Julio A. Pardo; Leandro Pardo; María Del Carmen Pardo; K. Zografos

An exploratory canonical analysis approach for multinomial populations based on the ϕ-divergence measure

Kybernetika, Vol. 40 (2004), No. 6, [757]--776

Persistent URL: http://dml.cz/dmlcz/135632

KYBERNETIKA — VOLUME 40 (2004), NUMBER 6, PAGES 757–776

AN EXPLORATORY CANONICAL ANALYSIS APPROACH FOR MULTINOMIAL POPULATIONS BASED ON THE ϕ–DIVERGENCE MEASURE*

J. A. Pardo, L. Pardo, M. C. Pardo and K. Zografos

In this paper we consider an exploratory canonical analysis approach for multinomial population based on the ϕ-divergence measure. We define the restricted minimum ϕ-divergence estimator, which is seen to be a generalization of the restricted maximum likelihood estimator. This estimator is then used in ϕ-divergence goodness-of-fit statistics which is the basis of two new families of statistics for solving the problem of selecting the number of significant correlations as well as the appropriateness of the model.

Keywords: canonical analysis, restricted minimum ϕ-divergence estimator, minimum ϕ-divergence statistic, simulation, power divergence

AMS Subject Classification: 62H17, 62H20, 62B10

1. INTRODUCTION

Let X and Y denote two categorical response variables with I and J levels respectively. When we classify subjects on both variables, there are IJ possible combinations of classification. The responses (X,Y) of a subject randomly chosen from some population have a probability distribution $p_{ij} = P(X=i, Y=j)$, with $p_{ij} > 0$, $i = 1,\dots,I$; $j = 1,\dots,J$ and we denote by $p = (p_{11},\dots,p_{IJ})^T$ the joint distribution of X and Y. We usually display this distribution in a rectangular table having I rows for the categories of X and J columns for the categories of Y. Let $M = \min(I-1, J-1)$ and we denote by $p_{i.} = \sum_{j=1}^{J} p_{ij}$ and $p_{.j} = \sum_{i=1}^{I} p_{ij}$. Here and in the sequel, "T" denotes the vector or matrix transpose.

Canonical analysis explores the structure of a contingency table. It is based on the fact, see Lancaster [22], that the bivariate probability p_{ij} can always be expanded for each i and j as

$$p_{ij} = p_{i.}p_{.j}\left(1 + \sum_{l=1}^{M} \lambda_l u_{il} v_{jl}\right), \quad 1 \le i \le I,\ 1 \le j \le J, \tag{1.1}$$

*The research in this paper was supported in part by Greek General Secretary of Research and Technology and Spanish Foreign Office, through a bilateral program of scientific and technologic cooperation (2000–2001) and DGI BMF2003-0892.

where

$$\begin{aligned}\sum_{i=1}^{I} u_{il} p_{i.} &= \sum_{j=1}^{J} p_{.j} v_{jl} = 0 \\ \sum_{i=1}^{I} p_{i.} u_{il} u_{il'} &= \sum_{j=1}^{J} p_{.j} v_{jl} v_{jl'} = \delta_{ll'}\end{aligned} \tag{1.2}$$

with $1 \leq l \leq M$, $1 \leq l' \leq M$. Being $\delta_{ll'}$ the Kronecker delta. The decomposition (1.1) is called the canonical form of the bivariate distribution $p = (p_{11}, \dots, p_{IJ})^T$.

In the previous representation, the u_{il} $(1 \leq i \leq I, 1 \leq l \leq M)$ are canonical scores assigned to X such that the canonical variables $U_l = (u_{1l}, \dots, u_{Il})^T$ have means 0 and variances 1, and are uncorrelated. The v_{jl} $(1 \leq j \leq J, 1 \leq l \leq M)$ are canonical scores assigned to Y such that the canonical variables $V_l = (v_{1l}, \dots, v_{Jl})^T$ also have means 0 and variances 1, and are uncorrelated. Thus, λ_l is the canonical correlation of U_l and V_l. More details about canonical analysis from contingency tables can be seen in Anderson [3], Greenacre [18, 20] and references there in.

The purpose of canonical analysis of contingency tables is the determination of the dimensionality of (1.1); that is, the determination of the number of significant correlations. If the scores are ordered so that $|\lambda_1| \geq |\lambda_2| \geq \cdots . \geq |\lambda_M|$ we can obtain the appropriate dimension by the values of the ratios

$$r_m^2 = \frac{\lambda_1^2 + \cdots + \lambda_m^2}{\lambda_1^2 + \cdots + \lambda_M^2} \tag{1.3}$$

for $m = 1, 2, \dots, M$. The ratio (1.3) is interpreted as the amount of the variation accounted for the first m dimensions. The choice of dimension $m = M_0$ is considered satisfactory if (1.3) is close to 1 and largely unchanged if M_0 is further increased. However, the choice of M_0 and the appropriateness of the model can also be based on statistical inference principles. Gilula and Haberman [17] presented a development of canonical analysis that for the first time exploits general results concerning restricted maximum likelihood estimators. That approach permits use of confidence intervals and estimated asymptotic standard deviations as well as to study the likelihood ratio test and chi-squared test based on restricted maximum likelihood estimators to select M_0 and the appropriateness of the model.

The kind of cross-tabulated frequency data, considered by the canonical analysis, often arise in Biometry. For example, in ecology, the species appear in different communities and the relation between them can be found by canonical analysis. Some interesting applications of the canonical analysis can be seen in Fasham [16] and Dahdouh et al [15]. A list of important references in the field of ecology can be seen in Greenacre [18, Section 9.12, p. 318]. In genetics many features can be related (eyes color and hair color, genes and populations) and its canonical analysis representation is also a useful tool, especially in studying polymorphism in population genetics. In medicine one can relate the effectiveness of several drugs, where each drug is rated on a verbal scale (poor, fair, good, very good) for a group of hospital patients. In Greenacre [19] various applications of canonical analysis to biomedical data are presented: On the relationship between headache types and age; On the association between personality types and various medical diagnostic groups; On the categorical rating scales such as an efficacy scale for a medication or a scale pain; On a collection of bacterial isolates with the object of comparing bacterial types and understanding the inter-relationships of the different tests. Another interesting application in this field is given in Greenacre [21]. In this paper the canonical analysis is used to explore relationship between variables in a complex survey and to suggest models for these relationship. The paper of N. J. Crichton and J. P. Hinde [13] is also an interesting paper in this area. Notice that the domain of application of the canonical analysis goes far beyond that of Biometry. In Greenacre [18] there can be seen many applications published in canonical analysis classified by field of application.

In this paper we present a generalization of the results given by Gilula and Haberman [17], in estimation and testing, using the family of ϕ-divergences. In Section 2 we present the restricted minimum ϕ-divergence estimator as a generalization of the restricted maximum likelihood estimator studied by the cited authors. Our generalization is in the sense that the minimum ϕ-divergence estimator with $\phi(x) = x\log x - (x-1)$ gives the maximum likelihood estimator. In Section 3 we introduce two new families of statistics based on the ϕ-divergence for testing the dimensionality of (1.1). One of them (see Theorem 3.1) contains the classical likelihood ratio statistic (for $\phi(x) = x\log x - (x-1)$) as well as the chi-squared statistic (for $\phi(x) = \frac{1}{2}(x-1)^2$), studied in Gilula and Haberman [17], as particular cases. Finally, in Section 4, we present two examples to demonstrate how the results of Sections 2 and 3 can be applied in practice.

