.01	0.10	0.32	0.03
8	4	0.02	0.11	0.31	0.03
9	2	0.00	0.13	0.00	0.00
9	3	0.09	0.06	0.95	0.03
9	4	0.00	0.15	0.61	0.01
9	5	0.05	0.15	0.86	0.05

Table 5: Comparison of Parametric Bootstrap and the Patterns Method according of Type I error, considering the number of manifest variables, the number of latent classes and a data density of 2 $(d_2 = 2 * 2^p)$.

Manifest	Latent	Parametric Bootstrap				Method Patterns			
variables (p)	Class (c)	G^2	χ^2	FT	RC	G^2	χ^2	FT	RC
5	2	0.07	0.09	0.06	0.07	0.01	0.09	0.04	0.06
5	3	0.03	0.04	0.03	0.04	0.01	0.05	0.02	0.01
6	2	0.03	0.14	0.02	0.13	0.00	0.21	0.00	0.04
6	3	0.08	0.06	0.06	0.09	0.00	0.09	0.02	0.03
7	2	0.03	0.11	0.03	0.09	0.00	0.21	0.00	0.03
7	3	0.08	0.06	0.06	0.06	0.00	0.09	0.01	0.02
7	4	0.05	0.05	0.06	0.05	0.00	0.07	0.01	0.02
6	3	0.08	0.06	0.06	0.09	0.00	0.09	0.02	0.03
8	2	0.07	0.05	0.08	0.05	0.00	0.17	0.00	0.10
8	3	0.06	0.07	0.09	0.08	0.00	0.10	0.00	0.00
8	4	0.04	0.07	0.05	0.03	0.00	0.10	0.00	0.01
9	2	0.04	0.06	0.04	0.05	0.00	0.21	0.00	0.00
9	3	0.05	0.05	0.05	0.07	0.00	0.13	0.00	0.00
9	4	0.04	0.04	0.06	0.04	0.00	0.12	0.00	0.00
9	5	0.08	0.08	0.06	0.05	0.00	0.08	0.00	0.00

7 Statistical tables of critical values

The Patterns Method does not derive from supposed parametrics with respect to the probability distribution of the original sample's data. On the contrary, it only generates data groups of identical size as the original sample, using the same number of manifest binary variables in order to estimate the empirical distribution of the GFS. Thus, the construction of statistical tables of critical values for GFS is made possible in order to contrast a hypothetical latent class model with sparse data which is suitable for the data from the original sample.

Due to the fact that sample sizes can vary widely, in the statistical tables we have decided to represent the n/R factor, where n is the size of the original sample, and R is the number of response patterns. The particular value n/R represents the estimate of the density data for the fit of the latent class models. The tabulated values of n/R are comprised from 1 and 10.

In the tables of critical values, accumulated probabilities are shown in the top row; the rationale for n/R appears in the first column, followed by the GFS. The intersection of the row with the column corresponds to the θ goodness-of-fit statistic. The level of theoretical significance is calculated as, $p_{value} = 1 - P(\theta < \theta^*)$.

In order to use the tables of critical values, this procedure should be followed:

- Establish the null hypothesis (H_0) for the latent classes.
- Adjust the latent class model to obtain the goodness-of-fit statistic model values of the sample.
- Calculate the rationale for n/R and find it in the table.
- For each goodness-of-fit statistic determine the p-value.
- Discard (H_0) if the fixed level of significance (α) is greater than the p-value.

For these purposes, two tables of critical values are presented to demonstrate the application of this approach to contrast a hypothetical latent class model with sparse data. The tables of critical values below present 6 manifest variables and 2 latent classes (see Table 6), as well as 8 manifest variables and 3 latent classes (see Table 7).

Tables of critical values of the Patterns Method are simple and practical alternative to the Parametric Bootstrap for the diagnosis of latent class models with sparse data. This facilitates the study of problems in the framework of latent class models with binary variables in sparse data. However, the critical value should be viewed with caution when n/R represented in the table is very different from the real value.

