KYBERNETIKA — VOLUME 35 (1999), NUMBER 3, PAGES 265 - 280

INFERENCE ABOUT STATIONARY DISTRIBUTIONS OF MARKOV CHAINS BASED ON DIVERGENCES WITH OBSERVED FREQUENCIES*

MARÍA LUISA MENÉNDEZ, DOMINGO MORALES, LEANDRO PARDO AND IGOR VAJDA

For data generated by stationary Markov chains there are considered estimates of chain parameters minimizing ϕ–divergences between theoretical and empirical distributions of states. Consistency and asymptotic normality are established and the asymptotic covariance matrices are evaluated. Testing of hypotheses about the stationary distributions based on ϕ–divergences between the estimated and empirical distributions is considered as well. Asymptotic distributions of ϕ–divergence test statistics are found, enabling to specify asymptotically α-level tests.

1. INTRODUCTION

Methods of statistical inference established for stationary independent data are often applied to dependent data. The effect of dependence on the Pearson goodness of fit tests using the Pearson statistics has been studied by Moore [11] and Glesser and Moore [6, 7]. Tavaré and Altham [15] evaluated for stationary Markov observations, under simple hypotheses about the state space distributions, asymptotic distribution of the corresponding Pearson statistic X^2. Moore [11] evaluated the asymptotic distribution of the maximum likelihood and minimum chi–square estimators of parameters of discrete distributions defined by a quantization in the state space of some stationary stochastic processes. Glesser and Moore [6, 7] evaluated for "positively dependent" observations, and for maximum likelihood estimators of parameters, asymptotic distribution of Pearson X^2 in the case where the hypotheses about the state space distribution are composite. They also mentioned possible extensions of their results to the Pearson–type statistic obtained as special ϕ–divergences (the so–called power divergences) between the estimated and empirical distributions. These divergences have been previously studied in the case of independence observations by Cressie and Read [4] (cf. also Read and Cressie [13],

*This work was supported by the DGICYT grant PB 96–0635, by grant 1075709 of the Academy of Sciences of the Czech Republic, and by grant 102/99/1137 of Grant Agency of the Czech Republic.

Salicrú et al [14] and Menéndez et al [9]). In Menéndez et al [10], we applied the ϕ–divergences in testing simple hypotheses about stationary irreducible aperiodic Markov chains. In this manner we extended the results of Tavaré and Altham to an infinite variety of ϕ–divergence goodness-of-fit test statistics. We also proposed a method for choice a best ϕ–divergence test statistic and numerically illustrated it by an example.

In this paper we study simple as well as composite hypotheses about irreducible aperiodic Markov observations. For arbitrary regular convex functions ϕ and ϕ^* we evaluate asymptotic distributions of the minimum ϕ^*–divergence estimator, and of the ϕ–divergence statistic employing the minimum ϕ^*–divergence estimator if the hypothesis is composite. This paper thus significantly extends the previous results of Menéndez et al [10], and precises and in some sense also extends the ideas of Glesser and Moore [6, 7].

2. BASIC CONCEPTS AND EXAMPLES

We consider a stationary irreducible aperiodic Markov chain $\mathcal{X} = (X_0, X_1, \ldots)$ with the state space $\{1, \ldots, m\}$. By $P = (p_{ij})_{i,j=1}^m$ we denote the matrix of transition probabilities of this chain and by $p = (p_1, \ldots, p_m)$ a stationary distribution, i. e. solution of the equation $p = pP$. Thus the Markov chains under consideration are described by pairs $\langle p, P \rangle$.

Assumption 1. P is from the class $\boldsymbol{P}$ of all irreducible aperiodic stochastic matrices with one ergodic class.

The aperiodicity and ergodicity imply the existence and unicity of the solution of equation $p = pP$. The irreducibility means that the solution p belongs to the set

$$\Pi_m = \{(p_1, \ldots, p_m) : p_i > 0,\ p_1 + \cdots + p_m = 1\}$$

which is an open subset of a hyperplane in R^m.

Assumption 2. On an open subset $\Theta \subset R^s$, there is given a continuous invertible mapping

$$\theta \mapsto p(\theta) = (p_1(\theta), \ldots, p_m(\theta)) \in \Pi_m$$

with a continuous inverse $p \mapsto \theta(p) \in \Theta$.

Under this Assumption, $p(\theta)$ and $\theta(p)$ are one-to-one mappings between Θ and an open subset $\Pi \subset \Pi_m$.

Assumption 3. The stationary distribution p belongs to Π considered in Assumption 2.

