JOURNAL OF MULTIVARIATE ANALYSIS 54, 329–354 (1995)

Divergence-Based Estimation and Testing of Statistical Models of Classification*

M. MENÉNDEZ

Department of Applied Mathematics, Technical University of Madrid, Madrid, Spain

D. MORALES AND L. PARDO

Department of Statistics and Operations Research, Complutense University of Madrid, Madrid, Spain

AND

I. VAJDA

Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague, Czech Republic

The problems of estimating parameters of statistical models for categorical data, and testing hypotheses about these models are studied. Asymptotic properties of estimators minimizing ϕ-divergence between theoretical and empirical vectors of means are established. Asymptotic distributions of ϕ-divergences between empirical and estimated vectors of means are explictly evaluated, and tests based on these statistics are studied. The paper extends results previously established in this area.

1. INTRODUCTION

A frequent problem of categorical data analysis is that a fixed number n of samples $X=(X_1, ..., X_n) \in \mathcal{X}^n$ is taken from each of N different populations (families of individuals, clusters of objects). The sample space $\mathcal{X}$ is classified into r categories by a rule

$$\rho: \mathcal{X} \to \{1, ..., r\}.$$

Received June 29, 1994; revised February 1995

Key words and phrases: statistical classification, categorical data, clustered data, minimum divergence estimation, minimum divergence testing, asymptotic theory, optimality of testing.

* This work was supported by Grants DGICYT PB 93-0068, PB 93-0022, and GA CR 201/93|0232, and by the Sabatical Program of Complutense University of Madrid.

0047-259X/95 $12.00

Let $Y=(Y_1, ..., Y_r)$ be the classification vector with the components representing counts of the respective categories in the sample vector X; i.e., let

$$Y_j = \#\{1 \leqslant k \leqslant n: \rho(X_k) = j\}, \qquad 1 \leqslant j \leqslant r. \tag{1}$$

The sample space of the vector Y is denoted by $S_{n,r}$; i.e.,

$$S_{n,r} = \{y=(y_1, ..., y_r) \in \{0, 1, ..., n\}^r: y_1 + \cdots + y_r = n\}.$$

Populations $i=1, ..., N$ generate different sample vectors $X^{(i)}$ and the corresponding classification vectors $Y^{(i)}$. The sampled populations are assumed to be independent and homogeneous in the sense that $X^{(i)}$, and consequently $Y^{(i)}$, are independent realizations of the above considered X and Y. The i.i.d. property of the components $X_1, ..., X_n$ is included as a special case.

The aim of this paper is to present an extended class of methods for estimating parameters of statistical models of vectors Y and for testing statistical hypotheses about these models. Our methods are based on the so-called ϕ-divergences of probability distributions. They include as particular cases the well-known maximum likelihood method of estimation and Pearson's X^2-method of testing.

The classical statistical model of classification vectors $Y=(Y_1, ..., Y_r)$ is based on the assumption that the components of sample vectors $X=(X_1, ..., X_n)$ are i.i.d. Then the distribution $(p(y): y \in S_{n,r})$ of Y is multinomial (r-nomial) with parameters n and

$$\pi_j = E(1_{\rho^{-1}(j)}(X_1)) = \Pr(\rho(X_1)) = j), \qquad 1 \leqslant j \leqslant r, \tag{2}$$

where $\pi=(\pi_1, ..., \pi_r)$ is from the set

$$\Pi_r = \{\pi=(\pi_1, ..., \pi_r): \pi_j \in (0, 1), \pi_1 + \cdots + \pi_r = 1\}.$$

Further, the maximum likelihood estimator (MLE) $\hat{\pi}=(\hat{\pi}_1, ..., \hat{\pi}_r)$ of π, which is given by the formula

$$\hat{\pi}_j = \frac{1}{nN} \sum_{i=1}^{N} \sum_{k=1}^{n} 1_{\rho^{-1}(j)}(X_k^{(i)}),$$

takes on values belonging essentially to Π_r too. The well-known asymptotic properties of the estimator follow from the Bernoulli law and the Moivre–Laplace theorem. Analogously, the Pearson statistic

$$nN \sum_{j=1}^{r} \frac{(\hat{\pi}_j - \pi_{0j})^2}{\pi_{0j}}$$

is under the null hypothesis $H: \pi = \pi_0$, $\pi_0 = (\pi_{01}, ..., \pi_{0r}) \in \Pi_r$ asymptotically distributed as the well-known χ^2_{r-1}.

Extension to the case where the null hypothesis is composite, of the form

$$H: \pi \in (\pi(\theta) \in \Pi_r : \theta \in \Theta), \qquad \Theta \subset R^s \text{ open}, \qquad s < r-1, \tag{3}$$

is easy, namely by the method of Birch [3] developed for arbitrary discrete stochastic models. This method employs the MLE $\hat{\theta}$ of the unknown parameter θ_0 defined by the condition $\pi = \pi(\theta_0)$, and the Pearson statistic with $\pi(\hat{\theta})$ plugged-in for the unknown $\pi_0 = \pi(\theta_0)$,

$$nN \sum_{j=1}^{r} \frac{(\hat{\pi}_j - \pi_j(\hat{\theta}))^2}{\pi_j(\hat{\theta})}. \tag{4}$$

Under standard regularity conditions for $\pi(\theta)$, the estimator $\hat{\theta}$ is asymptotically normal with known asymptotic covariance matrix, and the statistic (4) is asymptotically distributed as χ^2_{r-s-1} (cf. Birch [3], or Bishop, Fienberg, and Holland [4]). The last result, in fact, holds for any best asymptotically normal (BAN) estimator; cf. Read and Cressie [14].

The classical multinomial model is not realistic enough. In practice, the samples within one population (relatives within one family, objects within one cluster) are often dependent. If members of the same cluster tend to respond similarly, then responses from the member of the same cluster will not be independent and hence the multinomial distribution will not be the correct distribution for the observed counts.

The first alternative to the multinomial model of $Y = (Y_1, ..., Y_r)$, resulting from a certain model of mutually dependent variables $\rho(X_1), ..., \rho(X_n)$, has been proposed by Cohen [6] and Altham [2]. In fact, they proposed the mixed model $p(y) = (1-a)\, p^*(y) + a p^{**}(y)$, $0 < a \leqslant 1$, where p^* is the multinomial with parameters $\pi_1, ..., \pi_r$ and p^{**} has the mass π_j residing at the jth extremal point $(0, ..., 0, n, 0, ..., 0)$ of $S_{n,r}$. Later Brier [5] proposed the Dirichlet-multinomial model $p(y)$, analyzed formerly by Mosimann [13], with parameters n, $\pi = (\pi_1, ..., \pi_r)$, and a nuissance parameter $K > 0$.

Within the frameworks of their respective models, the cited authors have considered null hypotheses under which the normed expectation $n^{-1}E(Y)$ is a Π_r-valued function $\pi(\theta) = (\pi_1(\theta), ..., \pi_r(\theta))$ on the same parameter space Θ as that considered in (3). They restricted themselves to the MLE $\hat{\theta}$ and to the normed sample means $\hat{\pi} = n^{-1}\bar{Y}$. By employing methods parallel to that of Birch [3], they proved that the statistic formally identical with (4) has, under the same regularity conditions for $\pi(\theta)$ as those considered by Birch, the same asymptotic distribution as (4).

The present paper exploits the observation that the last result remains true for more general models of classification vectors Y than those considered by the

cited authors. Therefore, we consider all models with regular normed Π_r-valued expectations $\pi(\theta)$ and suitable covariance matrices $V(\theta)$. Moreover, the MLE is replaced by a quasi-likelihood estimate $\hat{\theta}$ (cf., e.g., McCullagh and Nelder [11]), where the estimating function is an arbitrary ϕ-divergence between the observed sample means $\hat{\pi}$ and the theoretical means $\pi(\theta)$. The MLE is a special case obtained for an appropriate ϕ. Also the X^2-statistic (4) is replaced by a more general ϕ-divergence between the observed and estimated means $\hat{\pi}$ and $\pi(\hat{\theta})$. It is proved that if the estimate-defining ϕ is smooth enough then the estimate is BAN. If also the divergence-defining ϕ is smooth then the asymptotic distribution of the statistic can be evaluated. This means that the corresponding ϕ-divergence statistics can be used in testing hypotheses about the classification models under consideration. Asymptotically α-level goodness-of-fit tests are found for these statistics. Optimality of ϕ for these tests is investigated in the last part of the paper.

