Monografías del Semin. Matem. García de Galdeano. **27:** 469–482, (2003).

Loglinear Models: An approach based on ϕ-Divergences

Leandro Pardo

Department of Statistics and O.R., Complutense University of Madrid.

e-mail: Leandro_Pardo@Mat.ucm.es

Abstract

In this paper we present a review of some results about inference based on ϕ-divergence measures, under assumptions of multinomial sampling and loglinear models. The minimum ϕ-divergence estimator, which is seen to be a generalization of the maximum likelihood estimator is considered. This estimator is used in a ϕ-divergence measure which is the basis of new statistics for solving three important problems of testing regarding loglinear models: Goodness-of-fit, nested sequence of loglinear models and nonadditivity in loglinear models.

Keywords: Chi-squared distribution; contiguous alternatives; multinomial distribution; nested hypotheses; noncentral chi-squared distribution; ϕ-divergence statistic, power-divergence statistic.

1 Introduction

Let $Y_1, Y_2, ..., Y_n$ be a sample of size $n \geq 1$, with realizations from $\mathcal{X} = \{1, 2, ..., M\}$ and independent and identically distributed (i.i.d.) according to a probability distribution $\boldsymbol{p}(\theta_0)$. This distribution is assumed to be unknown, but belonging to a known family

$$\mathcal{P} = \left\{ \boldsymbol{p}(\theta) = (p_1(\theta), ..., p_k(\theta))^T : \theta \in \Theta \right\},$$

of distributions on $\mathcal{X}$ with $\Theta \subseteq \mathbb{R}^{M_0}$ $(M_0 < M - 1)$, and

$$\mathcal{P} \subset \Delta_M = \left\{ \boldsymbol{p} = (p_1, ..., p_M)^T : 0 < p_i < 1, \sum_{i=1}^{M} p_i = 1 \right\}.$$

In other words, the true value θ_0 of parameter $\theta = (\theta_1, ..., \theta_{M_0})^T \in \Theta \subseteq \mathbb{R}^{M_0}$ is assumed to be fixed but unknown. We denote $\boldsymbol{p} = (p_1, ..., p_M)^T$ and $\widehat{\boldsymbol{p}} = (\widehat{p}_1, ..., \widehat{p}_M)^T$ with

$$\widehat{p}_j = \frac{N_j}{n} \text{ and } N_j = \sum_{i=1}^{n} I_{\{j\}}(Y_i)\,;\; j = 1, ..., M. \tag{1}$$

The statistic $(N_1, ..., N_M)$ is obviously sufficient for the statistical model under consideration and is multinomially distributed; that is,

$$P(N_1 = n_1, ..., N_M = n_M) = \frac{n!}{x_1!...x_M!} p_1(\theta)^{n_1} \times ... \times p_M(\theta)^{n_M}, \tag{2}$$

for integers $n_1, ..., n_M \geq 0$ such that $n_1 + ... + n_M = n$.

In what follows, we assume that $\boldsymbol{p}(\theta)$ belongs to the general class of loglinear models. That is, we assume :

$$p_u(\theta) = \exp\left(w_u^T \theta\right) / \sum_{v=1}^{M} \exp\left(w_v^T \theta\right);\ u = 1, ..., M, \tag{3}$$

where the $M \times M_0$ matrix $W = (w_1, ..., w_M)^T$ is assumed to have full column rank $M_0 < M - 1$ and columns linearly independent of the $M \times 1$ column vector $(1, ..., 1)^T$. This will be the model we shall consider for the theoretical results in the next sections.

If we denote by $u^* = -\log\left(\sum\limits_{v=1}^{k} \exp\left(w_v^T \theta\right)\right)$, we can consider the matricial expression of the loglinear model given in (3) given by,

$$\log \boldsymbol{p}(\theta^*) = X\theta^* \tag{4}$$

where X is a $M \times (M_0 + 1)$ matrix with

$$X = (\mathbf{1}_{M \times 1}, W_{M \times M_0}),$$

$\log \boldsymbol{p}(\theta^*) = (\log p_1(\theta^*), ..., \log p_M(\theta^*))^T$ and $\theta^* = (u^*, \theta_1, ..., \theta_{M_0})^T$. We can express the loglinear model (4) by

$$\log \boldsymbol{m}(\theta^{**}) = X\theta^{**},$$

where $m(\theta^{**}) = np(\theta^*)$, $\theta^{**} = (u, \theta_1, ..., \theta_t)^T$ with $u = u^* + \log n$ and

$$\log \boldsymbol{m}(\theta^{**}) = (\log m_1(\theta^{**}), ..., \log m_M(\theta^{**}))^T.$$

Therefore given a $M \times t_A$ matrix X_A with rank$(X_A) = t_A$, the set

$$C(X_A) = \left\{\log \boldsymbol{p}(\theta) : \log \boldsymbol{p}(\theta) = X_A\theta;\ \theta \in \mathbb{R}^{t_A}\right\},$$

represent the class of loglinear models associated to the matrix X_A. An important assumption for the purpose of normalization is that the $M \times 1$ dimensional vector $J_M \equiv (1, ..., 1)^T \in C(X_A)$. We can observe that $\boldsymbol{p}(\theta) \in C(X_A)$ is equivalent to $\log \boldsymbol{m}(\theta) \in C(X_A)$ with $\boldsymbol{m}(\theta) = n\boldsymbol{p}(\theta)$ because

$$\log \boldsymbol{m}(\theta) = \log \boldsymbol{p}(\theta) + \log n J_M.$$

Kullback-Leibler divergence measure, between two loglinear models $\boldsymbol{p}(\theta_1)$ and $\boldsymbol{p}(\theta_2) \in \mathcal{P}$ and verifying (3) is given by

$$D_{Kull}(\boldsymbol{p}(\theta_1), \boldsymbol{p}(\theta_2)) = \sum_{i=1}^{k} p_i(\theta_1) \log \frac{p_i(\theta_1)}{p_i(\theta_2)}.$$