Table 6: The tables of critical values: 6 manifest variables and 2 latent classes

n/R	GFS	Cumulative Probabilities								
		0,25	0,50	0,75	0,90	0,95	0,96	0,97	0,98	0,99
	G^2	51,7	57,3	63,0	68,2	71,0	72,2	73,3	74,6	76,8
1	χ^2	42,8	48,3	54,0	60,2	63,7	65,1	67,2	69,1	72,4
	FT	81,9	90,3	99,1	106,6	110,7	112,5	113,0	116,1	118,6
	RC	43,1	48,4	53,4	58,9	61,9	63,2	65,1	66,4	69,8
	G^2	51,5	58,0	66,4	73,0	77,3	78,3	79,9	82,3	86,9
2	χ^2	44,2	49,9	56,9	63,5	68,7	70,0	72,7	74,3	77,5
	FT	73,3	82,8	96,2	108,6	114,0	116,7	119,2	123,6	129,7
	RC	44,6	50,3	57,5	63,3	68,1	69,3	71,5	72,8	76,5
	G^2	45,1	53,1	61,5	69,5	73,7	75,0	78,1	79,1	83,7
3	χ^2	40,3	47,2	54,7	61,7	64,7	66,1	68,0	69,7	74,2
	FT	56,2	68,5	80,9	92,5	100,8	105,9	109,8	114,0	115,5
	RC	40,7	47,5	55,1	62,2	65,8	66,3	67,2	70,3	74,2
	G^2	43,3	50,1	57,8	64,1	68,0	69,1	71,0	73,8	76,4
4	χ^2	40,6	45,6	52,6	58,5	62,4	63,1	65,4	67,9	72,8
	FT	48,7	57,8	70,2	79,9	87,1	89,1	90,5	94,9	99,2
	RC	40,8	46,2	53,3	58,8	61,9	63,7	65,4	68,2	71,8
	G^2	42,7	48,8	55,2	63,7	66,0	67,0	69,2	71,7	76,0
5	χ^2	40,9	46,1	52,1	58,3	62,4	64,0	65,5	68,2	71,1
	FT	45,2	53	62,4	72,9	78,4	79,8	82,0	86,8	97,1
	RC	41,1	46,5	52,4	59,0	62,6	64,4	65,1	68,3	70,6
	G^2	42,2	48,6	56,0	63,5	67,7	68,7	69,8	72,1	75,1
6	χ^2	40,6	46,6	54,3	61,0	64,9	65,9	66,8	68,3	72,1
	FT	44,1	51,1	59,6	70,2	74,7	75,8	77,8	81,7	87,1
	RC	41,0	46,9	54,5	60,8	65,2	65,8	67,4	69,0	72,6
	G2	41,8	47,6	54,3	62,4	68,3	69,0	70,7	74,1	78,3
7	χ^2	40,6	46,1	52,3	60,7	64,4	65,8	67,8	70,0	73,0
	FT	43,7	50,2	57,2	67,1	73,8	78,3	79,7	82,0	87,5
	RC	40,8	46,4	52,7	60,4	65,2	66,1	67,4	71,1	73,1
	G^2	40,5	47,6	53,7	61,5	65,2	68,5	68,7	74,6	78,4
8	χ^2	39,7	45,8	51,9	59,3	63,0	64,4	66,5	71,2	75,1
	FT	41,8	49,2	56,2	64,6	69,6	71,6	74,6	80,4	86,7
	RC	39,7	46,1	52,3	60,0	63,2	64,4	66,8	71,9	75,5
	G^2	41,0	47,6	55,2	62,2	67,1	69,0	70,0	72,2	76,3
9	χ^2	40,4	46,8	54,1	59,7	64,7	66,3	68,4	70,0	73,3
	FT	42,2	49,2	57,0	64,9	70,7	72,9	74,5	75,9	82,3
	RC	40,5	46,5	54,3	59,9	64,8	67,0	68,3	69,8	72,8
	G^2	40,9	47,4	54,4	61,8	65,0	66,1	67,5	69,9	74,2
10	χ^2	40,3	46,6	53,1	60,3	63,0	63,3	64,8	68,2	72,9
	FT	41,7	48,4	56,2	63,7	67,9	69,4	70,9	73,6	78,5
	RC	40,5	46,5	53,5	60,5	63,1	63,7	64,7	69,2	72,8

Table 7: The tables of critical values: 8 manifest variables and 3 latent classes.