ϕ-divergences $D_\phi(p, q)$ of probability distributions p, q have been introduced by Csiszar [8] and, independently, by Ali and Silvey [1] as the expected values of likelihood ratio functions $\phi(p/q)$, where $\phi(u)$ is convex. Quasi-likelihood estimates $\hat{\theta}$ which maximize the ϕ-divergence $D_\phi(\hat{p}, p(\theta))$ between empirical and theoretical probability distributions have been studied by a number of authors; see Morales *et al.* [12] and references therein. The MLE $\hat{\theta}$ is the ϕ-divergence estimator for $\phi(u) = u \ln u$. The ϕ-divergence alternatives $D_\phi(\hat{p}, p(\hat{\theta}))$ to the statistic $\sum_j (\hat{p}_j - p_j(\hat{\theta}))^2 / p_j(\hat{\theta})$, which is the ϕ-divergence for $\phi(u) = (u-1)^2$, with the MLE estimate $\hat{\theta}$, have been introduced by Cressie and Read [7] and studied later in Read and Cressie [14] Salicrú *et al.* [15], and Morales *et al.* [12]. Note that in this paper the ϕ-divergence estimators and statistics are applied to properly normalized means $\hat{\pi}$ and $\pi(\theta)$ and not to probability distributions.

2. The Model and the Problem

Let us look in more detail at stochastic models of classification vectors $Y = (Y_1, ..., Y_r)$ represented by probability distributions $p = (p(y): y \in S_{n,r})$. As explained in Section 1, each model p is obtained from the distribution

$$q(j_1, ..., j_n) = \Pr(\rho_1 = j_1, ..., \rho_n = j_n), \qquad (j_1, ..., j_n) \in \{1, ..., r\}^n,$$

of classification $(\rho_1, ..., \rho_n) = (\rho(X_1), ..., \rho(X_n))$ of individuals observed in each population, according to

$$Y_j = \sum_{k=1}^{n} 1_{\{j\}}(\rho_k), \qquad 1 \leqslant j \leqslant r \quad \text{(cf. (1))}.$$

We consider the normed means

$$\pi = (\pi_1, ..., \pi_r) = n^{-1} E(Y) \tag{5a}$$

taking on values in Π_r, and the variances-covariances

$$V = (V_{ij})_{i,j=1}^{r} = E((Y - E(Y))'(Y - E(Y))). \tag{6a}$$

It is easy to see that the formulas

$$\pi_j = \frac{1}{n} \sum_k q_k(j), \tag{5b}$$

$$V_{jj} = \sum_k q_k(j)(1 - q_k(j)) + \sum_{k \neq m} (q_{k,m}(j, j) - q_k(j)\, q_m(j)), \tag{6b}$$

and

$$V_{ij} = \sum_{k \neq m} (q_{k,m}(i, j) - q_k(i)\, q_m(j)) - \sum_k q_k(i)\, q_k(j) \tag{6c}$$

hold for the marginals

$$q_k(j) = \Pr(\rho_k = j), \qquad 1 \leqslant k \leqslant n,$$

$$q_{k,m}(i, j) = \Pr(\rho_k = i, \rho_m = j), \qquad 1 \leqslant k, \quad m \leqslant n, \quad k \neq m,$$

of the primary distribution q.

We shall need the normed sample means

$$\hat{\pi} = n^{-1} \bar{Y} = (nN)^{-1} \sum_{i=1}^{N} Y^{(i)} \tag{7}$$

for the sample $Y^{(1)}, ..., Y^{(N)}$ i.i.d. by p. By the strong law of large numbers, $\hat{\pi} \to \pi$ a.s. (here, as well as everywhere in the sequel, the asymptotics are considered for $N \to \infty$). Therefore we may assume that

$$\pi \in \Pi_r, \qquad \hat{\pi} \in \Pi_r,$$

where the second relation holds asymptotically, on sets with probabilities approaching 1 exponentially.

The estimation and testing in this paper are based on the ϕ-divergence $D_\phi(\hat{\pi}, \pi)$ between stochastic vectors $\hat{\pi}$ and π. We recall that the

ϕ-divergence between arbitrary probability distributions has been introduced by Csiszar [8] and Ali and Silvey [1] for functions $\phi(u)$ convex on $(0, \infty)$. If $\mu, \nu \in \Pi_r$ then

$$D_\phi(\mu, \nu) = \sum_{j=1}^{r} \nu_j \phi\left(\frac{\mu_j}{\nu_j}\right). \tag{8}$$

Well-known are the I-divergence

$$I(\mu, \nu) = \sum_{j=1}^{r} \mu_j \ln \frac{\mu_j}{\nu_j}, \qquad \phi(u) = u \ln u, \tag{9}$$

and the X^2-divergence

$$X^2(\mu, \nu) = \sum_{j=1}^{r} \frac{(\mu_j - \nu_j)^2}{\nu_j}, \qquad \phi(u) = (1-u)^2. \tag{10}$$

Both (9) and (10) are particular cases of divergences considered by Cressie and Read [7] and defined in (21a)–(22b) below. For more examples and general properties of ϕ-divergences we refer to Liese and Vajda [9]. In this paper we restrict ourselves to ϕ-divergences (8) with $\phi(u)$ continuously differentiable in a neighborhood of $u = 1$ and $\phi''(1) \neq 0$.

Special attention will be paid to models for which there exists $c_n > 0$ such that

$$V = c_n(\operatorname{diag} \pi - \pi'\pi). \tag{11}$$

Note that here and in the sequel $\operatorname{diag}(y_1, ..., y_r)$ denotes the diagonal matrix with entries $y_1, ..., y_r$ at the diagonal.

The following assertion is essentially due to Cohen [6] ($n = 2$) and Altham [2] (general n).

LEMMA. *If there exist* $\alpha = (\alpha_1, ..., \alpha_r)$ *and* $(\alpha_{ij})_{i,j=1}^{r}$ *such that*

$$\alpha_j = q_k(j), \qquad \alpha_{ij} = q_{k,m}(i, j) \tag{12}$$

for all k *and* m *considered in* (5b) *and* (6b, 6c), *respectively, and if there exists* $0 \leqslant a \leqslant 1$ *such that*

$$\alpha_{ij} = (1-a)\,\alpha_i \alpha_j + a\alpha_i\,\delta_i(j), \tag{13}$$

then the corresponding model p *satisfies* (11) *for* $c_n = n + n(n-1)\,a$.

Proof. Under (12) it follows from (6b) that

$$V_{jj} = n\alpha_j(1-\alpha_j) + n(n-1)(\alpha_{jj} - \alpha_j^2)$$

and from (6c) that

$$V_{ij} = n(n-1)(\alpha_{ij} - \alpha_i \alpha_j) - n\alpha_i \alpha_j.$$

Inserting (13) one obtains

$$V_{jj} = c_n \alpha_j (1 - \alpha_j), \qquad V_{ij} = -c_n \alpha_i \alpha_j$$

for c_n given by the lemma. Finally, (5b) implies $\alpha = \pi$ which completes the proof.

In the following example, and in the sequel,

$$\binom{\gamma}{\gamma_1 \cdots \gamma_n} = \frac{\Gamma(\gamma+1)}{\Gamma(\gamma_1+1)\cdots\Gamma(\gamma_n+1)}.$$

EXAMPLE 1 (Multinomial). Let for $\pi = (\pi_1, ..., \pi_r) \in \Pi_r$

$$q(j_1, ..., j_n) = \pi_{j_1} \cdots \pi_{j_n}.$$

Then (12) holds for $\alpha = \pi$ (so that (5a) holds; i.e., our notation is consistent with (5a)) and $\alpha_{ij} = \pi_i \pi_j$. Therefore (13) is satisfied with $a = 0$. By Lemma, (11) holds for $c_n = n$. Here p is r-nomial with parameters n and π, i.e.,

$$p(y) = \binom{n}{y_1 \cdots y_n} \pi_1^{y_1} \cdots \pi_r^{y_r}, \qquad y \in S_{n,r}.$$

EXAMPLE 2 (Cohen [6]; Altham [2]). Let q^* be the same as q in Example 1 and

$$q^{**}(j_1, ..., j_n) = \begin{cases} \pi_j & \text{if } j_1 = \cdots = j_n = j, \\ 0, & \text{otherwise.} \end{cases}$$