2. THE RESTRICTED MINIMUM ϕ-DIVERGENCE ESTIMATOR OF CANONICAL PARAMETERS

We can observe that the probability distribution $p = (p_{ij})$, $(i,j) \in I \times J$, with p_{ij} given in (1.1) could be written as a function of $a = (I+J+1)(M+1) - 1$ parameters. That is to say,

$$p_{ij}(\beta) = \beta_i \beta_{j+I} \left(1 + \sum_{l=1}^{M} \beta_{l+I+J} \beta_{l+I+J+iM} \beta_{l+I+J+(I+j)M} \right)$$

with

$$\beta = \left(\beta_1, \ldots, \beta_{(I+J+1)(M+1)-1}\right) \in B \subset R^a$$

and the β_r are defined in the following way: $\beta_r = p_{r.}$ if $1 \le r \le I$; $\beta_r = p_{.(r-I)}$ if $I+1 \le r \le I+J$; $\beta_r = \lambda_{r-(I+J)}$ if $I+J+1 \le r \le I+J+M$; $\beta_r = u_{1(r-(I+J+M))}$ if $I+J+M+1 \le r \le I+J+2M$; $\ldots$; $\beta_r = u_{I(r-(I+J+IM))}$ if $I+J+IM+1 \le r \le I+J+(I+1)M$; $\beta_r = v_{1,(r-(I+J+(I+1)M))}$ if $I+J+(I+1)M+1 \le r \le I+J+(I+1)M+M$; $\ldots$; $\beta_r = v_{J(r-(I+J)(M+1))}$ if $(I+J)(M+1)+1 \le r \le a$.

The number of constraints in (1.2) is $b = (M+1)(M+2)$ and they are given by

$$\begin{aligned}
\sum_{i=1}^{I} p_{i.} - 1 &= 0 \\
\sum_{j=1}^{J} p_{.j} - 1 &= 0 \\
\sum_{i=1}^{I} p_{i.} u_{il} &= 0 \qquad 1 \le l \le M \\
\sum_{j=1}^{J} p_{.j} v_{jl} &= 0 \qquad 1 \le l \le M \\
\sum_{i=1}^{I} p_{i.} u_{il} u_{il'} - \delta_{ll'} &= 0 \qquad 1 \le l \le l' \le M
\end{aligned}$$

and

$$\sum_{j=1}^{J} p_{.j} v_{jl} v_{jl'} - \delta_{ll'} = 0 \qquad 1 \le l \le l' \le M.$$

Therefore, we can write them in the following way

$$f_s(\beta) = 0, \quad s = 1, \ldots, b$$

and it is not difficult to establish that the matrix

$$L(\beta_0) = \left(\frac{\partial f_s(\beta_0)}{\partial \beta_k} \right)_{\substack{s=1,\ldots,b \\ k=1,\ldots,a}},$$

where β_0 is the true value of the parameter, has the full rank, i. e., b.

First of all, we describe the maximum likelihood procedure, introduced in this context by Gilula and Haberman [17], to estimate the parameter $\beta \in B$ restricted to $f_s(\beta) = 0$, $s = 1, \ldots, b$. Secondly, from this procedure we introduce the restricted minimum ϕ-divergence estimator.

Consider a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with realizations from

$$\mathcal{X} = \{(i,j), \quad i = 1, \ldots, I, \ j = 1, \ldots, J\}$$

independent and identically distributed according to a probability distribution $p(\beta_0) = (p_{11}(\beta_0), \ldots, p_{IJ}(\beta_0))^T$. This distribution is assumed to be unknown, but belonging to a known family

$$\mathcal{P} = \left\{ p(\beta) = (p_{11}(\beta), \ldots, p_{IJ}(\beta))^T : \beta \in B \right\}$$

of distributions on $\mathcal{X}$ with $B \subset R^a$. In other words, the true value β_0 of the parameter $\beta = (\beta_1, \ldots, \beta_a)^T \in B$ is assumed to be unknown. We denote $\widehat{p} = (\widehat{p}_{11}, \ldots, \widehat{p}_{IJ})^T$ with

$$\widehat{p}_{ij} = \frac{N_{ij}}{n} \quad \text{and} \quad N_{ij} = \sum_{k=1}^{n} I_{\{(i,j)\}}((X_k, Y_k)), \quad i = 1, \ldots, I, \ j = 1, \ldots, J.$$

The statistic $(N_{11}, \dots, N_{IJ})$ is obviously sufficient for the statistical model under consideration and it is multinomially distributed; that is

$$P(N_{11} = n_{11}, \dots, N_{IJ} = n_{IJ}) = \frac{n!}{n_{11}! \dots n_{IJ}!} p_{11}(\beta_0)^{n_{11}} \dots p_{IJ}(\beta_0)^{n_{IJ}}$$

for integers $n_{11}, \dots, n_{IJ} \geq 0$ such that $n_{11} + \cdots + n_{IJ} = n$.

If

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \widehat{p}_{ij} \log p_{ij}(\beta)$$

is almost surely (a. s.) maximized over B, under the constraints

$$f_s(\beta) = 0, \quad s = 1, \dots, b$$

at some $\widehat{\beta}^{(r)}$, then $\widehat{\beta}^{(r)}$ is the restricted maximum likelihood estimator (RMLE). For more details see Gilula and Haberman [17].

However, the RMLE can equivalently be defined by the condition

$$\widehat{\beta}^{(r)} = \arg \min_{\beta \in B^*} D^{Kullback}(\widehat{p}, p(\beta)) \quad \text{a. s.}$$

where

$$B^* = \{\beta \in B \subset R^a : f_s(\beta) = 0, s = 1, \dots, b\} \tag{2.1}$$

and

$$D^{Kullback}(p, q) = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

is the Kullback–Leibler divergence between the probability distributions $p = (p_{11}, \dots, p_{IJ})^T$ and $q = (q_{11}, \dots, q_{IJ})^T$. This divergence measure is a particular case of the ϕ-divergence introduced independently by Csiszár [14] and Ali and Silvey [2]. The ϕ-divergence between two probability distributions $p = (p_{11}, \dots, p_{IJ})^T$, $q = (q_{11}, \dots, q_{IJ})^T$ is defined as follows:

$$D_\phi(p, q) = \sum_{i=1}^{I} \sum_{j=1}^{J} q_{ij} \phi\left(\frac{p_{ij}}{q_{ij}}\right), \quad \phi \in \Phi^*,$$

where Φ^* is the class of all convex functions $\phi : [0, \infty) \to R \cup \{\infty\}$, such that at $x = 1$, $\phi(1) = 0, \phi''(1) > 0$, and at $x = 0$,

$$0\phi\left(\frac{0}{0}\right) \equiv 0 \quad \text{and} \quad 0\phi\left(\frac{p}{0}\right) \equiv \lim_{u \to \infty} \frac{\phi(u)}{u}.$$

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

$$D_\psi(p,q) = D_\phi(p,q),$$

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 this paper, as a generalization of the RMLE, $\widehat{\beta}^{(r)}$, we consider the restricted minimum ϕ-divergence estimator,

$$\widehat{\beta}_\phi^{(r)} = \arg\min_{\beta \in B^*} D_\phi(\widehat{p}, p(\beta)) \quad \text{a.s.},$$

where B^* is defined in (2.1). We can observe that the RMLE is a particular case of the restricted minimum ϕ-divergence estimator because for $\phi(x) = x\log x - (x-1)$ we obtain the RMLE.