But this measure of divergence is a particular case of the $\phi-$divergence measures defined by Csiszár (1963) and Ali and Silvey (1966), by

$$D_\phi(\boldsymbol{p}(\theta_1), \boldsymbol{p}(\theta_2)) \equiv \sum_{i=1}^{k} p_i(\theta_2)\, \phi\left(\frac{p_i(\theta_1)}{p_i(\theta_2)}\right); \phi \in \Phi^*, \tag{5}$$

where Φ^* is the class of all convex functions $\phi(x)$, $x > 0$, such that at $x = 1, \phi(1) = 0, \phi''(1) > 0$, and at $x = 0$, $0\phi(0/0) = 0$ and $0\phi(p/0) = \lim_{u\to\infty} \phi(u)/u$. For every $\phi \in \Phi^*$ that is differentiable at $x = 1$, the function

$$\psi(x) \equiv \phi(x) - \phi'(1)(x-1)$$

also belongs to Φ^*. Then we have $D_\psi(\boldsymbol{p}(\theta_1), \boldsymbol{p}(\theta_2)) = D_\phi(\boldsymbol{p}(\theta_1), \boldsymbol{p}(\theta_2))$, and ψ has the additional property that $\psi'(1) = 0$. Because the two divergence measures are equivalent, we can consider the set Φ^* to be equivalent to the set

$$\Phi \equiv \Phi^* \cap \{\phi : \phi'(1) = 0\}.$$

In what follows, we give our theoretical results for $\phi \in \Phi$, but we often apply them to choices of functions in Φ^*. In the next several paragraphs, we give the essential details of the framework for estimation and hypothesis testing on loglinear models based on ϕ-divergences. We can observe that for $\phi(x) = x\log x - x + 1$ we obtain the Kullback-Leibler' divergence.

An important family of $\phi-$divergences in statistical problems is the power-divergence family,

$$\begin{aligned} &\phi_{(\lambda)}(x) = (\lambda(\lambda+1))^{-1}\left(x^{\lambda+1} - x\right);\ \lambda \neq 0, \lambda \neq -1, \\ &\phi_{(0)}(x) = \lim_{\lambda\to 0} \phi_{(\lambda)}(x) = x\log x - x + 1, \\ &\phi_{(-1)}(x) = \lim_{\lambda\to -1} \phi_{(\lambda)}(x) = \log x - x + 1, \end{aligned} \tag{6}$$

which was introduced and studied by Cressie and Read (1984). We can observe that the functions $\phi_{(\lambda)}(x)$ and $\psi_{(\lambda)}(x) \equiv \phi_{(\lambda)}(x) - (x-1)(\lambda+1)^{-1}$define the same divergence measure. In the following, we shall denote the power-divergence measures by,

$$I^\lambda(\boldsymbol{p}(\theta_1), \boldsymbol{p}(\theta_2)) \equiv D_{\phi_{(\lambda)}}(\boldsymbol{p}(\theta_1), \boldsymbol{p}(\theta_2)) = D_{\psi_{(\lambda)}}(\boldsymbol{p}(\theta_1), \boldsymbol{p}(\theta_2)).$$

We can observe that

$$I^0(\boldsymbol{p}(\theta_1), \boldsymbol{p}(\theta_2)) = D_{\phi_{(0)}}(\boldsymbol{p}(\theta_1), \boldsymbol{p}(\theta_2))$$

coincides with $D_{Kull}(\boldsymbol{p}(\theta_1), \boldsymbol{p}(\theta_2))$.

2 Minimum $\phi-$divergence estimators under the Loglinear Model

In this Section we present some asymptotic results for the minimum ϕ-divergence estimator under the loglinear model (3). It is well-known that the Fisher information matrix in the multinomial model considered (2) is given by $\boldsymbol{I}_F(\theta) = A(\theta)^T A(\theta)$, where $A(\theta)$ is a $M \times M_0$ matrix given by

$$A(\theta) = diag\left(\boldsymbol{p}(\theta)^{-\frac{1}{2}}\right)_{M\times M}\left(\frac{\partial p_i(\theta)}{\partial \theta_r}\right) \quad i = 1, ..., M, \quad r = 1, ..., M_0.$$

For the loglinear models we have

$$\frac{\partial p_i(\theta)}{\partial \theta_r} = p_j(\theta)\, w_{rj} - p_j(\theta)\sum_{v=1}^{k} w_{rv} p_v(\theta).$$