Consider $0 < c \leqslant 1$ and put $q = (1-c)\, q^* + cq^{**}$. Then (12) holds for $\alpha = \pi$ (the notational consistency of the previous example holds) and

$$\alpha_{ij} = (1-c)\, \pi_i \pi_j + c\pi_i \delta_i(j).$$

Therefore (13) holds for $a = c$ and, consequently, (11) holds for $c_n = n + n(n-1)\, c$. In this model

$$p(y) = (1-c)\, p^*(y) + cp^{**}(y), \qquad y \in S_{n,r},$$

where p^* is the r-nomial distribution of the previous example and

$$p^{**}(y) = \begin{cases} \pi_i & \text{if } y_j = n, \\ 0, & \text{otherwise.} \end{cases}$$

EXAMPLE 3 (Brier [5]). Denote the distribution of Example 1 by $q(j_1, ..., j_n/\pi)$, consider on Π_r the Dirichlet density

$$f_{\pi, K}(x) = \binom{K-1}{K\pi_1 - 1 \cdots K\pi_r - 1} x_1^{K\pi_1 - 1} \cdots x_r^{K\pi_r - 1}$$

with parameters π and $K > 0$, and define

$$q(j_1, ..., j_n) = \int_{\Pi_r} q(j_1, ..., j_n/x) f_{\pi, K}(x)\, dx.$$

Then, in accordance with Example 1,

$$p(y) = \binom{n}{y_1 \cdots y_r} \int_{\Pi_r} x_1^{y_1} \cdots x_r^{y_r} f_{\pi, K}(x_1, ..., x_r)\, dx_1 \cdots dx_r$$

$$= \binom{n}{y_1 \cdots y_r} \binom{K-1}{K\pi_1 - 1 \cdots K\pi_r - 1} \Big/ \binom{n+K-1}{y_1 + K\pi_1 - 1 \cdots y_r + K\pi_r - 1}.$$

This is the so-called Dirichlet-r-nomial distribution with parameters n, π, and K. As shown by Brier [5] with a reference to Mosimann [13], the vector of parameters π in this example satisfies (5a), and (11) holds for $c_n = n(n+K)/(K+1)$. In the lemma this corresponds to $a = (K+1)^{-1}$.

If $n = 1$ then the classification model considered in this paper is equivalent to the classical discrete model of statistical inference with sample space $\{1, ..., r\}$ and probabilities $\pi_1, ..., \pi_r$. Indeed, there is a one–one correspondence between the random vectors

$$Y \in S_{1,r} = \{(1, 0, ..., 0), (0, 1, ..., 0), ..., (0, ..., 0, 1)\}$$

and the random variables $Z = p(X)$ taking on values $1 \leqslant j \leqslant r$ with probabilities $\Pr(Z = j)$ equal to

$$\pi_j = E(1_{(j)}(Z)) = \Pr(Y_j = 1) \qquad \text{(cf. (1), (5a))}.$$

To formulate precisely the problems studied in this paper, consider a family $\mathscr{P} = (p_\theta : \theta \in \Theta)$ of probability distributions on $S_{n,r}$ with the same Θ as in (3). We consider the composite hypothesis

$$H: p \in \mathscr{P}. \tag{14}$$

The exact meaning of (14) is that the true distribution p of Y equals p_{θ_0}, where $\theta_0 \in \Theta$ is an unknown true parameter value. One of the two problems we solve is to find a suitable statistical α-level test of H. The other problem is to find under H a suitable consistent point estimator of θ_0. The

tests and estimators are assumed to be based on the sample $Y^{(1)}, ..., Y^{(N)}$ with independent components distributed by p_{θ_0}.

Since the unknown parameter varies over Θ, we consider the parameteric functions

$$\pi = \pi(\theta), \qquad V = V(\theta), \qquad \theta \in \Theta,$$

defined by (5a) and (6a) with p replaced by p_θ.

The only restriction on the hypothesis H in this paper is the regularity of $\pi(\theta)$ summarized in assumptions (A1)–(A3). Such assumptions have been considered already for Birch [3]. The first regularity assumption is quite natural.

(A1) The mapping inverse to $\theta \to \pi(\theta)$ exists and is continuous at θ_0.

In the next assumption we consider the gradient $\nabla = (\partial/\partial\theta_1, ..., \partial/\partial\theta_s)$, the $r \times s$ Jacobian of $\pi(\theta)$ with row vectors $\nabla\pi_j(\theta)$, denoted by $J(\theta)$ and the particular values $J_0 = J(\theta_0)$ and $\pi_0 = (\pi_{01}, ..., \pi_{0r}) = \pi(\theta_0)$.

(A2) The mapping $\pi(\theta)$ is continuously differentiable in an open neighborhood of θ_0 so that

$$(\pi(\theta) - \pi_0)^t = J_0(\theta - \theta_0)^t + o(\|\theta - \theta_0\|) \qquad \text{for} \quad \theta \to \theta_0.$$

In the last assumption we consider the vectors $\pi_0^\alpha = (\pi_{01}^\alpha, ..., \pi_{0r}^\alpha)$ defined for all real α; more precisely, we refer to the $r \times s$-matrix $A = \operatorname{diag} \pi_0^{-1/2} J_0$.

(A3) The $r \times s$-matrix $A^t A$ is positive definite.

3. Estimation

In this section we consider the estimation problem introduced in Section 2. An estimator $\hat{\theta}$ is a sequence of measurable functions $\hat{\theta}_N(Y^{(1)}, ..., Y^{(N)})$ taking on values in Θ. This estimator is consistent if $\hat{\theta}$ tends in probability to θ_0. It is c_N-consistent if $\|\hat{\theta} - \theta_0\| = O_p(c_N^{-1})$.

An estimator $\hat{\theta}$ is said to be a minimum ϕ-divergence estimator (briefly $M\phi E$) if

$$\hat{\theta} = \arg\min D_\phi(\hat{\pi}, \pi(\theta)).$$

Note that the estimators minimizing ϕ-divergence between the sample-based empirical distribution and theoretical distributions under the hypothesis (14) were considered by many authors (see, e.g., the references in Vajda [6]). The estimators minimizing ϕ-divergences between sample means and their theoretical expectations seem to be new. This approach is

a variant of the well-known method of moments and for $\phi(u) = u \ln u$ in Example 1 it coincides with the maximum likelihood method.

The following theorem summarizes asymptotic properties of minimum divergence estimators $\hat{\theta}$ and, also, properties of the function $\pi(\hat{\theta})$ for references later.

THEOREM 1. *Let assumptions* (A1)–(A3) *hold and let us consider the matrices*

$$B = \operatorname{diag} \pi_0^{-1/2} \frac{V(\theta_0)}{n^2} \operatorname{diag} \pi_0^{-1/2},$$
$$C = A(A^t A)^{-1}, \qquad D = CA^t = A(A^t A)^{-1} A^t. \tag{15}$$

Then for all convex functions $\phi(u)$ *under consideration the* $M\phi E\hat{\theta}$ *satisfies the asymptotic relations*

$$\hat{\theta} \to \theta_0 \qquad \text{a.s.}, \tag{16a}$$

$$\hat{\theta} = \theta_0 + (\hat{\pi} - \pi_0) \operatorname{diag} \pi_0^{-1/2} C + (1 + o_p(1)), \tag{16b}$$

$$N^{1/2}(\hat{\theta} - \theta_0) \to N(0, C^t BC) \qquad \text{in law}, \tag{16c}$$

$$\pi(\hat{\theta}) = \pi_0 + (\hat{\pi} - \pi_0) \operatorname{diag} \pi_0^{-1/2} D \operatorname{diag} \pi_0^{1/2} + (1 + o_p(1)), \tag{16d}$$

$$N^{1/2}(\pi(\hat{\theta}) - \pi_0) \to N(0, \operatorname{diag} \pi_0^{-1/2} DBD \operatorname{diag} \pi_0^{1/2}) \qquad \text{in law}. \tag{16e}$$

Proof. (a) By (A1), (16a) holds if

$$|\pi_0 - \pi(\hat{\theta})| \to 0 \qquad \text{a.s.},$$

where $|\cdot|$ denotes the absolute norm of vectors from R^r. The triangle inequality and symmetry of this norm imply

$$|\pi_0 - \pi(\hat{\theta})| \leqslant |\pi_0 - \hat{\pi}| + |\hat{\pi} - \pi(\hat{\theta})|.$$

By the strong law of large numbers, $|\pi_0 - \hat{\pi}| \to 0$ a.s. By the definition of $\hat{\theta}$ and the theorem about the range of ϕ-divergences in Liese and Vajda [9],

$$\phi(1) \leqslant D_\phi(\hat{\pi}, \pi(\hat{\theta})) \leqslant D_\phi(\hat{\pi}, \pi_0),$$

where $D_\phi(\hat{\pi}, \pi_0) \to \phi(1)$ a.s. Therefore $D_\phi(\hat{\pi}, \pi(\hat{\theta})) \to \phi(1)$ a.s. But, by Proposition 9.49 in Vajda [16], this implies $|\hat{\pi} - \pi(\hat{\theta})| \to 0$ a.s. As said above, this is sufficient for (16a).