The restricted minimum ϕ-divergence estimator, $\widehat{\beta}_\phi^{(r)}$, can be obtained as the solution of the following equation system

$$\begin{cases} \dfrac{\partial D_\phi(\widehat{p}, p(\beta))}{\partial \beta_k} + \displaystyle\sum_{s=1}^{b} l_s \dfrac{\partial f_s(\beta)}{\partial \beta_k} = 0, & k = 1, \ldots, a \\ \\ f_s(\beta) = 0, & s = 1, \ldots, b. \end{cases}$$

The unrestricted minimum ϕ-divergence estimator was studied for the first time in Morales et al [27] and the restricted minimum ϕ-divergence estimator was introduced and its properties was studied by Pardo et al [29].

Minimum distance estimation was presented by Wolfowitz [32] and it provides a convenient method of consistently estimating unknown parameters. An extensive bibliography for minimum distance estimates can be found in Parr [28], some additions in Read and Cressie [30] and Morales et al [27], Lindsay [24], Basu and Lindsay [5], Basu and Basu [4] and references there in. Wolfowitz was motivated by the desire to provide consistent parameter estimators in cases where other methods had not proved successful. Other desirable features of minimum distance estimators are natural robustness properties, a concrete interpretation for the value to which the estimator converges even when the model is wrong, ease of application to problems not involving symmetries or invariance properties, extremely competitive small-sample behavior in the several situations thus far explored by the Monte Carlo method (cf. Parr [28]). In the case where the model is discrete, or where the initial information about the data and hypothetical parametrized model is reduced by partitioning the observation space the minimum ϕ-divergence estimators are first order efficient under the model. Several of them have considerable robustness property under moderate contaminations. For more details see Lindsay [24] and Basu and Sarkar [6, 7].

Maximum likelihood estimation subject to constraints is considered for the first time by Aitchison and Silvey [1] in general populations. Matthews and Crowther [25, 26] present a procedure for the exponential family and Pardo et al [29] in multinomial models. For more details see the cited papers and references there in.

In the following theorems we establish some asymptotic properties of the restricted minimum ϕ-divergence estimator of canonical parameters.

Theorem 2.1. If we assume that the canonical correlations satisfy

$$|\lambda_1| > \cdots > |\lambda_K| > 0, \quad \lambda_m = 0, \; m > K, \tag{2.2}$$

then the restricted minimum ϕ-divergence estimator, $\widehat{\beta}_\phi^{(r)}$, satisfies

$$\widehat{\beta}_\phi^{(r)} = \beta_0 + H(\beta_0)\, I_F(\beta_0)^{-1} A(\beta_0)^T \operatorname{diag}\left(p(\beta_0)^{-1/2}\right)(\widehat{p} - p(\beta_0)) + o_p(\|\widehat{p} - p(\beta_0)\|) \tag{2.3}$$

where $\widehat{\beta}_\phi^{(r)}$ is unique in a neighbourhood of the true value of the parameter β_0;

$$A(\beta_0) = \operatorname{diag}\left(p(\beta_0)^{-1/2}\right)\left(\frac{\partial p_{ij}(\beta_0)}{\partial \beta_k}\right)_{\substack{(i,j)\in I\times J\\ k=1,\ldots,a}}$$

and

$$H(\beta_0) = I - I_F(\beta_0)^{-1} L(\beta_0)^T \left(L(\beta_0)\, I_F(\beta_0)^{-1} L(\beta_0)^T\right)^{-1} L(\beta_0)$$

where $I_F(\beta_0) = A(\beta_0)^T A(\beta_0)$ is the Fisher information matrix associated to the multinomial model.

Proof. In Appendix B of Gilula and Haberman [17] it is proved that the conditions given by Birch [9] about the model are satisfied and that the matrix $L(\beta_0)$ has the full rank if condition (2.2) is true. Then the proof is straightforward from Theorem 2.1 in Pardo et al [29]. □

In the following theorem we present the asymptotic distribution of the restricted minimum ϕ-divergence estimator.

Theorem 2.2. Under the assumptions of Theorem 2.1, it holds,

a)

$$\sqrt{n}\left(\widehat{\beta}_\phi^{(r)} - \beta_0\right) \xrightarrow[n\to\infty]{L} N\left(0, H(\beta_0)\, I_F(\beta_0)^{-1}\right)$$

b)

$$\sqrt{n}\left(p\left(\widehat{\beta}_\phi^{(r)}\right) - p(\beta_0)\right) \xrightarrow[n\to\infty]{L} N(0, \Sigma_1)$$

where

$$\Sigma_1 = \operatorname{diag}\left(p(\beta_0)^{1/2}\right) A(\beta_0)\, H(\beta_0)\, I_F(\beta_0)^{-1} A(\beta_0)^T \operatorname{diag}\left(p(\beta_0)^{1/2}\right).$$

Proof. The proof is straightforward from Theorem 2.2 in Pardo et al [29]. □

3. MINIMUM ϕ-DIVERGENCE STATISTIC IN CANONICAL ANALYSIS

The Pearson chi-squared statistic given by

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - np_{ij}\left(\widehat{\beta}^{(r)}\right)\right)^2}{np_{ij}\left(\widehat{\beta}^{(r)}\right)} \tag{3.1}$$

and the likelihood ratio statistic given by

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \log\left(\frac{\widehat{p}_{ij}}{p_{ij}\left(\widehat{\beta}^{(r)}\right)}\right) \tag{3.2}$$

are asymptotically distributed as a chi-squared distribution with $IJ - a + b - 1$ degrees of freedom under the hypothesis

$$H_0 : p = p(\beta)$$

and assuming that

$$|\lambda_1| > \cdots > |\lambda_K| > 0, \quad \lambda_m = 0,\ m > K.$$

It will be better to use the notation β^m instead of β to indicate that β^m is the parameter vector when it is considered in the adding of (1.1) only the first m terms. Under these assumptions, the procedure described in Gilula and Haberman [17] to choose the dimensionality of (1.1), M_0 is to test

$$H_0 : p = p\left(\beta^1\right),$$

that is to say $m = 1$. If we reject the null hypotheses we test

$$H_0 : p = p\left(\beta^2\right),$$

that is to say $m = 2$, until we find the value $M_0 \leq K$ such that H_0 will be not rejected. These tests were carried out using the statistics X^2 and G^2 described above.

Now we present a new family of statistics based on ϕ_1-divergence measures to test

$$H_0 : p = p\left(\beta^m\right), \tag{3.3}$$

which is defined by

$$T^m_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(\widehat{p}, p\left(\widehat{\beta}^{(r)}_{\phi_2,m}\right)\right) \tag{3.4}$$

where $\widehat{\beta}^{(r)}_{\phi_2,m}$ is the restricted minimum ϕ_2-divergence estimator considering m canonical correlations. First of all we obtain its asymptotic distribution.