Then

$$\frac{\partial \boldsymbol{p}(\theta)}{\partial(\theta)} = \left(diag(\boldsymbol{p}(\theta)) - \boldsymbol{p}(\theta)\boldsymbol{p}(\theta)^T\right)W = \Sigma_{\boldsymbol{p}(\theta)}W$$

and hence $A(\theta) = diag\left(\boldsymbol{p}(\theta)^{-\frac{1}{2}}\right)_{M\times M}\Sigma_{\boldsymbol{p}(\theta)}W$. Then the Fisher information matrix for a loglinear model is given by

$$\boldsymbol{I}_F(\theta) = W^T \Sigma_{\boldsymbol{p}(\theta)} W.$$

It is also well-known that the maximum likelihood estimator for multinomial model considered in (2) can be obtained as the value $\widehat{\theta} \in \Theta$ minimizing the Kullback-Leibler divergence measure, with respect to θ, between the loglinear model $\boldsymbol{p}(\theta) \in \mathcal{P}$ verifying (3) and the nonparametric estimator of the model $\widehat{\boldsymbol{p}}$, i.e., the estimator of the saturated model. As a generalization of the maximum likelihood estimator we can consider the minimum ϕ-divergence estimator as the value $\widehat{\theta}_\phi \in \Theta$ minimizing, with respect to θ, the ϕ-divergence measure, $D_\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\theta))$, i.e.,

$$\widehat{\theta}_\phi \equiv \arg\min_{\theta\in\Theta} D_\phi(\widehat{\boldsymbol{p}}, \boldsymbol{p}(\theta)). \tag{7}$$

Cressie, N. and Pardo, L (2000) established, based on a previous result given in Morales, D. et al. (1995) for multinomial models, the following BAN descomposition for the minimum ϕ-divergence estimator, $\widehat{\theta}_\phi$, of the parameter θ in the loglinear model $\boldsymbol{p}(\theta)$,

$$\widehat{\theta}_\phi = \theta_0 + \boldsymbol{I}_F(\theta_0)^{-1}\Sigma_{\boldsymbol{p}(\theta_0)} diag\left(\boldsymbol{I}(\theta_0)^{-1}\right)(\widehat{\boldsymbol{p}} - \boldsymbol{p}(\theta_0)) + o_P(\parallel \widehat{\boldsymbol{p}} - \boldsymbol{p}(\theta_0) \parallel) \tag{8}$$

where $\Sigma_{\boldsymbol{p}(\theta_0)} = diag(\boldsymbol{p}(\theta_0)) - \boldsymbol{p}(\theta_0)\boldsymbol{p}(\theta_0)^T$. Based on this result it was also established that

$$n^{1/2}\left(\widehat{\theta}_\phi - \theta_0\right) \xrightarrow[n\to\infty]{L} N\left(\boldsymbol{0}, \left(W^T\Sigma_{\boldsymbol{p}(\theta_0)}W\right)^{-1}\right).$$

Another interesting result, useful later, is the following

$$n^{1/2}\left(\boldsymbol{p}\left(\widehat{\theta}_{\phi}\right)-\boldsymbol{p}\left(\theta_{0}\right)\right)\xrightarrow[n\to\infty]{L}N\left(\mathbf{0},\Sigma_{\boldsymbol{p}(\theta_{0})}W\left(W^{T}\Sigma_{\boldsymbol{p}(\theta_{0})}W\right)^{-1}W^{T}\Sigma_{\boldsymbol{p}(\theta_{0})}\right).$$

¿From a practical point of view we have to solve the following system of equations

$$\left\{\begin{array}{l}\dfrac{\partial D_{\phi}\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\theta\right)\right)}{\partial\theta_{i}}=0\\ i=1,...,M_{0}\end{array}\right.,$$

to find the minimum ϕ-divergence estimator $\widehat{\theta}_{\phi}$.

These equations are nonlinear functions of the minimum $\phi-$divergence estimator, $\widehat{\theta}_{\phi}$. In order to solve these equations numerically the Newton-Raphson method is used. We have,

$$\begin{aligned}\left(\frac{\partial D_{\phi}\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\theta^{(t)}\right)\right)}{\partial\theta_{j}}\right)_{\theta=\theta^{(t)}} &= \sum_{l=1}^{M}\left\{\phi\left(\frac{\widehat{p}_{l}}{p_{l}(\theta^{(t)})}\right)-\phi'\left(\frac{\widehat{p}_{l}}{p_{l}(\theta^{(t)})}\right)\frac{\widehat{p}_{l}}{p_{l}(\theta^{(t)})}\right\}\times\\ &\times\left(p_{l}\left(\theta^{(t)}\right)w_{lj}-p_{l}\left(\theta^{(t)}\right)\sum_{u=1}^{M}w_{uj}p_{u}\left(\theta^{(t)}\right)\right),\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial\theta_{r}}\left(\frac{\partial D_{\phi}(\widehat{\boldsymbol{p}},\boldsymbol{p}(\theta^{(t)}))}{\partial\theta_{j}}\right) &= \sum_{l=1}^{M}\phi''\left(\frac{\widehat{p}_{l}}{p_{l}(\theta^{(t)})}\right)\frac{\widehat{p}_{l}}{p_{l}(\theta^{(t)})^{2}}\frac{\partial p_{l}(\theta^{(t)})}{\partial\theta_{r}}\frac{\partial p_{l}(\theta^{(t)})}{\partial\theta_{j}}\frac{\widehat{p}_{l}}{p_{l}(\theta^{(t)})}\\ &+\sum_{l=1}^{k}\frac{\partial^{2}p_{l}(\theta^{(t)})}{\partial\theta_{j}\partial\theta_{r}}\left(\phi\left(\frac{\widehat{p}_{l}}{p_{l}(\theta^{(t)})}\right)-\phi'\left(\frac{\widehat{p}_{l}}{p_{l}(\theta^{(t)})}\right)\frac{\widehat{p}_{l}}{p_{l}(\theta^{(t)})}\right).\end{aligned}\tag{9}$$