(b) By the central limit theorem, it follows from (7) and (6a) that

$$N^{1/2}(\hat{\pi} - \pi_0) \to N(0, n^{-2} V(\theta_0)) \qquad \text{in law}. \tag{17}$$

In particular,

$$\|\hat{\pi} - \pi_0\| = O_p(N^{-1/2}). \tag{18}$$

Now consider the vector function $\Psi(q, \theta) = \nabla D_\phi(q, \pi(\theta))$ of variables $q = (q_1, ..., q_r) \in \Delta_r$ and $\theta \in \Theta$. It holds that

$$\Psi(q, \theta) = h(q, \theta)\, J(\theta),$$

where $h(q, \theta) = (h_1(q, \theta), ..., h_m(q, \theta))$ is a vector function with components

$$h_j(q, \theta) = \phi\left(\frac{q_j}{\pi_j(\theta)}\right) - \frac{q_j}{\pi_j(\theta)}\,\phi'\left(\frac{q_j}{\pi_j(\theta)}\right). \tag{19}$$

Consequently,

$$\frac{\partial}{\partial q_j}\,\Psi(q, \theta) = -\frac{q_j}{\pi_j^2(\theta)}\,\phi''\left(\frac{q_j}{\pi_j(\theta)}\right)\nabla\pi_j(\theta).$$

Taking into account (18), the convergences $\hat{\pi} \to \pi_0$, $\pi(\hat{\theta}) \to \pi_0$, and $\nabla\pi_j(\hat{\theta}) \to \nabla\pi_j(\theta_0)$ (cf. (16a), and the continuous differentiability of $\pi(\theta)$ in (A2)), one obtains from the mean value theorem

$$\begin{aligned}\Psi(\hat{\pi}, \hat{\theta}) - \Psi(\pi_0, \hat{\theta}) &= -\phi''(1) \sum_{j=1}^{r} \frac{\nabla\pi_j(\theta_0)}{\pi_{0j}}(\hat{\pi}_j - \pi_{0j}) + o_p(N^{-1/2}) \\ &= -\phi''(1)(\hat{\pi} - \pi_0)\,\operatorname{diag}\pi_0^{-1/2} A + o_p(N^{-1/2}). \end{aligned} \tag{20}$$

Let us now evaluate $\Psi(\pi_0, \hat{\theta})$ for $\hat{\theta}$ from the neighborhood of θ_0 by applying the mean value theorem to the function $h(\pi_0, \hat{\theta})$ (cf. (19)). It holds that $h_j(\pi_0, \theta_0) = \phi(1) - \phi'(1)$ and, by (16a) and the mean value theorem,

$$h_j(\pi_0, \hat{\theta}) = \phi(1) - \phi'(1) + \phi''(1)\,\frac{\nabla\pi_j(\theta_0)(\hat{\theta} - \theta_0)^t}{\pi_{0j}} + (1 + o_p(1))$$

or, equivalently,

$$h(\pi_0, \hat{\theta}) = c + \phi''(1)(\hat{\theta} - \theta_0)\, A^t \operatorname{diag}\pi_0^{-1/2} + (1 + o_p(1)),$$

where c is the r-vector of constants $\phi(1) - \phi'(1)$. Any such vector satisfies the identity $cJ(\theta) = 0$ on Θ. Therefore, by definition,

$$\Psi(\pi_0, \hat{\theta}) = \phi''(1)(\hat{\theta} - \theta_0)\, A^t \operatorname{diag}\pi_0^{-1/2} J(\hat{\theta}) + (1 + o_p(1)),$$

where, by (16a) and (A2), the matrix $A^t \operatorname{diag}\pi_0^{-1/2} J(\hat{\theta})$ tends a.s. to the matrix A^tA which is positive definite by (A3).

Now, by the definition of estimate $\hat{\theta}$, $\Psi(\hat{\pi}, \hat{\theta}) = 0$. Hence we obtain from the last formula and (20) the identity

$$\phi''(1)(\hat{\theta} - \theta_0) = \phi''(1)(\hat{\pi} - \pi_0) \operatorname{diag} \pi_0^{-1/2} A(A^t A)^{-1} + 1 + o_p(1).$$

By the assumption $\phi''(1) \neq 0$ and (17), (16b) follows from here.

(c) The convergence (16c) follows from (16b) and (17).

(d) Employing as in (b) the mean value theorem and taking into account (16b), one obtains (16d).

(e) The convergence (16e) follows from (16d) and (17). Q.E.D.

Note that the assumptions of the present paper concerning ϕ are satisfied by the nonnegative convex functions

$$\phi_a(u) = \frac{u^a - au + a - 1}{a(a-1)}, \qquad a \neq 0, \qquad a \neq 1, \tag{21a}$$

as well as by their limits

$$\phi_0(u) = -\ln u + u - 1, \qquad \phi_1(u) = u \ln u - u + 1. \tag{21b}$$

The corresponding distances are defined in accordance with (8) by

$$D_a(\mu, \nu) = \frac{1}{a(a-1)} \left(\sum_{j=1}^{r} \mu_j^a \nu_j^{1-a} - 1 \right), \tag{22a}$$

or by the corresponding limits,

$$D_1(\mu, \nu) = \sum_{j=1}^{r} \mu_j \ln \frac{\mu_j}{\nu_j}, \qquad D_0(\mu, \nu) = D_1(\nu, \mu). \tag{22b}$$

This family of distances defines a family of estimators $\hat{\theta}^{(a)}$, $a \in R$.

An interesting problem is how to choose a convex function ϕ such that the corresponding $M\phi E$ $\hat{\theta}^{(\phi)}$ is optimal. The solution obviously depends on the optimality. If the optimality is represented by the asymptotic variance of $\hat{\theta}^{(\phi)}$ then it follows from Theorem 1 that, in the models satisfying (A1)–(A3), all $M\phi E$'s under consideration are equivalent. But for finite sample sizes N the variances of these estimators depend on ϕ and the stated problem becomes nontrivial. Similar methods as used in Section 5 can be employed, e.g., to demonstrate that there exist values $a \in R$ optimal for the above-considered estimators $\hat{\theta}^{(a)}$. An alternative approach based on a residual adjustment function has been presented recently by Lindsay [10].

4. Testing

In this section we consider the testing problem and related concepts and assumptions of Section 2. By Theorem 1, the point estimator $\hat{\theta}$ assumed in Theorem 2 below may be the $M\phi E$ for convex ϕ under consideration. The function ϕ in Theorem 1 may of course differ from the ϕ considered in Theorem 2.

THEOREM 2. *Let M be an $r \times r$-matrix and let $\hat{\theta}$ be a point estimator such that*

$$\pi(\hat{\theta}) = \pi_0 + (\hat{\pi} - \pi_0)\, M + o_p(N^{-1/2}). \tag{23}$$

Then for every convex ϕ under consideration

$$nN(D_\phi(\hat{\pi}, \pi(\hat{\theta})) - \phi(1)) \to \frac{\phi''(1)}{2} \sum_{j=1}^{r} \lambda_j Z_j^2 \qquad \text{in law,} \tag{24}$$

where Z_j are mutually independent standard normal random variables and $\lambda_j \geqslant 0$ are eigenvalues of the matrix

$$L = \operatorname{diag} \pi_0^{-1/2} (I - M)^t \frac{V(\theta_0)}{n} (I - M) \operatorname{diag} \pi_0^{-1/2}. \tag{25}$$

Proof. Under (23),

$$\hat{\pi} = \pi(\hat{\theta}) = (\hat{\pi} - \pi_0)(I - M) + o_p(N^{-1/2}).$$

It follows from here and (17) that

$$N^{1/2}(\hat{\pi} - \pi(\hat{\theta})) \operatorname{diag} \pi_0^{-1/2} \to N(0, n^{-1}L) \qquad \text{in law}$$

for the matrix L defined by (25), and also

$$(\hat{\pi} - \pi(\hat{\theta})) \operatorname{diag} \pi(\hat{\theta})^{-1/2} = (\hat{\pi} - \pi(\hat{\theta})) \operatorname{diag} \pi_0^{-1/2} + o_p(N^{-1/2}).$$