n/R	GFS	Cumulative Probabilities								
		0,25	0,50	0,75	0,90	0,95	0,96	0,97	0,98	0,99
1	G^2	260,1	272,3	284,9	292,6	296,6	299,2	303,0	306,2	313,3
	χ^2	221,9	235,7	249,9	262,3	269,7	271,3	273,5	277,5	285,6
	FT	414,5	431,4	448,5	459,2	464,4	468,9	474,8	480,7	487,9
	CR	220,8	232,7	245,9	255,3	260,6	262,0	264,3	269,8	275,6
2	G^2	251,2	265,5	281,3	295,5	304,1	306,2	308,8	310,7	316,1
	χ^2	218,4	231,1	244,6	257,8	265,6	268,1	269,9	276,8	281,3
	FT	359,6	384,3	411,4	434,7	446,8	452,8	455,8	458,1	459,8
	CR	219,5	231,3	244,9	257,1	265,6	267,2	270,0	273,9	278,1
	G^2	236,9	251,7	264,8	280,1	292,4	295,2	300,2	303,4	307,0
	χ^2	214,8	226,8	240,4	254,2	262,4	264,1	268,2	274,8	281,1
	FT	298,3	324,4	345,9	367,8	388,8	393,6	399,4	404,9	415,5
	CR	215,9	228,1	240,5	255,3	262,2	266,6	270,4	274,4	280,0
4	G^2	225,4	242,2	258,8	272,7	285,2	289,5	292,1	296,6	299,8
	χ^2	212,4	226,9	241,1	254,8	263,9	265,8	270,3	273,3	277,0
	FT	259,0	281,7	306,3	332,2	351,5	359,3	365,1	373,6	381,2
	CR	213,2	227,9	241,6	254,6	264,6	268,2	271,8	273,1	276,2
5	G^2	219,9	236,6	251,5	264,6	272,9	277,3	279,5	286,2	293,9
	χ^2	209,5	225,0	238,9	251,4	260,5	263,0	266,1	269,3	275,8
	FT	239,5	260,9	282,2	303,0	316,3	320,0	323,5	327,5	336,0
	CR	210,5	225,4	239,6	251,3	260,0	262,1	266,0	269,3	278,2
6	G2	223,0	235,1	253,0	264,6	271,5	275,4	277,1	283,0	298,5
	χ^2	214,2	226	242,1	254,0	262,5	265,3	266,7	273,1	277,6
	FT	234,6	253,1	273,8	289,8	297,4	306,8	311,3	318,6	339,1
	CR	215,1	226,7	243,2	254,4	261,5	264,1	266,2	272,1	278,2
7	G^2	218,4	232,2	247,9	260,4	271,3	274,1	280,3	283,4	291,3
	χ^2	212,2	225,3	239,7	252,6	262,1	266,2	270,3	277,5	284,1
	FT	227,1	244,0	262,6	277,8	291,5	295,6	300,5	303,4	312,8
	CR	212,5	225,5	240,4	253,1	263,4	265,7	270,9	275,5	282,3
8	G^2	216,7	231,2	244,6	255,7	265,3	267,8	271,0	273,5	277,0
	χ^2	211,8	225	238,7	251,2	258,4	260,4	263,5	265,1	270,2
	FT	224,1	240,6	255,5	269,9	279,4	281,9	283,8	288,9	296,9
	CR	212,1	225,6	238,7	250,6	258,9	261,6	262,8	265,2	270,1
9	G^2	214,7	232,2	245,5	261,7	270,3	271,7	274,9	278,2	284,5
	χ^2	210,3	226,5	241,0	254,6	264,8	266,8	268,9	270,6	280,5
	FT	222,5	240,0	255,8	274,3	281,5	282,4	285,0	290,2	300,4
	CR	210,4	226,8	241,1	255,4	264,5	266,6	268,5	271,5	279,5
10	G^2	216,1	229,0	245,0	259,1	266,3	268,7	271,1	275,2	278,6
	χ^2	211,6	225,4	239,7	252,6	262,0	265,1	267,6	270,3	274,1
	FT	222,0	236,3	254,2	267,9	276,2	278,5	280,8	285,0	289,5
	CR	211,5	225,6	240,5	252,8	261,0	265,0	266,6	269,8	273,3

8 Conclusions

In this study, we proposed a much faster alternative, which uses Patterns Method samples to construct the sampling distributions of the test statistic in sparse contingency tables where the number of response patterns, R, is large compared to sample size.