The set Π represents a *basic hypothesis* about the distribution p, Θ is a *parameter space* of distributions belonging to Π, and $\theta(p) \in \Theta$ is a *parameter* corresponding to $p \in \Pi$.

For every parameter $\theta \in \Theta$ we denote by $\boldsymbol{P}_\theta$ the set of all matrices $P \in \boldsymbol{P}$ such that their stationary distribution p coincides with $p(\theta)$. If $p(\theta)$ is uniform then $\boldsymbol{P}_\theta$ is the class of all doubly stochastic $m \times m$ matrices.

Example 1. Let $s = m-1$, $\Theta = \{\theta = (\theta_1, \ldots, \theta_{m-1}) \in (0,1)^{m-1} : \theta_1 + \cdots + \theta_{m-1} < 1\}$ and $p(\theta) = (\theta_1, \ldots, \theta_{m-1}, 1 - \sum_{i=1}^{m-1} \theta_i)$. Then $\Pi = \Pi_m$ and the parameters $\theta(p)$ of distributions $p \in \Pi_m$ are their first $m-1$ coordinates $p_1, \ldots, p_{m-1}$. In the particular case of $m = 2$ we obtain $\Theta = (0,1)$ and $\Pi_2 = \{(\theta, 1-\theta) : \theta \in (0,1)\}$. Here $\boldsymbol{P}$ is the set of all matrices

$$P = \begin{pmatrix} 1-\beta & \beta \\ \gamma & 1-\gamma \end{pmatrix} \quad \text{for} \quad 0 < \beta,\ \gamma \leq 1 \text{ and } \beta + \gamma < 2,$$

with the stationary distributions $p = (p_1, p_2) = (\theta, 1-\theta)$ given by the formula

$$\theta = \frac{\gamma}{\beta + \gamma}.$$

Therefore $\boldsymbol{P}_\theta$ is the set of all matrices

$$\begin{pmatrix} 1-\beta & \beta \\ \frac{\theta\beta}{1-\theta} & 1 - \frac{\theta\beta}{1-\theta} \end{pmatrix} \quad \text{for} \quad 0 < \beta \leq \min\left\{1, \frac{1-\theta}{\theta}\right\}, \quad \beta \neq 1.$$

This means that for every fixed $0 < \beta < 1$ these matrices belong to $\boldsymbol{P}_\theta$ for all $0 < \theta \leq \frac{1}{1+\beta}$. In particular, $\boldsymbol{P}_{\frac{1}{2}}$ is the set of all matrices

$$\begin{pmatrix} 1-\beta & \beta \\ \beta & 1-\beta \end{pmatrix} \quad \text{for} \quad 0 < \beta < 1.$$

Example 2. Let $s = m-1$ and $\Theta = \{\theta = (\theta_1, \ldots, \theta_{m-1}) : \theta_i \in (0,1), 1 \leq i \leq m-1\}$, and let $p(\theta) = (p_1(\theta), \ldots, p_m(\theta))$ be given for every $\theta \in \Theta$ by

$$p_1 = \frac{1}{1 + \theta_1 + \theta_1\theta_2 + \cdots + \theta_1 \cdots \theta_{m-1}}, \quad p_i = \theta_1 \cdots \theta_{i-1}\, p_1 \quad \text{for} \quad 1 \leq i \leq m.$$

Here Π is an $(m-1)$–dimensional variety in Π_m and $\theta(p_1, \ldots, p_m) = (p_2/p_1, \ldots$ $\ldots, p_m/p_{m-1})$. One of the matrices contained in $\boldsymbol{P}_\theta$ is $P(\theta) = (p_{ij})$ with $p_{m,1} = 1$ and

$$p_{i,1} = 1 - \theta_i, \qquad p_{i,i+1} = \theta_i, \qquad \text{for} \quad 1 \leq i \leq m-1.$$

Under Assumptions 1–3 and the basic hypothesis Π, the true stationary distribution of chain states is some $p_0 = (p_{01}, \ldots, p_{0m}) \in \Pi$. This means that the true chain parameter is $\theta_0 = \theta(p_0)$ from Θ.

Assumption 4. The true chain distribution is specified by an arbitrary initial distribution $p(\theta_0)$ and by a transition matrix $P(\theta_0) \in \boldsymbol{P}_{\theta_0}$.