Therefore,

$$(nN)^{1/2} (\hat{\pi} - \pi(\hat{\theta})) \operatorname{diag} \pi(\hat{\theta})^{-1/2} \to N(0, L) \qquad \text{in law.}$$

It is well known that if a sequence of random r-vectors $\hat{U} = \hat{U}^{(N)}$ and an $r \times r$-matrix L satisfy $\hat{U} \to N(0, L)$ in law, and L has eigenvalues $\lambda_1, \ldots, \lambda_r$, then

$$\hat{U}\hat{U}^t \to \sum_{j=1}^{r} \lambda_j Z_j^2 \qquad \text{in law,}$$

where Z_j are independent $N(0, 1)$. Applying this to

$$\hat{U} = (nN)^{1/2} (\hat{\pi} - \pi(\hat{\theta})) \operatorname{diag} \pi(\hat{\theta})^{-1/2}$$

and, using the obvious identity

$$\hat{U}\hat{U}^t = nNX^2(\hat{\pi}, \pi(\hat{\theta})) \qquad \text{(cf. (11))},$$

one obtains

$$nNX^2(\hat{\pi}, \pi(\hat{\theta})) \to \sum_{j=1}^{r} \lambda_j Z_j^2 \qquad \text{in law,} \tag{26}$$

where λ_j and Z_j are defined in accordance with Theorem 2. The non-negativity of eigenvalues follows from the fact that L is a covariance matrix.

Taking into account that the X^2-divergence defined by (11) is ϕ_*-divergence for $\phi_*(u) = (1-u)^2$, where $\phi_*(1) = 0$ and $\phi_*''(1) = 2$, we see that (24) is proved for $\phi = \phi_*$. The desired extension follows from the fact that if $\hat{\pi} = \hat{\pi}^{(N)}$ and $\hat{\hat{\pi}} = \hat{\hat{\pi}}^{(N)}$ are arbitrary sequences of random vectors with values in π_r, satisfying the asymptotic condition

$$\|\hat{\pi} - \pi_0\| = O_p(N^{-1/2}), \qquad \|\hat{\hat{\pi}} - \pi_0\| = O_p(N^{-1/2}), \tag{27}$$

then for every ϕ under consideration

$$D_\phi(\hat{\pi}, \hat{\hat{\pi}}) = \phi(1) + \frac{\phi''(1)}{2} X^2(\hat{\pi}, \hat{\hat{\pi}}) + o_p(N^{-1}). \tag{28}$$

Indeed, (24) follows from (26) and (28). Thus it remains to prove that (27) implies (28). To this end consider coordinates $\hat{\pi}_j$ and $\hat{\hat{\pi}}_j$ of vectors $\hat{\pi}$ and $\hat{\hat{\pi}}$. By Taylor's theorem,

$$\phi\left(\frac{\hat{\pi}_j}{\hat{\hat{\pi}}_j}\right) = \phi(1) + \phi'(1)\left(\frac{\hat{\pi}_j}{\hat{\hat{\pi}}_j} - 1\right) + \frac{1}{2}\phi''\left(\frac{\pi_j^*}{\pi_j^{**}}\right)\left(\frac{\hat{\pi}_j}{\hat{\hat{\pi}}_j} - 1\right)^2,$$

where π^*, π^{**} are r-vectors with the norms $\|\pi^* - \pi_0\|$, $\|\pi^{**} - \pi_0\|$ bounded above by $\|\hat{\pi} - \pi_0\|$, $\|\hat{\hat{\pi}} - \pi_0\|$, respectively. By (8),

$$\begin{aligned} D_\phi(\hat{\pi}, \hat{\hat{\pi}}) &= \phi(1) + \frac{1}{2}\sum_{j=1}^{r} \phi''\left(\frac{\pi_j^*}{\pi_j^{**}}\right)\frac{(\hat{\pi}_j - \hat{\hat{\pi}}_j)^2}{\hat{\hat{\pi}}_j} \\ &= \phi(1) + \frac{\phi''(1)}{2} X^2(\hat{\pi}, \hat{\hat{\pi}}) + Z^{(N)}, \end{aligned}$$

where

$$Z^{(N)} = \frac{1}{2}\left(\sum_{j=1}^{r} \phi'' \left(\frac{\pi_j^*}{\pi_j^{**}} \right) - \phi''(1) \right).$$

By (27), $Z^N = o_p(N^{-1})$. Q.E.D.

If (A1)–(A3) hold then the estimators $\hat{\theta}$ of Theorem 1 satisfy (23) for

$$M = \operatorname{diag} \pi_0^{-1/2} A (A^t A)^{-1} A^t \operatorname{diag} \pi_0^{1/2}. \tag{29}$$

In this case Theorem 2 can be made precise as follows.

COROLLARY. *Let* (A1)–(A3) *hold. If the model satisfies* (11), *i.e., if*

$$V(\theta_0) = c_n(\operatorname{diag} \pi_0 - \pi_0^t \pi_0) \tag{30}$$

for some positive real c_n, *then for all convex functions* ϕ, ϕ_* *under consideration and the* $M\phi_* E\,\hat{\theta}$,

$$nN(D_\phi(\hat{\pi}, \pi(\hat{\theta})) - \phi(1)) \to \frac{c_n \phi''(1)}{2n} \chi^2_{r-s-1} \qquad \textit{in law.} \tag{31}$$

Proof. Assumptions of Theorem 2 hold for M given by (29). We shall prove that, under (A1)–(A3), (29), and (30), the matrix (25) satisfies the relation

$$L = \frac{c_n}{n} (I - (\pi_0^{1/2})^t \pi_0^{1/2} - D), \tag{32}$$

where

$$\pi_0^{1/2} = (\pi_{01}^{1/2}, \ldots, \pi_{0r}^{1/2})$$

and D is given by (15). Analogously as on page 517 of Bishop, Fienberg, and Holland [4], under (32) it suffices to prove that the matrix $I - (\pi_0^{1/2})^t \pi_0^{1/2} - D$ is idempotent with the trace $r - s - 1$. The idempotence is clear from the relations

$$\pi_0^{1/2} D = D(\pi_0^{1/2})^t = 0, \qquad D^t D = D,$$

where the first one follows from the obvious equalities

$$\pi_0^{1/2} A = \pi_0^{1/2} \pi_0^{-1/2} J_0 = \left(\sum_{j=1}^{r} \frac{\partial \pi_j}{\partial \theta_1}(\theta_0), \ldots, \sum_{j=1}^{r} \frac{\partial \pi_j}{\partial \theta_1}(\theta_0) \right)$$

(cf. definitions of A and J_0 in Section 1) and the second one is clear from the definition of D. Futher,

$$\mathrm{Tr}(I) = r, \qquad \mathrm{Tr}((\pi_0^{1/2})^t \pi_0^{1/2}) = 1,$$
$$\mathrm{Tr}(D) = \mathrm{Tr}(A^t A)^{-1} A^t A) = s \qquad \text{(cf. (A3))}.$$

Therefore it suffices prove (32). To this end take into account that, by (29),

$$(I - M) \operatorname{diag} \pi_0^{-1/2} = (\operatorname{diag} \pi_0^{-1/2} \operatorname{diag} \pi_0^{1/2} - M) \operatorname{diag} \pi_0^{-1/2}$$
$$= \operatorname{diag} \pi_0^{-1/2} (I - D)$$

and, by (30),

$$\operatorname{diag} \pi_0^{-1/2} V(\theta_0) \operatorname{diag} \pi_0^{-1/2} = c_n (I - (\pi_0^{1/2})^t \pi_0^{1/2}).$$

Therefore, by (25),

$$L = \frac{c_n}{n} (I - D)^t (I - (\pi_0^{1/2})^t \pi_0^{1/2})(I - D).$$

By multiplying the matrices and taking into account the above relations for π_0 and D one obtains (32). Q.E.D.