Therefore the $(t+1)th$ step estimate, $\widehat{\theta}^{(t+1)}$, in a Newton-Raphson procedure is obtained from $\widehat{\theta}^{(t)}$ as

$$\widehat{\theta}^{(t+1)}=\widehat{\theta}^{(t)}-\left(\frac{\partial D_{\phi}\left(\widehat{p},p\left(\theta^{(t)}\right)\right)}{\partial\theta_{j}}\right)_{\theta=\theta^{(t)}}^{T}G\left(\theta^{(t)}\right)^{-1},$$

where $G\left(\theta^{(t)}\right)$ is the dimension matrix whose elements are defined in (9). A interesting simulation study to analyze behavior of the minimum power-divergence estimator, defined by

$$\widehat{\theta}_{(\lambda)}\equiv\arg\min_{\theta\in\Theta}I^{\lambda}\left(\widehat{\boldsymbol{p}},\boldsymbol{p}\left(\theta\right)\right),\tag{10}$$

in a three dimensional contingence table, has been considered in Pardo, L. and Pardo, M. C. (2003). Notice that $\widehat{\theta}_{(0)}$ is the MLE, as we observed in (6) . Other estimators (less well known than the MLE) that are members of the family of minimum power-divergence estimators are: the minimum chi-squared estimator (Neyman, 1949) for $\lambda=1$; the minimum modified chi-squared estimator (Neyman, 1949) for $\lambda=-2$; the modified

MLE or minimum discrimination information estimator (Kullback, 1985) for $\lambda = -1$; the minimum Matusita distance (or Hellinger distance) estimator (Matusita, 1954) for $\lambda = -1/2$; and the minimum Cressie-Read distance estimator (Cressie and Read, 1984) for $\lambda = 2/3$.

3 Testing hypotheses in loglinear models

For testing if our data can be justified by a loglinear model (goodness-of-fit test), i.e.,

$$H_{Null} : \boldsymbol{p} = \boldsymbol{p}(\theta) \in \mathcal{P} \text{ versus } H_{Alter} : \boldsymbol{p} \in \Delta_M - \mathcal{P} \tag{11}$$

we can use the test statistics

$$T_n^{\phi_1}\left(\widehat{\theta}_{\phi_2}\right) = \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}\left(\widehat{\theta}_{\phi_2}\right)\right).$$

When $T_n^{\phi_1}\left(\widehat{\theta}_{\phi_2}\right) > c$, we should reject H_{Null} in (11), where c is specified so that the size of the test is α :

$$\Pr\left(T_n^{\phi_1}\left(\widehat{\theta}_{\phi_2}\right) > c \mid H_{l+1}\right) = \alpha;\ \alpha \in (0,1).$$

The result estasblished in Morales, D.et al. (1995) for general multinomial models can be adapted to the context of loglinear models. Under (2), (3) and $H_{Null} : \boldsymbol{p} = \boldsymbol{p}(\theta) \in \mathcal{P}$, the test statistic $T_n^{\phi_1}\left(\widehat{\theta}_{\phi_2}\right)$ converges in distribution to a chi-squared distribution with $M - M_0 - 1$ degrees of freedom ($\chi^2_{M-M_0-1}$). Therefore, c could be chosen as the $(1-\alpha)$-th quantile of a $\chi^2_{M-M_0-1}$ distribution:

$$c = \chi^2_{M-M_0-1}(1-\alpha),$$

where $\Pr\left(\chi^2_f \le \chi^2_f(p)\right) = p$.

One of the main problems in loglinear models is to test a nested sequence of hypotheses,

$$H_l : \boldsymbol{p} = \boldsymbol{p}(\theta);\ \theta \in \Theta_l;\ l = 1, ..., m,\ m \le M_0 < M - 1, \tag{12}$$

where $\Theta_m \subset \Theta_{m-1} \subset ... \subset \Theta_1 \subset \mathbb{R}^{M_0}$; $M_0 < M - 1$ and $\dim(\Theta_l) = d_l$; $l = 1, ..., m$, with

$$d_m < d_{m-1} < ... < d_1 \le M_0. \tag{13}$$

Our strategy will be to test successively the hypotheses

$$H_{l+1} \text{ against } H_l; l = 1, ..., m-1, \tag{14}$$

as null and alternative hypotheses respectively. We continue to test as long as the null hypothesis is accepted and choose the loglinear model Θ_l according to the first l for which H_{l+1} is rejected (as a null hypothesis) in favor of H_l (as an alternative hypothesis). This strategy is quite standard for nested models (Read and Cressie, 1988, p. 42). The nesting occurs naturally because of the hierarchical principle, which says that interactions should not be fitted unless the corresponding main effects are present (e.g., Collett, 1994, p.78).

Theorem 1 *Suppose that data $(N_1, ..., N_M)$ are multinomially distributed according to (2) and (3). Consider the nested sequence of hypotheses given by (12) and (13). Choose the two functions $\phi_1, \phi_2 \in \Phi$. Then, for testing hypotheses,*

$$H_0 : H_{l+1} \text{ against } H_1 : H_l,$$

the asymptotic null distribution of the test statistic,

$$T^{(l)}_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(\boldsymbol{p}\left(\widehat{\theta}^{(l+1)}_{\phi_2}\right), \boldsymbol{p}\left(\widehat{\theta}^{(l)}_{\phi_2}\right)\right) \tag{15}$$

is chi-squared with $d_l - d_{l+1}$ degrees of freedom; $l = 1, ..., m-1$. In (15), $\widehat{\theta}^{(l)}_{\phi_2}$ and $\widehat{\theta}^{(l+1)}_{\phi_2}$ are the minimum ϕ_2-divergence estimators under the models H_l and H_{l+1}, respectively, where the minimum ϕ-divergence estimators are defined by (5).