From the results of the simulations, for the latent class models with binary manifest variables, using densities of $d \leq 5$, we have shown that the Type I error probabilities are lower than 5% ($\alpha < 0.05$) for the likelihood ratio (G^2) and Read-Cressie (RC) statistics, using our Patterns Method proposal. In light of this, we recommend the diagnosis of latent class models for sparse data using these statistics. In contrast, the Freeman-Tukey statistic provides acceptable results when the data density are $d \leq 4$.

In the case of data density $d \leq 5$, Pearson's statistic $\left(\chi^2\right)$ should not be used to select latent class models using the Patterns Method, given that this has the probability of Type I error being greater than 5%. In the same way, Langeheine et al. [20] in the context of parametric bootstrapping concluded that the Pearson statistic puts a much more severe penalty on an observation in a cell with a very low model-expected probability than the likelihood ratio G^2 does.

As a side product, by comparing the Patterns Method and the Parametric Bootstrap for data density $d = 2$, we show that the Patterns Method has more accurate Type I error probabilities since the likelihood ratio, Read-Cressie and Freeman-Tukey statistics afford values of $\alpha < 0.05$. In contrast, the Parametric Bootstrap provides values in these statistics that surpass 5%. But further study is required to be certain about these results.

The parametric bootstrap require both knowledge of advanced statistics and this method is computationally intense since in order to obtain a stable p-value several hundred bootstrap resamples are needed for each model the researcher is interested in comparing. Meanwhile, the Patterns Method is presented as a rapid, simple, and labour-saving technique to provide tables of critical values to diagnose latent class models.

Finally, for future research, the Patterns Method will be tested for the analysis of ordinal data, in order to study the effectiveness of the latent class model on sparse data.

Acknowledgments

The author thanks the Sede de Occidente of the Universidad de Costa Rica for assistance in pursuing his PhD studies, and for a grant to stay at the University of Salamanca (Spain) while this work was under preparation.

References

[1] Agresti, A. (2007) *An Introduction to Categorical Data Analysis*, 2nd Edition. Wiley Interscience, Hoboken NJ.

[2] Agresti, A.; Yang, M.C. (1987) "An empirical investigation of some effects of sparseness in contingency tables", *Computational Statistics & Data Analysis* **5**(1): 9–21.

[3] Bartholomew, D.J.; Knott, M.; Moustaki, I. (2011) *Latent Variable Models and Factor Analysis: A Unified Approach*. John Wiley & Sons, Chichester UK.

[4] Bartholomew, D.J.; Leung, S.O. (2002) "A goodness of fit test for sparse 2p contingency tables", *British Journal of Mathematical and Statistical Psychology*, **55**(1): 1–15.

[5] Bartholomew, D.J.; Tzamourani, P. (1999) "The goodness of fit of latent trait models in attitude measurement", *Sociological Methods & Research* **27**(4): 525–546.

[6] Cochran, W.G. (1952) "The χ^2 test of goodness of fit", *The Annals of Mathematical Statistics* **23**(3): 315–345.

[7] Cochran, W.G. (1954) "Some methods for strengthening the common χ^2 tests", *Biometrics* **10**(4): 417–451.

[8] Collins, L.M.; Fidler, P.L.; Wugalter, S.E.; Long, J.D. (1993) "Goodness-of-fit testing for latent class models", *Multivariate Behavioral Research* **28**(3): 375–389.

[9] Cramér, H. (1946) *Mathematical Methods of Statistics*. Princeton University Press, New York.

[10] Davison, A.C.; Fraser, D. ; Reid, N.; Sartori, N. (2013) "Accurate directional inference for vector parameters in linear exponential families", *Journal of the American Statistical Association* **109**: 302-314.

[11] Dayton, C.M. (1998) *Latent Class Scaling Analysis*. Sage Publications, Thousand Oaks CA.

[12] Dias, J.G.; Vermunt, J.K. (2006) "Bootstrap methods for measuring classification uncertainty in latent class analysis", in: *Compstat 2006-Proceedings in Computational Statistics*, Physica-Verlag HD: 31–41.