The corollary offers a family of asymptotically α-level tests $(T^{(\phi)}, \chi^2_{r-s-1}(1-\alpha)\colon \phi \in \Phi)$ of the hypothesis (14) for the class Φ of convex functions considered in this paper. The test statistics are defined by

$$T^{(\phi)} = \frac{2n^2 N}{c_n \phi''(1)} (D_\phi(\hat{\pi}, \pi(\hat{\theta})) - \phi(1)) \tag{33}$$

and the critical values are the $(1-\alpha)$-quantiles $\chi^2_{r-s-1}(1-\alpha)$ of χ^2-distributed random variable with $r-s-1$ degrees of freedom. The estimates $\hat{\theta}$ figuring in (33) are the $M\phi_* E$'s for $\phi_* \in \Phi$. There are thus two "free parameters" of the testing procedure under consideration, namely the functions $\phi, \phi_* \in \Phi$. A particular solution of the problem of optimal choice of these parameters, namely the optimal choice of ϕ, is considered in Section 5 below.

Examples 1–3 illustrate situations where the above considered tests can be applied. In practical applications, however, the value c_n defined by (11) and figuring in (33) might be unknown. In the models satisfying the assumptions of the lemma in Section 2, one can use the fact that $n \leqslant c_n \leqslant n^2$. If one is uncertain about the assumptions of the lemma, he can use the sample-based approximation

$$\hat{c}_{n,N} = \frac{n^2}{(N-1)(r-1)} \sum_{i=1}^{N} \frac{(\hat{\pi}_j - \pi_j(\hat{\theta}))^2}{\pi_j(\hat{\theta})}.$$

As shown by Brier [5], $\hat{c}_{n,N}$ consistently estimates c_n under (11). Therefore the replacement of c_n by $\hat{c}_{n,N}$ preserves the considered asymptotic distribution of the statistic (33).

Remark. Throughout the paper we suppose that a fixed number n of samples $X_1, ..., X_n$ is taken from each of N independent populations. This assumption is often violated (e.g., due to missing values in some populations). But the extension of our results to this more general situation is relatively easy. Suppose that the populations can be clustered into a fixed number of groups indexed by γ which differ only in that the number of samples $n = n_\gamma$ varies with the group. Denote by N_γ the number of populations in the group γ (so that $N = \sum_\gamma N_\gamma$) and replace the assumption $N \to \infty$ by $N_\gamma \to \infty$ by $N_\gamma \to \infty$ for each γ. Then all above stated results hold separately for each population group.

We shall describe the extension of the corollary to the union of all groups under the assumption that in each group the model satisfies (11) with the vector of parameter $\pi = (\pi_1, ..., \pi_r)$ not depending on γ (or on n, which is, in the given context, equivalent). Put for brevity

$$\Sigma_0 = \operatorname{diag} \pi_0 - \pi_0^t \pi_0 \qquad \text{for} \quad \pi_0 = \pi(\theta_0)$$

and

$$M_0 = \operatorname{diag} \pi_0^{-1/2} \Sigma_0 \operatorname{diag} \pi_0^{-1/2}, \qquad B_\gamma = \frac{c_{n_\gamma}}{n_\gamma^2} M_0.$$

Denote by $\hat{\pi}_\gamma$ the normed sample mean in the group γ defined by (7) with n, N replaced by n_γ, N_γ. Let $\hat{\theta}_\gamma$ be an estimator satisfying (16d) with B replaced by B_γ. Finally, define

$$\hat{\pi} = \frac{\Sigma_\gamma N_\gamma \hat{\pi}_\gamma}{\Sigma_\gamma n_\gamma N_\gamma}, \qquad \tilde{\pi} = \frac{\Sigma_\gamma n_\gamma N_\gamma \pi(\hat{\theta}_\gamma)}{\Sigma_\gamma n_\gamma N_\gamma}, \qquad \nu = \frac{\Sigma_\gamma n_\gamma N_\gamma}{(\Sigma_\gamma c_{n_\gamma} N_\gamma)^{1/2}}.$$

By (16d)

$$\hat{\pi} = \pi_0 + (\hat{\pi} - \pi_0) \operatorname{diag} \pi_0^{-1/2} D \operatorname{diag} \pi_0^{1/2} + o_p(\min_\gamma N_\gamma^{-1/2}), \quad (16d*)$$

$$\hat{\pi} - \tilde{\pi} = (\hat{\pi} - \pi_0)(I - D) + o_p(\min_\gamma N_\gamma^{-1/2}), \quad (16d**)$$

and by (17)

$$\nu(\hat{\pi} - \pi_0) \to N(0, \Sigma_0) \qquad \text{in law.} \quad (17*)$$

Combining (16d**) and (17*) one obtains

$$v(\hat{\pi} - \tilde{\pi}) \operatorname{diag} \pi_0^{\,1/2} \to N(0, \tilde{L}),$$

where

$$\tilde{L} = \operatorname{diag} \pi_0^{-1/2}(I - D)^t \Sigma_0 (I - D) \operatorname{diag} \pi_0^{-1/2}$$

and, by a similar argument as in the proof of the corollary,

$$\tilde{L} = (I - (\pi_0^{-1/2})^t \pi_0^{-1/2} - D).$$

Combining (16d*) and (16d**) one obtains from here

$$v(\hat{\pi} - \tilde{\pi}) \operatorname{diag} \tilde{\pi}^{-1/2} - N(0, \tilde{L}).$$

Finally, by repeating essentially the argument of the proof of Theorem 2 and its Corollary, one obtains from here that for every ϕ under consideration

$$v^2(D_\phi(\hat{\pi}, \tilde{\pi}) - \phi(1)) \to \frac{\phi''(1)}{2} \chi^2_{r-s-1} \qquad \text{in law.}$$

This is an extension of (31) which leads to generalized test statistics (33). If there is only one group with $n_y = n$ then $v^2 = Nn^2/c_n$ and the last result reduces to (31).

5. Choice of ϕ

In this section we investigate the problem of optimal specification of ϕ in the test statistic (33) for a fixed sample size N. The aim is to demonstrate that the choice of ϕ has significant practical consequences, rather than to present a practically significant solution.

As pointed out in Section 2, the model of classification considered in this paper is for $n = 1$, equivalent to the classical discrete statistical model with sample space $\{1, ..., r\}$ and probabilities $\pi_1, ..., \pi_r$. In the classical discrete model the problem of optimal choice of ϕ from the special family $(\phi_a : a \in R)$ defined by (21) has been studied by Cressie and Read [7]. We extend their study to the classification model of the present paper. Our results obtained for arbitrary n may thus be verified by putting $n = 1$ and comparing with their results.

The methods used here are similar to those of Cressie and Read. We therefore omit motivation or justification of these methods. In particular, (14) is replaced by the simple hypothesis $H: \pi = \pi_0 \in H_r$.

Further, we restrict ourselves to the model of Brier [5] described in Example 3. The test statistics are thus assumed to be defined by (33) for $c_n = n(n+K)/(K+1)$ (cf. Example 3) and for distances $D_a(\mu, \nu)$, $a \in R$, defined by (22).

Some of these distances are well known. As follows from (22b) and (10),

$$D_1(\mu, \nu) = I(\mu, \nu), \qquad D_0(\mu, \nu) = I(\nu, \mu)$$

(in general, $D_a(\mu, \nu) = D_{1-a}(\nu, \mu)$ for all $a \in R$). Further,

$$D_2(\mu, \nu) = X^2(\mu, \nu),$$
$$D_{-1}(\mu, \nu) = X^2(\nu, \mu) \qquad \text{(cf. (11) and (22))}.$$

It is easy to see that $D_{1/2}(\mu, \nu) = 2H^2(\mu, \nu)$ for the Hellinger metric distance

$$H(\mu, \nu) = \left(\sum_{j=1}^{r} (\mu_j^{1/2} - \nu_j^{1/2})^2 \right)^{1/2}.$$

In general, $D_a(\mu, \nu)$ are monotone functions of the distances of Renyi [17] given (in the variant extended to all real $a \neq 0$, $a \neq 1$ by Liese and Vajda [9]) by the formula

$$R_a(\mu, \nu) = \frac{1}{a(a-1)} \ln \sum_{j=1}^{r} \mu_j^a \nu_j^{1-a},$$

coinciding with them at $a = 0$ and $a = 1$. The special case $R_{1/2}(\mu, \nu)/4$ is known as the Bhattacharyya distance. The sum $M(a) = \sum \mu_j^a \nu_j^{1-a}$ is the moment generating function of log likelihood ratio of μ and ν under the hypothesis ν.

The Renyi distances play an important role in information theory and in the statistics of random processes. All results obtained in this paper for D_a can easily be reformulated for R_a (the corresponding minimum distance estimators coincide).