A basic statistical problem is how to estimate in a consistent and asymptotically normal way the unknown true parameter $\theta_0 \in \Theta$ by using the data $S_n = (X_1, \ldots, X_n)$ about the states of the chain, i.e. how to find a measurable mapping

$$\widehat{\theta}_n = \widehat{\theta}_n(S_n) \tag{1}$$

taking on values in Θ such that

$$\begin{aligned} \widehat{\theta}_n &\to \theta_0 && \text{in probability} \\ n^{1/2}(\widehat{\theta}_n - \theta_0) &\to N(0, V_0) && \text{in law,} \end{aligned} \tag{2}$$

and how to evaluate the $s \times s$ matrix V_0 (note that all convergences in this paper are considered for $n \to \infty$).

Another important statistical problem is how to test a hypothesis about θ_0 by using the data S_n. The hypothesis may be represented by a subset $\Theta_0 \subset \Theta$ or, equivalently, by $\Pi_0 = \{p(\theta) : \theta \in \Theta_0\} \subset \Pi$. The alternative is $\Theta_1 = \Theta - \Theta_0$ or $\Pi_1 = \Pi - \Pi_0$. The problem is to find a measurable test statistic and a measurable critical region in the target space of this statistics,

$$T_n = T_n(S_n) \quad \text{and} \quad K_{n,\alpha} \quad \text{for} \quad 0 < \alpha < 1, \tag{3}$$

such that the tests $(T_n, K_{n,\alpha})$ are asymptotically of α-size in the sense

$$\Pr\{T_n \in K_{n,\alpha} | P(\theta_0)\} \to \alpha \quad \text{for all } \theta_0 \in \Theta_0. \tag{4}$$

Preferences between various tests satisfying (4) are usually based on the power functions

$$\pi_n(\theta) = \Pr\{T_n \in K_{n,\alpha} | P(\theta)\} \quad \text{for} \quad \theta \in \Theta_1. \tag{5}$$

Most preferred are those with a maximum test power where the "test power" means an asymptotic or nonasymptotic variant of the power function (5).

Both these problems, of estimation and testing, are solved in this paper. The solution is based on relative frequencies observed in the data S_n,

$$\widehat{p}_n = \left(\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{I}_{(1)}(X_k), \ldots, \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{I}_{(m)}(X_k) \right), \tag{6}$$

i.e., it in fact uses the ordered version of S_n and ignores the information about transitions contained in the original statistics S_n. This means a considerable loss of efficiency on the one hand, but also a considerable relative simplicity on the other hand.

By the strong law of large numbers, $\widehat{p}_n = (\widehat{p}_{n1}, \ldots, \widehat{p}_{nm})$ may be assumed to belong to the same open set Π_m as $p(\theta_0)$. We show that there exist an estimator (1)

satisfying (2) and a test (3) satisfying (4), both based on the ϕ-divergences

$$D_\phi(\widehat{p}_n,\, p(\theta)) = \sum_{i=1}^{m} p_i(\theta)\, \phi\left(\frac{\widehat{p}_{ni}}{p_i(\theta)}\right) \tag{7}$$

of stationary distributions $p(\theta)$ with the observed frequencies $\widehat{p}_n$.

The ϕ–divergences of probability distributions specified by convex functions $\phi\colon (0,\infty) \mapsto R$ have been used in the statistics by many authors, see the references in Liese and Vajda [8] and Read and Cressie [13]. Properties of ϕ-divergences were systematically studied in Liese and Vajda [8], where we refer for the details.

Our estimator $\widehat{\theta}_n = \widehat{\theta}_n^{(\phi)}$ minimizes the ϕ-divergence (7) over Θ, i.e.

$$\widehat{\theta}_n = \operatorname{argmin} D_\phi(\widehat{p}_n,\, p(\theta)). \tag{8}$$

For the particular function $\phi_*(t) = t \ln t$, $\widehat{\theta}_n^{(\phi_*)}$ is the partial maximum likelihood estimator (partial MLE), where "partial" means that it is using only the partial information contained in the ordered version of S_n. If the data are independent then it becomes to be the standard MLE. In the model of Example 1 with $p(\theta) \equiv \theta$ for all $\theta \in \Theta$, we obtain $\widehat{\theta}_n^{(\phi)} = \widehat{p}_n$ for any ϕ.