[13] Fisher, R.A. (1941) *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

[14] Gong, H. (2012) *Modeling and Measuring Association for Ordinal Data*. M.Sc. dissertation, Faculty of Graduate Studies and Research, University of Regina, Canada.

[15] Kendall, M.G. (1952) *The Advanced Theory of Statistics. Vol. 1: Distribution Theory*, 5th edition. Griffin, London.

[16] Kojadinovic, I.; Yan, J. (2012) "Goodness-of-fit testing based on a weighted bootstrap: A fast large-sample alternative to the parametric bootstrap", *Canadian Journal of Statistics* **40**(3): 480–500.

[17] Kraus, K. (2012) *On the Measurement of Model Fit for Sparse Categorical Data*. Doctoral dissertation, Disciplinary Domain of Humanities and Social Sciences, Faculty of Social Sciences, Department of Statistics, Uppsala University.

[18] Kunihama, T.; Dunson, D.B. (2013) "Bayesian modeling of temporal dependence in large sparse contingency tables", *Journal of the American Statistical Association* **108**(504): 1324–1338.

[19] Lancaster, H.O.; Seneta, E. (1969) "Chi-square distribution", in: *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd. Florida.

[20] Langeheine, R.; Pannekoek, J.; Van de Pol, F. (1996) "Bootstrapping goodness-of-fit measures in categorical data analysis", *Sociological Methods & Research* **24**(4): 492–516.

[21] Larntz, K. (1978) "Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics", *Journal of the American Statistical Association* **73**(362): 253–263.

[22] Lazarsfeld, P.F.; Henry, N.W. (1968) *Latent Structure Analysis*. Houghton Mifflin, Boston.

[23] Mielke P.W.; Berry, K.J. (2002) "Categorical independence tests for large sparse r-way contingency tables", *Perceptual and Motor Skills* **95**(2): 606–610.

[24] Milovanovic, J. (2011) *Chi-Square Orthogonal Components for Assessing Goodness-of-fit of Multidimensional Multinomial Data*. Doctoral dissertation, Arizona State University.

[25] Nylund, K.L.; Asparouhov, T.; Muthén, B.O. (2007) "Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study", *Structural Equation Modeling* **14**(4): 535–569.

[26] Papoulis, A.; Pillai, S.U. (2002) *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill Education.

[27] Radavičius, M.; Samusenko, P. (2011) "Profile statistics for sparse contingency tables under Poisson sampling", *Austrian Journal of Statistics* **40**(1-2): 115–123.

[28] Radavičius, M.; Samusenko, P. (2012) "Goodness-of-fit tests for sparse nominal data based on grouping", *Nonlinear Analysis: Modeling and Control* **17**(4): 489–501.

[29] Reiser, M.; Lin, Y. (1999) "A goodness-of-fit test for the latent class model when expected frequencies are small", *Sociological methodology* **29**(1): 81–111.

[30] Samusenko, P. (2012) *Nonparametric Criteria for Sparse Contingency Tables*. Doctoral dissertation, Vilnius Gediminas Technical Univerty, Lithuania.

[31] Tate, M.W.; Hyer, L.A. (1973) "Inaccuracy of the χ^2 test of goodness of fit when expected frequencies are small", *Journal of the American Statistical Association* **68**(344): 836–841.

[32] Tollenaar, N.; Mooijaart, A. (2003) "Type I errors and power of the parametric bootstrap goodness-of-fit test: full and limited information", *British Journal of Mathematical and Statistical Psychology* **56**(2): 271–288.

[33] Van Der Heijden, P.; Hart, H.; Dessens, J. (1997) "A parametric bootstrap procedure to perform statistical tests in a LCA of anti-social behaviour", in: J. Rost et al. (Eds.) *Applications of Latent Trait and Latent Class Models in the Social Sciences*, University of Michigan Library, Ann Arbor: 196–208.

[34] Von Davier, M. (1997) "Bootstrapping goodness-of-fit statistics for sparse categorical data-results of a Monte Carlo study", *Methods of Psychological Research* **2**(2): 29–48.

[35] Van Kollenburg, G.; Mulder, J.; Vermunt, K. (2015) "Assessing model fit in latent class analysis when asymptotics do not hold methodology", *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* **11**(2): 65–79.