In accordance with Cressie and Read [7] we use the modified parameter $\lambda = a - 1$; i.e., we consider the tests $(T^{(\lambda)}, \chi^2_{r-1}(1-\alpha))$ of Section 3, where $0 < \alpha < 1$, $\lambda \in R$, and $T^{(\lambda)}$ is defined by (33) with ϕ replaced by $\phi_{1+\lambda}$ and $\pi(\hat{\theta})$ replaced by π_0. In other words, we substitute in (33) $\phi(1) = 0$, $\phi''(1) = 1$, and

$$D_\phi(\hat{\pi}, \pi(\hat{\theta})) = D_{1+\lambda}(\hat{\pi}, \pi_0).$$

Therefore,

$$T^{(\lambda)} = \frac{2nN}{c_n} D_{1+\lambda}(\hat{\pi}, \pi_0) \tag{34}$$

for $c_n = n(n+K)/(K+1)$. Some well-known particular cases of statistics $ND_{1+\lambda}(\hat{\pi}, \pi_0)$ are presented in Table I, which is also helpful for orientation in the results that follow.

First we present a computer-based analysis of optimality in the class of tests

$$(T^{(\lambda)}, \chi^2_{r-1}(1-\alpha)), \qquad \lambda \in R. \tag{35}$$

This analysis consists in the computation of the actual size and power of these tests. The computations were restricted to the symmetric hypothesis H with the normed expectations $\pi_0 = (r^{-1}, ..., r^{-1})$ and to alternatives H_δ for $\delta \in (-1, r-1)$ with the normed expectations

$$\pi(\delta) = \pi_0 + \left(-\frac{\delta}{r(r-1)}, ..., -\frac{\delta}{r(r-1)}, \frac{\delta}{r} \right).$$

The computed quantity is the probability

$$P_{\delta,\lambda} = \Pr\left(\frac{2nN}{c_n} D_{1+\lambda}(\hat{\pi}, \pi_0) > \chi^2_{r-1}(1-\alpha) \,|\, H_\delta \right)$$

for

$$\hat{\pi} = \frac{1}{nN} \sum_{i=1}^{N} Y^{(i)},$$

where $Y^{(1)}, ..., Y^{(N)}$ are i.i.d. realizations of Y with

$$\Pr(Y = y \,|\, H_\delta) = \binom{n}{y_1 \cdots y_r} \frac{\binom{K-1}{K\pi_1(\delta)-1 \cdots K\pi_r(\delta)-1}}{\binom{n+K-1}{y_1+K\pi_1(\delta)-1 \cdots y_r+K\pi_r(\delta)-1}}$$

TABLE I

Statistics $ND_{1+\lambda}(\hat{\pi}, \pi_0) = ND_a(\hat{\pi}, \pi_0)$

λ	a	symbol	name
1	2	X^2	Pearson's X^2
0	1	G^2	Log likelihood ratio
$-\frac{1}{2}$	$\frac{1}{2}$	T^2	Freeman–Tukey
-1	0	MG^2	Modified log likelihood ratio
-2	-1	MX^2	Neyman modified X^2

for every $y=(y_1, ..., y_r) \in S_{n,r}$. The value $P_{0,\lambda}$ is thus the actual size, and $P_{\delta,\lambda}$ for $\delta \neq 0$ is the actual power of the test (35) applied in the model of Example 3.

Our computations have been carried out for $\alpha = 0.05$ and $r=4$, also considered by Cressie and Read. Further, we have chosen $N=5$ and $n=4$ in order to obtain the same overall sample size, $nN=20$, as considered by these authors. For K tending to infinity, the model of Example 3 tends to the r-nomial with parameters $\pi_1, ..., \pi_r$. Therefore the results of our computations for large K must coincide with those of Cressie and Read. The other extreme, namely $K=0$, leads again to the results of the type of Cressie and Read, but with the reduced sample size $N=5$. We decided for K "in the midle" between these two extremes, namely for $K=5$. The results of our computations are presented in Table II. The computation error is less than 0.01.

The results of Table II seem to be more interesting from the point of view of practical small-sample applications of the tests (35) than analogous results of Cressie and Read:

(a) The actual size of the tests (35) differs too much from the designed size $\alpha=0.05$ for λ outside the interval $-1 \leqslant \lambda \leqslant 2$. Inside this interval one can take the power as the criterion of test optimality

(b) The monotonicity of power in λ found by Cressie and Read is not observed here. Instead, for $\delta=-0.9$ we see a dramatic peak of power in the neighborhood of $\lambda=-\frac{1}{2}$.

TABLE II

Probability $P_{\delta,\lambda}$ for $N=K=5$ and $n=r=4$

	δ			
λ	1.5	0.5	0	−0.9
−5	0.83	0.57	0.48	0.39
−2	0.56	0.25	0.18	0.25
−1	0.72	0.16	0.08	0.22
$-\frac{1}{2}$	0.71	0.16	0.08	0.72
−0.3	0.69	0.15	0.08	0.71
0	0.70	0.14	0.07	0.65
0.3	0.69	0.11	0.05	0.36
$\frac{1}{2}$	0.71	0.11	0.05	0.29
1	0.70	0.11	0.04	0.22
2	0.78	0.16	0.07	0.22
5	0.89	0.31	0.19	0.51

Note. The test size for $\delta=0$; the test power for $\delta \neq 0$.

(c) All classical goodness-of-fit tests, defined by λ equal 1, 0, -1, and -2 (cf. Table I), are problematic from the point of view of the power. The celebrated Pearson X^2-test defined by $\lambda = 1$ seems to be most problematic.

(d) The least problematic (optimal in a mimimax sense) seems to be the Freeman–Tukey T^2-test defined by $\lambda = -\frac{1}{2}$.

One can also conclude from Table II, and to some extent also from the analogous table of Cressie and Read, that (i) the optimality of tests $(T^{(\phi)}, \chi^2_{r-1}(1-\alpha))$, $\phi \in \Phi$, for finite N significantly depends on the choice of ϕ, and (ii) there is no hope of finding a test universally best for a greater variety of statistical models. Such a test cannot be chosen even among the tests (35) for the relatively narrow class of Dirichlet-multinomial models. Therefore positive conclusions concerning some tests in Cressie and Read [7] or here are valid only within the framework of the considered assumptions and one has to be very careful in extrapolating these conclusions beyond this framework.

Our analytic method of evaluation of optimal $\lambda \in R$ for the tests (35) is based on the assumption that a test is better, the closer $T^{(\lambda)}$ is to χ^2_{r-1}. The peculiarity of this approach to the test optimality is further sharpened by the next step. Namely, aiming at compatibility with the approach of Cressie and Read, we measure the distance between $T^{(\lambda)}$ and χ^2_{r-1} by the absolute deviation of expectations. Thus by the optimization of λ we mean the minimization of

$$|E(T^{(\lambda)}) - E(\chi^2_{r-1})|.$$

Next we show that, for large N, zero deviation is achieved in the neighborhood of two points λ_1 and λ_2, where $\lambda_1 = 1$ is constant and λ_2 depends on the hypothesis parameters π_0 and model parameters (n, r, K).

By Taylor expansion of $D_{1+\lambda}(\hat{\pi}, \pi_0)$ around $\hat{\pi} = \pi_0$ we get

$$D_{1+\lambda}(\hat{\pi}, \pi_0) = \frac{1}{2}\sum_{j=1}^{r} \pi_{0j}^{-1} W_j^2 + \frac{\lambda - 1}{6}\sum_{j=1}^{r} \pi_{0j}^{-2} W_j^3 + \frac{(\lambda-1)(\lambda-2)}{24}\sum_{j=1}^{r} \pi_{0j}^{-3} W_j^4 + O_p(N^{-5/2}),$$

where

$$W_j = \hat{\pi}_j - \pi_{0j} = \frac{1}{nN}\sum_{i=1}^{N} (Y_j^{(i)} - E(Y_j^{(i)})).$$

By (6b) and (7),

$$E(W_j^2) = \frac{c_n}{nN}\pi_{0j}(1-\pi_{0j}),$$

so that

$$E\left(\frac{1}{2}\sum_{j=1}^{r}\pi_{0j}^{-1}W_j^2\right) = \frac{(r-1)\,c_n}{2nN} = \frac{c_n E(\chi_{r-1}^2)}{2nN}.$$

Therefore, $E(T^{(\lambda)}) - E(\chi_{r-1}^2)$ is equal to

$$\frac{(\lambda-1)\,nN}{3c_n}\sum_{j=1}^{r}\pi_{0j}^{-2}E(W_j^3) + \frac{(\lambda-1)(\lambda-2)}{12c_n}Nn\sum_{j=1}^{r}\pi_{0j}^{-3}E(W_j^4) + O(N^{-3/2}).$$