Our test statistics $T_n = T_n^{(\phi,\phi_*)}$ are defined for arbitrary convex ϕ, ϕ_*, with ϕ twice continuously differentiable in an open neighbourhood of 1, $\phi(1) = 0$ and $\phi''(1) \neq 0$, by

$$T_n = 2n\phi''(1)^{-1}\, D_\phi(\widehat{p}_n,\, p(\widehat{\theta}_n^{(\phi_*)})). \tag{9}$$

Here, obviously, $\widehat{\theta}_n^{(\phi_*)}$ is defined by (8) with ϕ replaced by ϕ_*. Sometimes it is convenient to employ this estimator in the version with the minimization in (8) restricted to the null space Θ_0. Then, if the hypothesis is simple, i.e. $\Theta_0 = \{\theta_0\}$, (9) reduces to

$$T_n = 2n\phi''(1)^{-1}\, D_\phi\,(\widehat{p}_n,\, p(\theta_0))\,. \tag{10}$$

For example, if $\phi(t) = (t-1)^2$ then (10) is the Pearson statistic

$$X^2\,(\widehat{p}_n,\, p(\theta_0)) = n \sum_{i=1}^{m} \frac{(\widehat{p}_{ni} - p_i(\theta_0))^2}{p_i(\theta_0)}$$

and (9) the Pearson statistic with $\widehat{\theta}_n^{(\phi_*)}$ plugged-in for the unknown θ_0.

Various particular cases of the mentioned ϕ-divergence estimators and ϕ-divergence tests have been extensively used in the literature dealing with discrete independent observations, in particular with testing hypotheses concerning such observations, cf. Read and Cressie [13], Salicrú et al [14], Morales et al [12], and further references therein. Versions important from our point of view, applicable to positive recurrent Markov chains, have been considered by Tavaré and Altham [15]. These authors solved among others the testing problem under consideration for the simple hypothesis $\Pi_0 = \{p_0\}$ by using the Pearson test statistic $T_n = X^2(\widehat{p}_n, p_0)$. Using

known facts about asymptotic distributions of irreducible aperiodic Markov chains, they found that for every model under consideration and every $\theta_0 \in \Theta$, the statistic $T_n = X^2(\widehat{p}_n, p(\theta_0))$ satisfies the asymptotic relation

$$T_n \to \sum_{i=1}^{m} \rho_i Z_i^2 \quad \text{in law}, \tag{11}$$

where Z_i are independent $N(0,1)$ and ρ_i are the eigenvalues of the matrix $D_0^{-1}\Omega_0$ for $D_0 = \operatorname{diag} p(\theta_0)$, (i.e. $d_{ii} = p_i(\theta_0)$ and $d_{ij} = 0$ for $i \neq j$),

$$\Omega_0 = D_0 C_0 + C_0^t D_0 - D_0 - p(\theta_0)^t p(\theta_0), \quad C_0 = \left(I - P(\theta_0) + \mathbf{1}^t p(\theta_0)\right)^{-1}$$

(here, obviously, I is the identity $m \times m$ matrix, i.e. $I = \operatorname{diag} \mathbf{1}$ where $\mathbf{1}$ is the row vector of m units).

In the present paper we are interested in the validity of (11) for more general test statistics (9) and (10). A generalization of (11) will lead us to the family of asymptotically α-level tests

$$\mathcal{T} = \{(T_n, K_{n,\alpha} = K_\alpha = (Q_\alpha(\rho_1, \ldots, \rho_m), \infty)) : \phi \in \Phi\}, \tag{12}$$

where T_n are the statistics (9) or (10) and the critical region $K_{n,\alpha} = K_\alpha$ is the interval (Q_α, ∞) for the $(1-\alpha)$-quantile $Q_\alpha(\rho_1, \ldots, \rho_m)$ of the random variable $\sum_{i=1}^m \rho_i Z_i^2$. In the sequel, the tests (T_n, K_α) figuring in (12) will be explicitly indexed by the elements ϕ of the class Φ of convex functions considered there.

3. TESTING SIMPLE HYPOTHESES

In Menéndez et al [10] we studied the simple hypothesis, i.e. the case $\Theta_0 = \{\theta_0\}$. We obtained the following extension of the Tavaré and Altham [15] version of (11). This extension also exploits the possibility of simpler evaluation of parameters ρ_i figuring in (11) for reversible chains. Remind that a chain P under consideration is said to be reversible if the probability of inverse transition q_{ij} (i.e. the conditional probability that the previous state was j given that the present state is i, formally $p_j p_{ji}/p_i$) coincides with the probability of ordinary transition p_{ij}, i.e. if for every $1 \leq i, j \leq m$

$$p_j p_{ji} = p_i p_{ij}$$

or, equivalently, $D P = P^t D$.