The two most commonly used test statistics in (15) are the Pearson statistic, corresponding to $\phi_1(x) = \frac{1}{2}(x-1)^2$ and $\phi_2(x) = x\log x - x + 1$, and the log-likelihood ratio statistic, corresponding to $\phi_1(x) = \phi_2(x) = x\log x - x + 1\phi_1 \equiv \phi_{(0)}$ (e.g., Christensen, 1997, p. 338). The asymptotic null distribution of both of these statistics is a central chi-squared distribution with $d_l - d_{l+1}$ degrees of freedom. We should reject the null hypothesis if $T^{(l)}_{\phi_1,\phi_2} > c$, where c is specified so that the size of the test is α :

$$\Pr\left(T^{(l)}_{\phi_1,\phi_2} > c \mid H_{l+1}\right) = \alpha; \;\; \alpha \in (0,1), \tag{16}$$

based on Theorem 1 c could be chosen as the $(1-\alpha)$-th quantile of a $\chi^2_{d_l - d_{l+1}}$ distribution:

$$c = \chi^2_{d_l - d_{l+1}}(1-\alpha). \tag{17}$$

The choice of (17) in (16) only guarantees an asymptotic size-α test. In the case of the Pearson and loglikelihood ratio statistics, some corrections to (17) have been proposed, and these have been discussed by Read and Cressie (1988), Ch. 5, in the context of power-divergence statistics for testing goodness-of-fit. In Cressie, N., Pardo, L. and Pardo, M.C. (2003), (17) was used to answer, in a finite-sample simulation study, what choices of λ in the family $T^{(l)}_{\phi_{(\lambda)},\phi_{(0)}}$ is the relation (16) most accurately attained? It was concluded with the interesting result that, for loglinear models, the test statistic based on the power-divergence measure for $\lambda = 2/3$ (Cressie-Read statistic), offers an attractive alternative to the classical Pearson-based ($\lambda = 1$) and likelihood-ratio-based ($\lambda = 0$) test statistics. The same value of $\lambda = 2/3$, was found by Cressie and Read (1984) to be at times preferable to $\lambda = 0, 1$ in problems of goodness-of-fit.

To test the nested sequence of hypotheses $\{H_l : l = 1, ..., m\}$ effectively, we need an asymptotic independence result for the sequence of test statistics $T^{(1)}_{\phi_1,\phi_2}, T^{(2)}_{\phi_1,\phi_2}, ..., T^{(m^*)}_{\phi_1,\phi_2}$, where m^* is the integer $1 \leq m^* \leq m$ for which H_{m^*} is true but H_{m^*+1} is not true . This result is given in the theorem below.

Theorem 2 *Suppose that data $(N_1, ..., N_M)$ are multinomially distributed according to (2) and (3). Suppose we wish to test first,*

$$H_{Null} : H_l \text{ against } H_{Alt} : H_{l-1},$$

followed by

$$H_{Null} : H_{l+1} \text{ against } H_{Alt} : H_l.$$

Then, under the hypothesis H_l, the statistics $T^{(l-1)}_{\phi_1,\phi_2}$ and $T^{(l)}_{\phi_1,\phi_2}$ are asymptotically independent and chi-squared distributed on $d_{l-1} - d_l$ and $d_l - d_{l+1}$ degrees of freedom, respectively.

The proof of this result can be seen in In Cressie, N., Pardo, L. and Pardo, M.C. (2003).

In general, theoretical results for the test statistic $T^{(l)}_{\phi_1,\phi_2}$ under alternative hypotheses are not easy to obtain. An exception to this is when there is a contiguous sequence of alternatives that approach the null hypothesis H_{l+1} at the rate of $O\left(n^{-1/2}\right)$. Regarding the alternative, Haberman (1974) was the first to study the asymptotic distribution of the Pearson statistic and log-likelihood ratio statistic under contiguous alternative hypotheses, establishing that the asymptotic distribution is non-centrally chi-squared distributed with $d_l - d_{l+1}$ degrees of freedom. Oler (1985) presented a systematic study of the contiguous alternative hypotheses in multinomial populations, obtaining as a special case the asymptotic distribution for the log-linear models. Through simulations, she also studied how closely the noncentral chi-squared distributions agree with the exact sampling distributions. Fenech and Westfall (1988) presented an interesting analytic study of the noncentrality parameter in the case of loglinear models. Now we generalize their results to tests based on the $\phi-$divergence statistic $T^{(l)}_{\phi_1,\phi_2}$ given by (15).

Consider the multinomial probability vector

$$\boldsymbol{p}_n(\theta) \equiv \boldsymbol{p}(\theta) + d/\sqrt{n};\ \theta \in \Theta_{l+1}, n \geq n_0 > 0, \tag{18}$$

where $\boldsymbol{d} \equiv (d_1, ..., d_M)^T$ is a fixed $M \times 1$ vector such that $\sum_{j=1}^M d_j = 0$, and recall that n is the total-count parameter of the multinomial distribution. As $n \to \infty$, the sequence of multinomial probabilities $\{\boldsymbol{p}_n(\theta)\}_{n \in N}$ converge to a multinomial probability in H_{l+1} at the rate of $O\left(n^{-1/2}\right)$. We call

$$H_{l+1,n} : \boldsymbol{p} = \boldsymbol{p}_n(\theta) = \boldsymbol{p}(\theta) + d/\sqrt{n};\ \theta \in \Theta_{l+1}, n \geq n_0 > 0, \tag{19}$$

a sequence of *contiguous alternative hypotheses,* here contiguous to the null hypothesis H_{l+1}.