Theorem 3.1. Under the null hypothesis (3.3) and the assumptions in Theorem 2.1, we have

$$T^m_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(\widehat{p}, p\left(\widehat{\beta}^{(r)}_{\phi_2,m}\right)\right) \xrightarrow[n\to\infty]{L} \chi^2_{IJ-a(m)+b(m)-1}$$

where by $a(m)$ and $b(m)$ we denote that the number of parameters as well as the number of restrictions depend on m.

Proof. It can be established, for more details see Theorem 3.1 in Pardo et al [29], that

$$\frac{2}{\phi_1''(1)} D_{\phi_1}\left(\widehat{p}, p\left(\widehat{\beta}^{(r)}_{\phi_2,m}\right)\right) = \frac{Z^TZ}{n} + o_p\left(n^{-1}\right)$$

where Z^TZ is asymptotically a chi-squared distribution with $IJ - a(m) + b(m) - 1$ degrees of freedom. □

Remark 1. If we consider $\phi_2(x) = x\log x - (x-1)$ and $\phi_1(x) = \frac{1}{2}(x-1)^2$ we get the Pearson chi-squared statistic, X^2, given in (3.1) and for $\phi_1(x) = \phi_2(x) = x\log x - (x-1)$ we get the likelihood ratio statistic, G^2, given in (3.2) (see, i.e., Anderson [3] and references there in).

Remark 2. There are important measures of divergence that can not be expressed as ϕ-divergences, for instance, the divergence measures given by Battacharya, Rényi, and Sharma and Mittal. However, such measures can be written in the following form:

$$D_{\phi,h}(p,q) = h\left(D_\phi(p,q)\right),$$

where h is a differentiable increasing function mapping from $[0,\infty)$ onto $[0,\infty)$, with $h(0) = 0$ and $h'(0) > 0$, and $\phi \in \Phi^*$. In the following table, we present these divergence measures:

Divergence	$h(x)$	$\phi(x)$
Rényi	$\frac{1}{r(r-1)}\log\left(r(r-1)x+1\right); r \neq 0,1$	$\frac{x^r - r(x-1)-1}{r(r-1)}; r \neq 0,1$
Sharma-Mittal	$\frac{1}{s-1}\left\{\left(1+r(r-1)x\right)^{\frac{s-1}{r-1}} - 1\right\}; s,r \neq 1$	$\frac{x^r - r(x-1)-1}{r(r-1)}; r \neq 0,1$
Battacharya	$-\log(-x+1)$	$-x^{1/2} + \frac{1}{2}(x+1)$

In the case of Rényi's divergence, we have

$$D_r(p,q) = \frac{1}{r(r-1)}\log\left(\sum_{j=1}^{k} p_j^r q_j^{1-r}\right); \quad r \neq 0,1,$$

and limiting cases for $r = 0$ and $r = 1$. That is,

$$D_1(P,Q) = \lim_{r\to 1} D_r(P,Q) = \sum_{j=1}^{k} p_j \log\frac{p_j}{q_j} = D^{Kullback}(p,q),$$

which is the Kullback–Leibler divergence.

Similarly,

$$D_0(P,Q)=\sum_{j=1}^{k} q_j \log\frac{q_j}{p_j} = D^{Kullback}(q,p).$$

Theorem 3.2. Under the assumptions given in Theorem 3.1, the asymptotic null distribution of the statistic,

$$T^{m}_{\phi_1,\phi_2,h_1,h_2} \equiv \frac{2n}{\phi_1''(1)\,h_1'(0)} h_1\left(D_{\phi_1}\left(\widehat{p}, p\left(\widehat{\beta}^{(r)}_{\phi_2,h_2,m}\right)\right)\right)$$

is a chi-squared distribution with $IJ-a(m)+b(m)-1$ degrees of freedom, where $\widehat{\beta}^{(r)}_{\phi_2,h_2,m}$ is the minimum (ϕ_2,h_2)-divergence estimator defined by

$$\widehat{\beta}^{(r)}_{\phi_2,h_2,m} = \arg\min_{\beta\in B^*} h_2\left(D_{\phi_2}\left(\widehat{p}, p(\beta)\right)\right) \quad \text{a.s.}$$

Proof. Using a similar approach to that given in the proof of Theorem 3.1, it can be established that

$$\frac{2}{\phi_1''(1)} D_{\phi_1}\left(\widehat{p}, p\left(\widehat{\beta}^{(r)}_{\phi_2,h_2,m}\right)\right) = \frac{Z^T Z}{n} + o_p\left(n^{-1}\right)$$

where the asymptotic distribution of Z^TZ is a chi-squared distribution with $IJ - a(m) + b(m) - 1$ degrees of freedom.

Further, because $h_1(x) = h_1(0) + h_1'(0)\,x + o(x)$, we have,

$$T^{m}_{\phi_1,\phi_2,h_1,h_2} = Z^T Z + o_p(1). \qquad \square$$

As an alternative to this procedure to choose the value M_0 we propose another one as follows. We consider the nested sequence of hypotheses,

$$H_m : \beta \in B^m \subset R^{a(m)}, \quad m = 1,\dots,K \tag{3.5}$$

where

$$B^1 \subset B^2 \subset \cdots \subset B^K \subset R^{a(K)}$$

and $\dim(B^m) = d_m = (I+J+1)(m+1)-1$, $m = 1,\dots,K$, with

$$d_1 < d_2 < \cdots < d_K.$$

Our strategy will be to test successively the hypothesis H_m against H_{m+1}; $m = 1,\dots,K-1$, as null and alternative hypotheses respectively. We go on testing as long as the null hypothesis is rejected and choose the model with $m = M_0$ canonical correlations for the first m for which H_m is accepted. This strategy is quite standard for nested models (Read and Cressie [30, p. 42]). An interesting application in loglinear models can be seen in Cressie and Pardo [10, 11]. To solve this problem we will consider the family of statistics

$$T^{(m)}_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(p\left(\widehat{\beta}^{(r)}_{\phi_2,m+1}\right), p\left(\widehat{\beta}^{(r)}_{\phi_2,m}\right)\right). \tag{3.6}$$

Its asymptotic distribution is given in the following theorem.

Theorem 3.3. Under the null hypothesis (3.5) and the assumptions of Theorem 2.1, we have

$$T^{(m)}_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(p\left(\widehat{\beta}^{(r)}_{\phi_2,m+1}\right), p\left(\widehat{\beta}^{(r)}_{\phi_2,m}\right)\right) \xrightarrow[n\to\infty]{L} \chi^2_{I+J+1}$$

for $m = 1, \dots, K-1$.