By the definition of W_j above

$$E(W_j^3) = n^{-3}N^{-2}[E(Y_j^3) - 3n\pi_{0j}E(Y_j^2) + 2n\pi_{0j}^3]$$

and

$$E(W_j^4) = n^{-4}N^{-3}[E(Y_j^4) - 4n\pi_{0j}E(Y_j^3) + 3(N-1)(E(Y_j^2))^2 + 6\pi_{0j}n^2(2-N)\,E(Y_j^2) + 3n^4(N-2)\,\pi_{0j}^4].$$

Neglecting the terms of order $O(N^{-3})$ one obtains from here

$$E(T^{(\lambda)}) - E(\chi_{r-1}^2) = N^{-1}a(\lambda, r, n, K) + O(N^{-3/2}),$$

where

$$a = a(\lambda, r, n, K) = \frac{\lambda-1}{3n^2c_n}\left(b + \frac{\lambda-2}{4n}c\right)$$

for

$$b = \sum_{j=1}^{r}\pi_{0j}^{-2}E(Y_j^3) - 3n\sum_{j=1}^{r}\pi_{0j}^{-1}E(Y_j^2) + 2n \tag{36}$$

and

$$c = 3\sum_{j=1}^{r}\pi_{0j}^{-3}(E(Y_j^2))^2 - 6n^2\sum_{j=1}^{r}\pi_{0j}^{-1}E(Y_j^2) + 3n^4. \tag{37}$$

The expectations figuring in (36) and (37) are explicitly given as

$$E(Y_j^2) = \frac{\pi_{0j}^2 n(n-1)K + \pi_{0j} n(n+K)}{K+1},$$

$$E(Y_j^3) = \frac{\begin{pmatrix} \pi_{0j}^3 n(n^2-3n)+2)K^2 + K3\pi_{0j}^2 n(n^2+n[K-1)-K] \\ +\pi_{0j} n(2n^2+3Kn+K^2) \end{pmatrix}}{(K+1)(K+2)}.$$

Thus our result is that in the classification model of Example 3 the optimal large-sample tests (35) are obtained for

$$\lambda_1 = 1 \qquad \text{and} \qquad \lambda_2 = 2 - \frac{4nb}{c}, \tag{38}$$

where b, c are given by (36) and (37).

We see that the value λ_1 defines the classical Pearson's X^2-test. The other test which is optimal in the stated sense is defined by λ_2, ranging over the whole real line when n, r, K, and π_0 vary arbitrarily in their respective domains. One can expect that in large-sample applications of tests (35), the size of optimal tests $(T^{(\lambda_2)}, \chi^2_{r-1}(1-\alpha))$ and $(T^{(1)}, \chi^2_{r-1}(1-\alpha))$ will be best fitted to the designed value $0 < \alpha < 1$. Table II indicates that the actual size of tests (35) may depart very significantly from α so that the present optimality is statistically meaningful. This table also implies that this optimality is lost when the sample size becomes small. Indeed, one obtains from (36)–(38) that $\lambda_2 = 2.22$ for $(n, r, k) = (4, 4, 5)$ and for the uniform π_0 considered in Table II, but the actual sizes of $(T^{(1)}, \chi^2_4(0.95))$ and $(T^{(2.22)}, \chi^2_4(0,95))$, deductible from Table II are not fitted to $\alpha = 0.05$.

The powers of tests $(T^{(\lambda_2)}, \chi^2_{r-1}(0.95))$ and $(T^{(1)}, \chi^2_{r-1}(0.95))$ need not be the same as those demonstrated to some extent by Table II too. The preference between them in practical applications can be decided by calculating an analogue of the corresponding two rows in Table II. The choice of alternatives H_δ used in these calculations may, of course, be not as simple as in Table II, which is dealing with the special symmetric hypothesis H.

Let us now look in more detail at the particular case of the symmetric hypothesis H with $\pi_0 = (r^{-1}, \ldots, r^{-1})$. Then

$$E(Y_j^2) = \frac{n(n-1)K + rn(n+K)}{r^2(K+1)}$$

and

$$E(Y_j^3) = \frac{\begin{pmatrix} n(n^2-3n+2)K^2 + 3rnK[n^2+n(K-1)-K] \\ +r^2 n(2n^2+3nK+K^2) \end{pmatrix}}{r^3(K+1)(K+2)}$$

so that, by (36) and (37),

$$b(K+1)(K+2) = n(n^2-3n+2)\,K^2 + 3rn[n^2+n(K-1)-K] + r^2n(2n^2+3nK+K^2) - 3n[n(n-1)\,K+rn(n+K)] \times (K+2) + 2n(K+1)(K+2)$$

and

$$c(K+1)^2 = 3[n(n-1)\,K+rn(n+K)]^2 - 6n^2[n(n-1)\,K+rn(n+K)] + 3n^4(Kn)^2.$$

Therefore,

$$\lim_{r\to\infty}\frac{b}{c} = \frac{n(2n^2+3nK+K^2)}{3n^2(n+K)^2}\,\frac{(K+1)^2}{(K+1)(K+2)} = \frac{(2n^2+3nK+K^2)(K+1)}{3n(n+K)^2\,(K+2)}$$

and, by (38),

$$\lim_{r\to\infty}\lambda_2 = 2-\frac{4}{3}\,\frac{(2n^2+3nK+K^2)(K+1)}{(n+K)^2\,(K+2)}. \tag{39}$$

Thus for large r the explicit specification of the optimal test $(T^{(\lambda_2)}, \chi^2_{r-1}(1-\alpha))$ is simple and easy.

For $K\to\infty$ (the case considered in Section 4 of Cressie and Read [7], with the overall sample size nN), as well as for $n=1$ and arbitrary $K>0$ (the same case, but with the overall sample size N), we obtain in (39) the limit value

$$\lambda_2 = 2-\tfrac{4}{3} = \tfrac{2}{3} \tag{40}$$

which coincides with the limit value found already by Cressie and Read. Hence the large-sample results of the present section are consistent with an extend those of Cressie and Read [7]. In particular, it remains true for the classification models under consideration that the test $(T^{(2/3)}, \chi^2_{r-1}(1-\alpha))$ is a practically interesting alternative to the Pearson's X^2-type test $(T^{(1)}, \chi^2_{r-1}(1-\alpha))$, provided the normed expectations in hypothesis (14) are close to $(r^{-1}, \ldots, r^{-1})$ and r is large.

REFERENCES

[1] ALI, M. S., AND SILVEY, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** 131–140.

[2] ALTHAM, P. M. E. (1976). Discrete variable analysis for individuals grouped into families. *Biometrika* **63** 263–129.

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

[4] BISHOP, Y. M. M., FIENBERG, S. E., AND HOLLAND, P. W. (1975). Discrete Multivariate Analysis: Theory and Practice, MIT Press, Cambridge, MA.

[5] BRIER, S. S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika* **67** 591–596.

[6] COHEN, J. E. (1976). The distribution of the chi-squared statistic under clustered sampling from contigency tables. *J. Am. Statist. Assoc.* **71** 665–670.

[7] CRESSIE, N., AND READ, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46** 440–464.

[8] CSISZAR, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizit = 84t von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci. Ser A* **8** 85–108.

[9] LIESE, F., AND VAJDA, I. (1987). *Convex Statistical Distances*. Teubner, Leipzig.

[10] LINDSAY, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Statist.* **22** 1081–1114.

[11] MCCULLAGH, P., AND J. A. NELDER (1989). *Generalized Linear Models*. Chapman & Hall, London.

[12] MORALES, D., PARDO, L., AND VAJDA, I. (1995). Asymptotic divergence of estimates of discrete distributions. *J. Statist. Planning Inference*, in print.

[13] MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution and correlation among proportions. *Biometrika* **49** 65–82.

[14] READ, T. R. C., AND CRESSIE, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, Berlin.

[15] SALICRÚ, M., MENÉNDEZ, M. L., MORALES, D., AND PARDO, L. (1994). On the applications of divergence type measures in testing statistical hypotheses. *J. Multivariate Anal.* **51** 372–391.

[16] VAJDA, I. (1988). *Theory of Statistical Inference and Information*. Kluwer, Dordrecht.

[17] RENYI (1961). On the measures of entropy and information. *Proc. 4th Berkeley Symp. Math. Statist. and Prob.* **1** 547–561.