Now consider testing

$$H_{Null} : H_{l+1} \text{ against } H_{Alt} : H_{l+1,n}, \tag{20}$$

using the test statistic $T_{\phi_1,\phi_2}^{(l)}$ given by (15). The power of this test is,

$$\pi_n^{(l)} \equiv \Pr\left(T_{\phi_1,\phi_2}^{(l)} > c\,|H_{l+1,n}\right). \tag{21}$$

In what to follow, we show that under the alternative $H_{l+1,n}$, and as $n \to \infty$, $T_{\phi_1,\phi_2}^{(l)}$ converges in distribution to a non-central chi-squared random variable with non-centrality parameter μ, where μ is given in Theorem 3, and $d_l - d_{l+1}$ degrees of freedom $(\chi^2_{d_l-d_{l+1},\mu})$. Consequently, as $n \to \infty$,

$$\pi_n^{(l)} \to \Pr\left(\chi^2_{d_l-d_{l+1},\mu} > c\right). \tag{22}$$

Theorem 3 *Suppose that data $(N_1, ..., N_k)$ are multinomially distributed according to (2) and (3). The asymptotic distribution of the statistic $T_{\phi_1,\phi_2}^{(l)}$, under the contiguous alternative hypotheses (19), is a chi-squared distribution with $d_l - d_{l+1}$ degrees of freedom and noncentrality parameter μ given by*

$$\mu = \boldsymbol{d}^T diag\left(\boldsymbol{p}(\theta_0)^{-1/2}\right)\left(A_{(l)} - A_{(l+1)}\right) diag\left(\boldsymbol{p}(\theta_0)^{-1/2}\right)\boldsymbol{d},$$

where $\boldsymbol{d} = (d_1, ..., d_M)^T$ is defined in (19) and satisfies $\sum_{i=1}^{M} d_i = 0$, and

$$A_{(i)} = diag(\boldsymbol{p}(\theta_0)^{-1/2})\Sigma_{\boldsymbol{p}(\theta_0)}W_{(i)}\left(W_{(i)}^T\Sigma_{\boldsymbol{p}(\theta_0)}W_{(i)}\right)^{-1}W_{(i)}^T\Sigma_{\boldsymbol{p}(\theta_0)} diag\left(\boldsymbol{p}(\theta_0)^{-1/2}\right); \quad i = l, l+1.$$

Remark 1 *Theorem 3 can be used to obtain an approximation to the power function of the test (14), as follows. Write*

$$\boldsymbol{p}\left(\theta^{(l)}\right) = \boldsymbol{p}\left(\theta^{(l+1)}\right) + \frac{1}{\sqrt{n}}\left(\sqrt{n}\left(\boldsymbol{p}\left(\theta^{(l)}\right) - \boldsymbol{p}\left(\theta^{(l+1)}\right)\right)\right),$$

and define

$$\boldsymbol{p}_n\left(\theta^{(l)}\right) \equiv \boldsymbol{p}\left(\theta^{(l+1)}\right) + \frac{1}{\sqrt{n}}d,$$

where $\boldsymbol{d} = \left(\sqrt{n}\left(\boldsymbol{p}\left(\theta^{(l)}\right) - P\left(\theta^{(l+1)}\right)\right)\right)$. Then substitute $\boldsymbol{p}$ into the definition of μ, and finally μ into the right hand side of (22), we have the approximation to the power function.

.

Remark 2 *If we consider the statistic $T_{\phi_1,\phi_2}^{(l)}$ with $\phi_2(x) = \psi_{(0)}(x) = x\log x - x + 1$ and $\phi_1(x) = \psi_{(1)}(x) = \frac{1}{2}(1-x)^2$, we obtain the classical Pearson statistic for testing loglinear models (e.g., Christensen, 1997, p.338),*

$$X^2 \equiv n\sum_{j=1}^{k}\frac{\left(p_j\left(\widehat{\theta}^{(l)}\right) - p_j\left(\widehat{\theta}^{(l+1)}\right)\right)^2}{p_j\left(\widehat{\theta}^{(l+1)}\right)},$$

where $\widehat{\theta}^{(i)}$ *is the MLE of* θ *in the model* H_i*. The asymptotic distribution of* X^2 *under a sequence of contiguous alternative hypotheses is given in Theorem 3.*

If we consider the statistic $T^{(l)}_{\phi_1,\phi_2}$ *with* $\phi_2(x) = \psi_{(0)}(x) = x\log x - x + 1$ *and* $\phi_1(x) = \psi_{(0)}(x) = x\log x - x + 1$, *we get the classical likelihood ratio statistic for testing loglinear models (e.g., Christensen, 1997, p.338),*

$$G^2 \equiv 2\sum_{j=1}^{k} p_j\left(\widehat{\theta}^{(l+1)}\right)\log\frac{p_j\left(\widehat{\theta}^{(l+1)}\right)}{p_j\left(\widehat{\theta}^{(l)}\right)}.$$

Again, the asymptotic distribution of G^2 *under a sequence of contiguous alternative hypotheses is given in Theorem 2. This particular result was obtained for the first time in Oler (1985).*

4 Nonadditivity in loglinear models

Let

$$\log \boldsymbol{p}(\theta) \in C(X_A), \tag{23}$$

be any loglinear model with $\dim(X_A) = M \times t_A$ and $rank(X_A) = t_A$. This loglinear model can be written as

$$\boldsymbol{p}^{A*}\left(\theta^A\right) \equiv \exp\left(X_A\theta^A\right)/n,$$

where θ^A is a $t_A \times 1$ vector contained in $\mathbb{R}^{t_A}$. Although model (23) seems adequate for our data it is possible to consider a more complete model. For instance Christensen and Utts (1992) expanded this model as follows

$$\boldsymbol{p}^{V*}\left(\theta^V\right) \equiv \exp\left(X_V\theta^V\right)/n, \tag{24}$$

where $X_V = \left(X_A, Z\left(N^{-1}\boldsymbol{m}\left(\theta^A\right)\right)\right)$ and $\theta^V = \left(\left(\theta^A\right)^T, \gamma^T\right)^T$. The matrix function $Z(.)$ is assumed to be differentiable and $rank((X_A, Z(.))) = t_V > t_A$ and the functional form of its elements is known but it is a function of the unknown estimable functions of θ^A. This model includes nonlinear terms from what it is not a loglinear model.