Proof. The second order expansion of

$$D_{\phi_1}\left(p\left(\widehat{\beta}^{(r)}_{\phi_2,m+1}\right), p\left(\widehat{\beta}^{(r)}_{\phi_2,m}\right)\right)$$

around $(p(\beta_0), p(\beta_0))$, gives

$$T^{(m)}_{\phi_1,\phi_2} = Z^T Z + o_p(1),$$

where

$$Z = \sqrt{n}\operatorname{diag}\left(p(\beta_0)^{-1/2}\right)\left(p\left(\widehat{\beta}^{(r)}_{\phi_2,m+1}\right) - p\left(\widehat{\beta}^{(r)}_{\phi_2,m}\right)\right),$$

and the first order expansion of $p\left(\widehat{\beta}^{(r)}_{\phi_2,i}\right)$ around $p(\beta_0)$, gives

$$p\left(\widehat{\beta}^{(r)}_{\phi_2,i}\right) - p(\beta_0) = \left(\frac{\partial p(\beta)}{\partial \beta}\right)_{\beta=\beta_0}\left(\widehat{\beta}^{(r)}_{\phi_2,i} - \beta_0\right) + o_p\left(\left\|\widehat{\beta}^{(r)}_{\phi_2,i} - \beta_0\right\|\right) \quad i = m, m+1.$$

By Theorem 2.1

$$\begin{aligned}\widehat{\beta}^{(r)}_{\phi_2,i} - \beta_0 &= H^{(i)}(\beta_0)\, I_F^{(i)}(\beta_0)^{-1}\left(A^{(i)}(\beta_0)\right)^T \operatorname{diag}\left(p(\beta_0)^{-1/2}\right)(\widehat{p} - p(\beta_0)) \\ &\quad + o_p(\|\widehat{p} - p(\beta_0)\|)\end{aligned}$$

where

$$A^{(i)}(\beta_0) = A^{(i)} = \operatorname{diag}\left(p(\beta_0)^{-1/2}\right)\left(\frac{\partial p(\beta)}{\partial \beta}\right)_{\beta=\beta_0}$$

$$I_F^{(i)}(\beta_0) = I_F^{(i)} = \left(A^{(i)}\right)^T A^{(i)}$$

$$H^{(i)}(\beta_0) = H^{(i)} = I - I_F^{(i)}(\beta_0)^{-1}\left(L^{(i)}\right)^T\left(L^{(i)} I_F^{(i)}(\beta_0)^{-1}\left(L^{(i)}\right)^T\right)^{-1} L^{(i)},$$

and

$$L^{(i)} = L^{(i)}(\beta_0) = \left(\frac{\partial f(\beta)}{\partial \beta}\right)_{\beta=\beta_0} \quad i = m, m+1.$$

Then

$$p\left(\widehat{\beta}^{(r)}_{\phi_2,m+1}\right)-p\left(\widehat{\beta}^{(r)}_{\phi_2,m}\right)=\left(\left(\frac{\partial p(\beta)}{\partial \beta}\right)_{\beta=\beta_0}H^{(m+1)}\left(I_F^{(m+1)}\right)^{-1}\left(A^{(m+1)}\right)^T\right.$$
$$\left.-\left(\frac{\partial p(\beta)}{\partial \beta}\right)_{\beta=\beta_0}H^{(m)}\left(I_F^{(m)}\right)^{-1}\left(A^{(m)}\right)^T\right)\operatorname{diag}\left(p(\beta_0)^{-1/2}\right)$$
$$\times\left(\widehat{p}-p(\beta_0)\right)+o_p\left(\left\|\widehat{\beta}^{(r)}_{\phi_2,m+1}-\beta_0\right\|\right)-o_p\left(\left\|\widehat{\beta}^{(r)}_{\phi_2,m}-\beta_0\right\|\right)$$

or equivalently

$$\begin{aligned}T &= \left(Q_{(m+1)}-Q_{(m)}\right)\operatorname{diag}\left(p(\beta_0)^{-1/2}\right)\left(\widehat{p}-p(\beta_0)\right)\\ &\quad +o_p\left(\left\|\widehat{\beta}^{(r)}_{\phi_2,m+1}-\beta_0\right\|\right)-o_p\left(\left\|\widehat{\beta}^{(r)}_{\phi_2,m}-\beta_0\right\|\right)\end{aligned}$$

where

$$T=\operatorname{diag}\left(p(\beta_0)^{-1/2}\right)\left(p\left(\widehat{\beta}^{(r)}_{\phi_2,m+1}\right)-p\left(\widehat{\beta}^{(r)}_{\phi_2,m}\right)\right)$$

and

$$Q_{(i)}=A^{(i)}H^{(i)}\left(I_F^{(i)}\right)^{-1}\left(A^{(i)}\right)^T\qquad i=m,m+1.$$

Therefore

$$Z\xrightarrow[n\to\infty]{L}N\left(0,\Sigma^*\right)$$

with

$$\begin{aligned}\Sigma^* &= \left(Q_{(m+1)}-Q_{(m)}\right)\left(I-\sqrt{p(\beta_0)}\sqrt{p(\beta_0)^T}\right)\left(Q_{(m+1)}-Q_{(m)}\right)\\ &= \left(Q_{(m+1)}-Q_{(m)}\right)\left(Q_{(m+1)}-Q_{(m)}\right)\end{aligned}$$

since

$$\sqrt{p(\beta_0)^T}A^{(i)}=0,\qquad i=m,m+1.$$

On being, $Q_{(m)}Q_{(m+1)}=Q_{(m+1)}Q_{(m)}=Q_{(m)}$, $Q^2_{(m)}=Q_{(m)}$ and $Q^2_{(m+1)}=Q_{(m+1)}$ we have that $\Sigma^*=Q_{(m+1)}-Q_{(m)}$ is symmetric and idempotent so all the eigenvalues of Σ^* are zero except for $d_{m+1}-d_m=I+J+1$ unit values and

$$Z^TZ\xrightarrow[n\to\infty]{L}\chi^2_{I+J+1}.\qquad\square$$

Theorem 3.4. Under the assumptions given in Theorem 3.3, the asymptotic null distribution of the statistic,

$$T^{(m)}_{\phi_1,\phi_2,h_1,h_2}\equiv\frac{2n}{\phi_1''(1)\,h_1'(0)}h_1\left(D_{\phi_1}\left(p\left(\widehat{\beta}^{(r)}_{\phi_2,h_2,m+1}\right),p\left(\widehat{\beta}^{(r)}_{\phi_2,h_2,m}\right)\right)\right)$$

is a chi-squared distribution with $I+J+1$ degrees of freedom.

Proof. Using a similar approach to that given in the proof of Theorem 3.3, it can be established that

$$\frac{2}{\phi_1''(1)} D_{\phi_1}\left(p\left(\widehat{\beta}^{(r)}_{\phi_2,h_2,m+1}\right), p\left(\widehat{\beta}^{(r)}_{\phi_2,h_2,m}\right)\right) = \frac{Z^T Z}{n} + o_p\left(n^{-1}\right).$$

Now a first-order expansion of $h(x)$, in a similar way to the Theorem 3.2 gives

$$T^{(m)}_{\phi_1,\phi_2,h_1,h_2} = Z^T Z + o_p(1)$$

where $Z^T Z$ is asymptotically a chi-squared distribution with $I+J+1$ degrees of freedom. □

Theorem 3.5. Under the assumptions given in Theorem 3.3 and 3.4, the asymptotic null distribution of each of the statistics,

$$\widetilde{T}^{(m)}_{\phi_1,\phi_2} \equiv \frac{2n}{\phi_1''(1)} D_{\phi_1}\left(p\left(\widehat{\beta}^{(r)}_{\phi_2,m}\right), p\left(\widehat{\beta}^{(r)}_{\phi_2,m+1}\right)\right)$$

and

$$\widetilde{T}^{(m)}_{\phi_1,\phi_2,h_1,h_2} \equiv \frac{2n}{\phi_1''(1)\, h_1'(0)} h_1\left(D_{\phi_1}\left(p\left(\widehat{\beta}^{(r)}_{\phi_2,h_2,m}\right), p\left(\widehat{\beta}^{(r)}_{\phi_2,h_2,m+1}\right)\right)\right)$$

is asymptotically a chi-squared distribution with $I+J+1$ degrees of freedom.

Proof. We consider the function $\varphi(x) = x\phi_1\left(x^{-1}\right)$. It is clear that $\varphi(x) \in \Phi$, $T^{(m)}_{\varphi,\phi_2} = \widetilde{T}^{(m)}_{\phi_1,\phi_2}$, $T^{(m)}_{\varphi,\phi_2,h_1,h_2} = \widetilde{T}^{(m)}_{\phi_1,\phi_2,h_1,h_2}$. Then the results follow directly from Theorems 3.3 and 3.4. □

4. NUMERICAL APPLICATION

To illustrate results, two examples previously analyzed by several authors will be considered in this section. We consider the power-divergence measure introduced by Cressie and Read [12], which is a particular case of the Csiszár divergence, for estimation as well as testing. That is to say, we consider the statistics $T^{(m)}_{\phi_{(a_1)},\phi_{(a_2)}}$ with

$$\phi_{(a)}(x) \equiv (a(a+1))^{-1}\left(x^{a+1} - x\right) - (x-1)(a+1)^{-1} \quad a \neq 0, a \neq -1$$

$$\phi_{(0)}(x) = \lim_{a\to 0} \phi_{(a)}(x), \qquad \phi_{(-1)}(x) = \lim_{a\to -1} \phi_{(a)}(x). \tag{4.1}$$

For more details about this family of divergences as well as its importance in statistical inference see Read and Cressie [30].

We estimate the parameters β by the RMLE $\widehat{\beta}^{(r)}_{0,m}$; the restricted minimum chi-square, $\widehat{\beta}^{(r)}_{1,m}$ and the new estimator obtained for $a_2 = 2/3$, $\widehat{\beta}^{(r)}_{2/3,m}$ with $\widehat{\beta}^{(r)}_{a_2,m} \equiv \widehat{\beta}^{(r)}_{\phi_{(a_2)},m}$. They have been obtained by means of a FORTRAN-90 program which uses the nag_nlp module of the NAG F-90 Numerical Libraries. As initial point for the minimization algorithm have been used the empirical row and column marginal probabilities $\widehat{p}_{i.} = \sum_{j=1}^{J} \widehat{p}_{ij}, i = 1, \dots, I$ and $\widehat{p}_{.j} = \sum_{i=1}^{I} \widehat{p}_{ij}, j = 1, \dots, J$ as estimators of $p_{i.}$, $i = 1, \dots, I$ and $p_{.j}, j = 1, \dots, J$ respectively as well as the estimators for the canonical correlations and scores obtained by ordinary canonical analysis which is described in the following.

Let further Λ be a diagonal matrix of dimension M with $\widehat{\lambda}_1, \dots, \widehat{\lambda}_M$ in the diagonal and U and V be matrices of dimension $I \times M$ and $J \times M$, respectively, which have the $\widehat{u}'_{im}s$ and $\widehat{v}'_{jm}s$ as elements. Then the elements of Λ are the square roots of the eigenvalues of the matrix

$$D = C_I^{-1} R C_J^{-1} R^T$$

or equivalently the square roots of the eigenvalues of the matrix

$$E = C_J^{-1} R^T C_I^{-1} R$$

where $R = (\widehat{p}_{ij} - \widehat{p}_{i.}\widehat{p}_{.j})$, $i = 1, \dots, I$, $j = 1, \dots, J$, C_I and C_J are the diagonal matrices given by $C_I = \operatorname{diag}(\widehat{p}_{1.}, \dots, \widehat{p}_{I.})$ and $C_J = \operatorname{diag}(\widehat{p}_{.1}, \dots, \widehat{p}_{.J})$. The columns of U are the normalized eigenvectors of D and the columns of V are the normalized eigenvectors of E according to constraints (1.2). For more details see for instance Anderson [3], Lebart et al [23], Benzecri [8], Greenacre [18]. Therefore, $\lambda_m, m = 1, \dots, M$ are replaced by the first canonical correlations $\widehat{\lambda}_m$, $m = 1, \dots, M$ and the u_{im} $(i = 1, \dots, I,\ m = 1, \dots, M)$ and v_{jm} $(j = 1, \dots, J,\ m = 1, \dots, M)$ are replaced by the first M pairs of canonical row and column scores $\widehat{u}_{im}$ $(i = 1, \dots, I,\ m = = 1, \dots, M)$ and $\widehat{v}_{jm}$ $(j = 1, \dots, J,\ m = 1, \dots, M)$.

Example 4.1. The first example is a 6×4 table from Srole et al [31, p. 213], which was analyzed by Gilula and Haberman in [17], among others. Table 1 shows a random sample of subjects in Midtown Manhattan cross-classified by mental health status and parental socioeconomic status.

Table 1.

	Parental socioeconomic status stratum					
Mental health category	A	B	C	D	E	F
Well	64	57	57	72	36	21
Mild symptom formation	94	94	105	141	97	71
Moderate symptom formation	58	54	65	77	54	54
Impaired	46	40	60	94	78	71

Firstly, we consider the model $H_1 : p_{ij} = p_{i.}p_{.j}(1 + \lambda_1 u_{i1} v_{j1})$ $i = 1, 2, 3, 4$, $j = 1, 2, 3, 4, 5, 6$. The restricted minimum $\phi_{(a_2)}$-divergence estimators for $a_2 = 0$, $a_2 =$

$2/3$ and $a_2 = 1$ of the unknown parameters are given in Table 2. We can observe that for $a_2 = 0$ we get the RMLE and for $a_2 = 1$ the minimum chi-square estimator.

In Tables 3, 4 and 5 we present the values of the statistic $T^{(m)}_{a_1,a_2} \equiv T^{(m)}_{\phi_{(a_1)},\phi_{(a_2)}}$ with $a_1 = -2, -1, -0.5, 0, 2/3$ and 1, given in (3.6), for $a_2 = 0, 2/3$ and 1, respectively. We use the subscript 2 in the values of the parameter a to indicate that these values will be used associated to the procedure of testing, i. e., associated to the function ϕ_2.

It is clear that the model with $M_0 = 1$ is quite adequate since the critical point at level 0.05 for the selection of an appropriate model is $\chi^2_{11,.05} = 19.675$.

Table 2. Estimates of the parameters of the model H_1 for $a_2 = 0, 2/3$ and 1.

Parameter	$a_2 = 0$	$a_2 = 2/3$	$a_2 = 1$
$p_{1.}$	.184940	.184881	.184852
$p_{2.}$	.362651	.362564	.362521
$p_{3.}$	.218072	.218303	.218417
$p_{4.}$	.234337	.234252	.234210
$p_{.1}$	.157831	.157803	.157789
$p_{.2}$	.147590	.147549	.147529
$p_{.3}$	.172892	.172830	.172799
$p_{.4}$	.231325	.231348	.231361
$p_{.5}$	.159639	.159662	.159677
$p_{.6}$	.130723	.130807	.130846
λ_1	.163016	.163016	.163016
u_{11}	−1.60321	−1.60532	−1.60636
u_{21}	−.187737	−.188727	−.189190
u_{31}	.086170	.091533	.094113
u_{41}	1.47561	1.47379	1.47290
v_{11}	−1.08699	−1.08752	−1.08780
v_{21}	−1.17351	−1.17373	−1.17385
v_{31}	−.370019	−.369479	−.369222
v_{41}	.053377	.053051	.052910
v_{51}	1.00962	1.00859	1.00809
v_{61}	1.79929	1.79919	1.79913

Table 3.