The problem of considering models in which nonlinear terms have been added was considered for the first time by Tukey (1949). He solved the problem of testing if there is interaction in the two-way classification model with one observation per cell. There have been several extensions of this test to models with different functions for interactions and to other designs. These are discussed by Harter and Lum (1962), Mandel (1959, 1961) Tukey (1955, 1962). Latter Milliken and Graybill (1970) extended these ideas to the general linear model and to the case where the interaction is any known function of the block and treatment effects. Johnson and Graybill (1972) considered the situation where the interaction may not be a function of the treatment and block effects. The papers of Snee (1982) and Petitt (1989) are also interesting in this area. The importance as well as some interesting references in relation with the problem of testing nonadditivity, in general, can be seen in these papers and, for loglinear models in particular, in Christensen and Utts (1992).

To obtain the maximum likelihood estimator for the parameters of the lognonlinear model (24) require specialized methods for fitting it. But it is possible to overcome this problem using the two-stage fitting procedure. In this procedure the parameters are first estimated using a loglinear model, and then the estimates are treated as known constants for the second stage of the test. Therefore, the objective of this Section is to test

$$H_{Null}: \gamma = 0 \text{ versus } H_{Alt}: \gamma \neq 0. \tag{25}$$

where γ is an unknown vector.

For testing (24), Christensen and Utts (1992) proposed the likelihood ratio test. In Pardo, L. and Pardo, M.C. (2003) some new families of test statistics for testing (25) were presented. These three new families are natural extensions of the likelihood ratio test and they are based on the $\phi-$divergence measures.

Following Christensen and Utts (1992), rather than fitting (24) directly, the two-stage estimation procedure consists of finding $\boldsymbol{p}\left(\widehat{\theta}^A\right)$ from model (23) and then fitting

$$\boldsymbol{p}^{V*}\left(\theta^V\right) \equiv \exp\left(X_A\theta^A + Z\left(N^{-1}\boldsymbol{m}\left(\widehat{\theta}^A\right)\right)\gamma\right)/N,$$

with $Z\left(N^{-1}\boldsymbol{m}\left(\widehat{\theta}^A\right)\right)$ treated as a known fixed matrix. In the second stage, we want to test

$$H_{Null}: \log \boldsymbol{p}(\theta) \in C\left(X_A\right) \text{ versus } H_{Alt}: \log \boldsymbol{p}(\theta) \in C\left(\widehat{X}_V\right), \tag{26}$$

where $\widehat{X}_V = \left(X_A, Z\left(m^{A_*}\left(\widehat{\theta}^A\right)\right)\right)$. It is clear that $C\left(X_A\right) \subset C\left(\widehat{X}_V\right)$.

For linear models, the validity of tests based on this two-stage fitting procedure was established by Milliken and Graybill (1970) and in loglinear models by Christensen and Utts (1992). These authors proposed the likelihood ratio test for testing (26), which is given by

$$G^2 = -2\log\prod_{j=1}^{k}\left(\frac{m_j^A\left(\widehat{\theta}^A\right)}{m_j^V\left(\widehat{\theta}^V\right)}\right)^{n_{Nj}}. \tag{27}$$

This statistic can be written as

$$G^2 = 2n\sum_{j=1}^{k} p_j^{V_*}\left(\widehat{\theta}^V\right)\log\left(\frac{p_j^{V_*}\left(\widehat{\theta}^V\right)}{p_j^{A_*}\left(\widehat{\theta}^A\right)}\right) \tag{28}$$

which is the usual form. We can observe that the expression of G^2 in (28) can be written as

$$G^2 = 2ND_{Kull}\left(\boldsymbol{p}^{V_*}\left(\widehat{\theta}^V\right), \boldsymbol{p}^{A_*}\left(\widehat{\theta}^A\right)\right). \tag{29}$$

As a generalization of (29) in Pardo, L. and Pardo, M. C. (2003) it was considered the following family of statistics for testing (26),

$$T_\phi \equiv \frac{2N}{\phi''(1)} D_\phi\left(\boldsymbol{p}^{A*}\left(\widehat{\theta}^A\right), \boldsymbol{p}^{V*}\left(\widehat{\theta}^V\right)\right). \tag{30}$$

We can also write (27) as

$$G^2 = 2\sum_{j=1}^{k} n_{Nj} \log\left(\frac{m_j^V\left(\widehat{\theta}^V\right)}{m_j^A\left(\widehat{\theta}^A\right)}\right) \tag{31}$$

so

$$G^2 = 2N\left(D_{Kull}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}^{A_*}\left(\widehat{\theta}^A\right)\right) - D_{Kull}\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}^{V_*}\left(\widehat{\theta}^V\right)\right)\right). \tag{32}$$

Then as a generalization of (32) we can also consider the family of statistics

$$S_\phi \equiv \frac{2N}{\phi''(1)}\left(D_\phi\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}^{A_*}\left(\widehat{\theta}^A\right)\right) - D_\phi\left(\widehat{\boldsymbol{p}}, \boldsymbol{p}^{V_*}\left(\widehat{\theta}^V\right)\right)\right). \tag{33}$$

In the following Theorem we consider the asymptotic distribution of the familes of statistics T_ϕ and S_ϕ.