H_m v. H_{m+1}	$T^{(1)}_{-2,0}$	$T^{(1)}_{-1,0}$	$T^{(1)}_{-.5,0}$	$T^{(1)}_{0,0}$	$T^{(1)}_{2/3,0}$	$T^{(1)}_{1,0}$
1 versus 2	2.277942	2.273007	2.271424	2.270432	2.270036	2.270237

Table 4.

H_m v. H_{m+1}	$T^{(1)}_{-2,2/3}$	$T^{(1)}_{-1,2/3}$	$T^{(1)}_{-.5,2/3}$	$T^{(1)}_{0,2/3}$	$T^{(1)}_{2/3,2/3}$	$T^{(1)}_{1,2/3}$
1 versus 2	2.284696	2.276312	2.272991	2.270250	2.267502	2.266517

Table 5.

H_m v. H_{m+1}	$T^{(1)}_{-2,1}$	$T^{(1)}_{-1,1}$	$T^{(1)}_{-.5,1}$	$T^{(1)}_{0,1}$	$T^{(1)}_{2/3,1}$	$T^{(1)}_{1,1}$
1 versus 2	2.288831	2.278760	2.274592	2.271001	2.267112	2.265553

Example 4.2. As another example consider the data in Table 6 on the connection between frequency of attending meetings and social rank in Denmark. This example has been solved by the classical way without using statistical inference principles in Anderson [3].

Table 6.

Social group	Attend meetings outside working hours: One or more times a week	One or more times a month	Approx. once every second month	A few times a year	Never
I	17	27	13	24	25
II	25	57	17	49	55
III	38	91	41	217	213
IV	22	33	21	133	222
V	9	21	17	87	305

In Tables 7, 8 and 9 we present the values of the statistic $T^{(m)}_{a_1,a_2}$ for $a_2 = 0, 2/3$ and 1, respectively. These tables strongly suggests that a model with $M_0 = 2$ fits the data well while a model with $M_0 = 1$ would not suffice since the critical point at level 0.05 for the selection of an appropriate model is $\chi^2_{11,.05} = 19.675$.

Table 7.

H_m v. H_{m+1}	$T^{(m)}_{-2,0}$	$T^{(m)}_{-1,0}$	$T^{(m)}_{-.5,0}$	$T^{(m)}_{0,0}$	$T^{(m)}_{2/3,0}$	$T^{(m)}_{1,0}$
1 versus 2	35.426666	34.050673	33.504004	33.042161	32.548267	32.350210
2 versus 3	3.057359	3.070039	3.079596	3.091364	3.110592	3.121764

Table 8.

H_m v. H_{m+1}	$T^{(m)}_{-2,2/3}$	$T^{(m)}_{-1,2/3}$	$T^{(m)}_{-.5,2/3}$	$T^{(m)}_{0,2/3}$	$T^{(m)}_{2/3,2/3}$	$T^{(m)}_{1,2/3}$
1 versus 2	36.417790	34.587473	33.845204	33.204456	32.493337	32.194242
2 versus 3	3.109573	3.108382	3.110927	3.115607	3.125234	3.131527

Table 9.

H_m v. H_{m+1}	$T^{(m)}_{-2,1}$	$T^{(m)}_{-1,1}$	$T^{(m)}_{-.5,1}$	$T^{(m)}_{0,1}$	$T^{(m)}_{2/3,1}$	$T^{(m)}_{1,1}$
1 versus 2	37.061622	34.970345	34.117991	33.377952	32.548286	32.194872
2 versus 3	3.139479	3.131519	3.130671	3.131940	3.136977	3.140951

The model with $M_0 = 2$ corresponds to H_2, i.e., the model given by

$$p_{ij} = p_{i.}p_{.j}\left(1 + \lambda_1 u_{i1}v_{j1} + \lambda_2 u_{i2}v_{j2}\right), \quad i,j = 1,\dots,5. \tag{4.2}$$

The estimates of the parameters are shown in Table 10 and its analysis gives important information about the original data in Table 6. The different estimates obtained for $a_1 = 0$, $2/3$ and 1 are similar. For this reason we present only some comments for $a_1 = 0$. The estimated canonical correlation $\widehat{\lambda}_1 = 0.354498$ is fairly large, so "social group" and "attend meeting outside working hour" are related. The principal scores u_{11} and u_{21} as well as u_{12} and u_{22} are fairly similar this means that, as regards attending meeting, persons in social groups I and II behave in a similar fashion. The same happens for "one or more times a week" and "one or more times a month". From the relation (4.2), we have

$$\frac{p_{ij}}{p_{i.}p_{.j}} - 1 = \lambda_1 u_{i1}v_{j1} + \lambda_2 u_{i2}v_{j2}, \quad i,j = 1,\dots,5. \tag{4.3}$$

Then big values of the term $\lambda_1 u_{i1}v_{j1} + \lambda_2 u_{i2}v_{j2}$ correspond to dependence between the levels i and j of the categorical variables "social group" and "attend meeting outside working hour". On the basis of (4.3) we can conclude that persons in the two highest social groups attend meeting rather frequently and that persons in the two lowest social groups almost never attend meeting. These conclusions coincide with that of Example 11.2 in Anderson [3].

It should be noted, in the two examples, that the choice of different test statistics as well as estimators yields different values but no difference in model choice. It was of waiting, nevertheless it is not guaranteed for any problem and for every choice of ϕ and h. Asymptotically, the statistics have the same distribution, but in finite samples their performances will differ. To choose the "best" ϕ-divergence measure in estimation and testing, in the sense of efficiency and robustness, in these large classes of divergence measures depends on both finite-sample and asymptotic comparisons. Read and Cressie [30] give comparative results for the power divergence measures, although not for canonical analysis. Cressie and Read [12] concluded that for simple models, the test statistic based on power-divergence statistic, $a = 2/3$, offered an attractive alternative to the classical Pearson-based, $a = 1$, or likelihood-ratio-based, $a = 0$, test statistics. It will be interesting to see if this will also be the case for canonical analysis. In future work we will compare important family members, inter alia those considered here, through a simulation study for estimation and testing and for small samples.

Table 10. Estimates of the parameters of the model H_2 for $a_2 = 0, 2/3$ and 1.

parameter	$a_1 = 0$	$a_1 = 2/3$	$a_1 = 1$
$p_{1.}$	.059584	.059781	.059876
$p_{2.}$	.114109	.114125	.114135
$p_{3.}$	.337268	.337170	.337125
$p_{4.}$	.242271	.242308	.242322
$p_{5.}$	.246768	.246616	.246540
$p_{.1}$	.062394	.062678	.062816
$p_{.2}$	.128724	.128871	.128950
$p_{.3}$	.061270	.061367	.061414
$p_{.4}$	.286678	.286476	.286375
$p_{.5}$	.460933	.460607	.460445
λ_1	.354498	.354719	.354835
λ_2	.139138	.139187	.139214
u_{11}	-1.78609	-1.79056	-1.79270
u_{12}	-1.62033	-1.62754	-1.63090
u_{21}	-1.51095	-1.50779	-1.50625
u_{22}	-1.33422	-1.32677	-1.32312
u_{31}	-.463818	-.462214	-.461345
u_{32}	1.03137	1.03224	1.03261
u_{41}	.458414	.456868	.456187
u_{42}	.535603	.534382	.533897
u_{51}	1.31381	1.31475	1.31518
u_{52}	-.927256	-.927810	-.928145
v_{11}	-1.61581	-1.60943	-1.60629
v_{12}	-.881311	-.862700	-.853350
v_{21}	-1.65781	-1.65631	-1.65560
v_{22}	-1.01037	-1.00951	-1.00918
v_{31}	-.922196	-.928161	-.930940
v_{32}	-.401942	-.420273	-.428891
v_{41}	-.233528	-.231709	-.230786
v_{42}	1.55170	1.55336	1.55415
v_{51}	.949528	.950187	.950505
v_{52}	-.510189	-.510283	-.510355

5. CONCLUSIONS

Analysis of data by canonical analysis involves two steps. First, to estimate the unknown parameters in the model from the data. Second, to use these parameters estimates in statistical tests for the determination of the number of significant correlations. From the classical point of view the unknown parameters are estimated by the restricted maximum likelihood and Pearson and likelihood chi-square tests are used for testing. The first purpose of this paper is to present an analogue procedure but using the restricted minimum ϕ-divergence estimator jointly with a new family of statistics also based on a ϕ-divergence measure. The second one is to introduce a new procedure based on testing a sequence of nested hypotheses for the determination of the number of significant correlation. Finally, we apply the second procedure to two data sets studied previously by some authors.

ACKNOWLEDGEMENT

We would like to thank the referees for their comments and suggestions which helped to improve the paper.

(Received May 30, 2003.)

REFERENCES

[1] J. Aitchison and S. D. Silvey: Maximum-likelihood estimation of parameters subject to constraints. Ann. Math. Statist. *29* (1958), 813–828.

[2] S. M. Ali and S. D. Silvey: A general class of coefficients of divergence of one distribution from another. J. Roy. Statist. Soc. Ser. B *26* (1966), 131–142.

[3] E. B. A. Anderson: The Statistical Analysis of Categorical Data. Springer-Verlag, Berlin 1990.

[4] A. Basu and S. Basu:. Penalized minimum disparity methods in multinomials models. Statistica Sinica *8* (1998), 841–860.

[5] A. Basu and B. G. Lindsay: Minimum disparity estimation for continuous models. Ann. Inst. Statist. Math. *46* (1994), 683–705.

[6] A. Basu and S. Sarkar: Minimum disparity estimation in the errors-invariables model. Statist. Probab. Lett. *20* (1994), 69–73.

[7] A. Basu and S. Sarkar: The trade-off between robustness and efficiency and the effect of model smoothing. J. Statist. Comput. Simul. *50* (1994), 173–185.

[8] J. P. Benzecri: L'Analyse des Données. Tome 2: L'Analyse des Correspondances. Dunod, Paris 1973.

[9] M. W. Birch: A new proof of the Pearson–Fisher theorem. Ann. Math. Statist. *35* (1964), 817–824.

[10] N. A. C. Cressie and L. Pardo: Minimum ϕ-divergence estimator and hierarchical testing in loglinear models. Statistica Sinica *10* (2000), 867–884.

[11] N. A. C. Cressie and L. Pardo: Model checking in loglinear models using ϕ-divergences and MLEs. J. Statist. Plann. Inference *103* (2002), 437–453.

[12] N. A. C. Cressie and T. R C. Read: Multinomial goodness-of-fit tests. J. Roy. Statist. Soc. Ser. B *46* (1984), 440–464.

[13] N. J. Crichton and J. P. Hinde: Correspondence analysis as a screening method for indicants for clinical diagnosis. Statistics in Medicine *8* (1989), 1351–1362.

[14] I. Csiszár: Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität on Markhoffschen Ketten. Publ. Math. Inst. Hungar. Acad. Sci. Ser. A *8* (1963), 85–108.

[15] B. Dahdouh, J. F. Durantan, and M. Lecoq: Analyse des donnée sur l'ecologie des acridients d'Afrique de lóuest. Cahiers de l' Analyse des Données *3* (1978), 459–482.

[16] M. J. R. Fasham: A comparison of nonmetric multidimensional scaling, principal components averaging for the ordination of simulated coenocicles, and coenoplanes. Ecology *58* (1977), 551–561.

[17] Z. Gilula and J. Haberman: Canonical Analysis of Contingency Tables by Maximum Likelihood. J. Amer. Statist. Assoc. *81* (1986), 395, 780–788.

[18] M. J. Greenacre: Theory and Applications of Correspondence Analysis. Academic Press, New York 1984.
[19] M. Greenacre: Correspondence analysis in medical research. Statist. Meth. Medic. Res. *1* (1992), 97–117.
[20] M. J. Greenacre: Correspondence Analysis in Practice. Academic Press, London 1993.
[21] M. J. Greenacre: Correspondence Analysis of the Spanish National Health Survey. Department of Economics and Business, Universitat Pompeu Fabra, Barcelona 2002.
[22] H. O. Lancaster: The Chi-squared Distribution. Wiley, New York 1969.
[23] L. Lebart, A. Morineau, and K. Warwick: Multivariate Descriptive Statistical Analysis. Wiley, New York 1984.
[24] B. G. Lindsay: Efficiency versus robustness. The case for minimum Hellinger distance and other methods. Ann. Statist. *22* (1994), 1081–1114.
[25] G. B. Matthews and N. A. S. Crowther: A maximum likelihood estimation procedure when modelling categorical data in terms of cross-product ratios. South African Statist J. *31* (1997), 161–184.
[26] G. B. Matthews and N. A. S. Crowther: A maximum likelihood estimation procedures when modeling in terms of constraints. South African Statist. J. *29* (1995), 29–50.
[27] D. Morales, L. Pardo, and I. Vajda: Asymptotic divergence of estimates of discrete distributions. J. Statist. Plann. Inference *48* (1995), 347–369.
[28] W. C. Parr: Minimum distance estimation: a bibliography. Comm. Statist. Theory Methods *10* (1981), 1205–1224.
[29] J. A. Pardo, L. Pardo, and K. Zografos: Minimum ϕ-divergence estimators with constraints in multinomial populations. J. Statist. Plann. Inference *104* (2002), 221–237.
[30] T. R. C. Read and N. A. C. Cressie: Goodness-of-fit Statistics for Discrete Multivariat Data. Springer, New York 1988.
[31] L. Srole, T. S. Langner, S. T. Michael, M. K. Opler, and T. A. C. Reannie: Mental Health in the Metropolis: The Midtown Manhattan Study. McGraw-Hill, New York 1962.
[32] J. Wolfowitz: Estimation by minimum distance method. Ann. Inst. Statist. Math. *5* (1953), 9–23.

Julio A. Pardo, Leandro Pardo, and María Carmen Pardo, Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Plaza de Ciencias 3, 28040 Madrid. Spain.
e-mails: julio_pardo@mat.ucm.es, leandro_pardo@mat.ucm.es, mcapardo@mat.ucm.es

Konstantinos Zografos, Department of Mathematics, University of Ioannina, 451 10 Ioannina. Greece.
e-mail: kzograf@cc.uoi.gr