Theorem 4 *Suppose that data $(N_1, ..., N_k)$ are multinomially distributed according to (2) and (3). For testing hypotheses H_A versus H_V given by (26), with $C(X_A) \subset C\left(\widehat{X}_V\right)$, $C(X_0) \subset C(X_A)$, the asymptotic null distribution of the test statistics T_ϕ and S_ϕ under the null hypothesis given in (26) is chi-squared with $t_V - t_A$ degrees of freedom.*

In the cited paper of Pardo, L. and Pardo, M. C. (2003) it can be seen an interesting simulation study to get the best value of λ in the power divergence family. The results presented in this review assume a multinomial random sampling. In Cressie, N. and Pardo, L. (2002) the results given in Theorems 2 and 3 were extended under the assumption of either Poisson, multinomial or product multinomial sampling. In Pardo, L. and Pardo, M. C. (2003) the Theorem 4 was presented under the same types of random sampling.

References

[1] Ali, S.M. and Silvey, S.D. (1966). A general class of coefficient of divergence of one distribution from another. *Journal of Royal Statistical Society, Series B*, **286**, 131–142.

[2] Christensen, R. and Utts, J. (1992). Testing for nonadditivity in log-linear and logit models. *Journal of Statistical Planning and Inference,* **33**, 333-343.

[3] Christensen, R. (1997). *Log-Linear Model and Logistic Regression.* Springer-Verlag, New York.

[4] Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society*, *Series B*, **46,** 440-464.

[5] Cressie, N. and Pardo, L. (2000). Minimum $\phi-$divergence estimator and hierarchical testing in loglinear models. *Statistica Sinica,* **10,** 867-884 .

[6] Cressie, N. and Pardo, L. (2002).Model checking in loglinear models using ϕ-divergences and MLEs. *Journal of Statistical Planning and Inference,* **103,** 437-453.

[7] Cressie, N. , Pardo, L. Pardo, M. C. (2003). Size and power considerations for testing loglinear models using ϕ-divergence test statistics. *Statistica Sinica,* **13,** 555-570.

[8] Csiszár, I. (1963). Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of Hungarian Academy of Sciences* **8**, *Ser. A*, 85–108.

[9] Fenech, A. P. and Westfall, P. H. (1988). The power function of conditional log-linear model tests. *Journal of the American Statistical Association,* **83,** 198-203.

[10] Haberman, S. J. (1974). *The Analysis of Frequency Data.* University of Chicago Press, Chicago

[11] Harter, H.L. and Lum, M.D. (1962). An interpretation and extension of Tukey's one degree of freedom for nonadditivity. *Aeronautical Research Laboratory Technical Report*, ARL 62-313.

[12] Johnson, D.E. and Graybill, F.A. (1972). An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association*, ***67***, 862-868.

[13] Kullback, S. (1985). Kullback information. In *Encyclopedia of Statistical Sciences, Volume* 4 (editors S. Kotz and N. L. Johnson), 421-425. John Wiley & Sons, New York.

[14] Mandel, J. (1959). The analysis of Latin Squares with a certain type of row-column interaction. *Technometrics*, ***1***, 379-387.

[15] Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association*, ***56***, 878-888.

[16] Matusita, K. (1954). On the estimation by the minimum distance method. *Annals of the Institute of Statistical Mathematics* **5,** 59-65.

[17] Milliken, G. A. and Graybill, F. A. (1970). Extensions of the general linear hypothesis model. *Journal of the American Statistical Association,* **65**, 797-807.

[18] Morales, D., Pardo, L., and Vajda, I. (1995). Asymptotic divergences of estimates of discrete distributions. *Journal of Statistical Planning and Inference,* **48,** 347-369.

[19] Neyman, J. (1949). Contribution to the theory of the χ^2 test. *Proceeding of the First Berkeley Symposium on Mathematical Statistics and Probability,* 239-275.

[20] Oler, J. (1985). Noncentrality parameters in chi-squared goodness-of-fit analyses with an application to log-linear procedures. *Journal of the American Statistical Association,* **80,** 181-189.

[21] Pardo, L. and Pardo, M. C. (2003). Minimum power-divergence estimator in three-way contingency tables. *J. Statist. Comput. Simul. (*In print).

[22] Pardo, L. and Pardo, M. C. (2003). Nonadditivity in loglinear models using ϕ-divergences and MLEs. *Journal of Statistical Planning and Inference (*In print*).*

[23] Pettit, A.N. (1989). One degree of freedom for nonadditivity. Application with Generalized Linear Models and Link Functions. *Biometrics*, **45**, 1153-1162.

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

[25] Searle, S. R. (1971). *Linear Models.* John Wiley & Sons, New York.

[26] Snee, R.D. (1982). Non-additivity in two-way classification. Is it interaction or nonhomogeneous variance? *Journal of the American Statistical Association*, **77**, 515,519.

[27] Tukey J. W. (1949). One degree of freedom for nonadditivity. *Biometrics, **5**,* 232-242.

[28] Tukey J. W. (1955). Answer to Query 113. *Biometrics, **11**,* 113-123.

[29] Tukey J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics, **33**,* 1-62.