Theorem 1. (Menéndez et al [10]) Let Assumptions 1–4 hold. Then relation (11) takes place for all statistics (10). If the chain $P(\theta_0)$ is reversible then the parameters ρ_i in (11) are given by the formula

$$\rho_i = \frac{1+\lambda_i}{1-\lambda_i} \quad \text{for} \quad 1 \leq i \leq m-1, \quad \rho_m = 0, \tag{13}$$

where $\lambda_1, \ldots, \lambda_{m-1}$ are the non-unit eigenvalues of $P(\theta_0)$.

This result was obtained in [10] by proving that for every ϕ under consideration the statistic (10) is expansible as follows

$$T_n = X^2(\widehat{p}_n, p(\theta_0))(1 + o_p(1)), \tag{14}$$

and by a subsequent application of the mentioned special result of Tavaré and Altham [15]. Let us briefly mention some consequences useful in the sequel.

Corollary 1. If P has identical rows (i.e. $P = \mathbf{1}^t p$ where $p = (p_1, \ldots, p_m)$ is a stochastic vector) then it is reversible and all its nonunit eigenvalues are zero. Thus Theorem 1 implies that if the data $X_1, \ldots, X_n$ are independent then all statistics (10) are asymptotically χ^2-distributed with $m - 1$ degrees of freedom (in symbols, χ^2_{m-1}). More generally, if $P = (1 - \pi)I + \pi \mathbf{1}^t p$ where $0 < \pi \leq 1$ then the nonunit eigenvalues of P are all equal to $1 - \pi$. Therefore all statistics (10) tend in law to $\chi^2_{m-1}(2 - \pi)/\pi$.

Remark 1. Using Theorem 1 one can argue that (12) with the statistics T_n defined by (10) is a family of asymptotically α-level tests. This is true however only if the matrix $P(\theta_0) \in \boldsymbol{P}_{\theta_0}$ is known, i.e. only if the eigenvalues $\rho_1, \ldots, \rho_m$ needed to specify the critical value Q_α are available. If this assumption is not satisfied then one can use the relative frequencies

$$\widehat{p}_{nij} = \frac{\sum_{k=2}^n \boldsymbol{I}_{(i,j)}(X_{k-1}, X_k)}{\sum_{k=2}^n \boldsymbol{I}_{(i)}(X_{k-1})}$$

as consistent estimates of elements $p_{ij}(\theta_0)$ of the matrix $P(\theta_0)$ (cf. Billingsley [1]). Since the eigenvalues $\rho_1, \ldots, \rho_m$ are continuous functions of elements of the matrix P, the eigenvalues $\rho_{n1}, \ldots, \rho_{nm}$ obtained by replacing $p_{ij}(\theta_0)$ by $\widehat{p}_{nij}$ are consistent estimates of the unknown values $\rho_1, \ldots, \rho_m$. This together with Theorem 1 implies the following fact.

Corollary 2. All tests in the family (12) with T_n given by (10) and $Q_{n,\alpha} = Q_\alpha(\rho_{n1}, \ldots, \rho_{nm})$ are asymptotically α-level tests of the simple hypothesis $\{\theta_0\}$ under consideration.

Corollary 2 provides a variety of tests. In Menéndez et al [10] we considered a class of ϕ_a–divergence tests, using the ϕ_a–statistics for functions

$$\phi_a(t) = \frac{t^a - 1}{a(a-1)} \qquad \text{for} \quad a \neq 0, a \neq 1, \tag{15}$$

leading to the Hellinger–type divergences

$$D_a(\widehat{p}, p) = \frac{\sum_{i=1}^m \widehat{p}_i^a p_i^{1-a} - 1}{a(a-1)} \qquad \text{for} \quad a \neq 0, a \neq 1.$$

The limits

$$D_1(\hat{p}, p) = \sum_{i=1}^{m} \hat{p}_i \ln \frac{\hat{p}_i}{p_i}$$

and

$$D_0(\hat{p}, p) = \sum_{i=1}^{m} p_i \ln \frac{p_i}{\hat{p}_i}$$

of these divergences for $a \uparrow 1$ and $a \downarrow 0$ are the ϕ_a–divergences for functions $\phi_1(t) = t \ln t$ and $\phi_0(t) = -\ln t$. From (10) one obtains in this manner the statistics

$$T_n^a = \frac{2n}{a(a-1)} \left(\sum_{i=1}^{m} \hat{p}_{ni}^a p_i(\theta_0)^{1-a} - 1 \right) \qquad \text{for} \quad a \neq 0, 1, \tag{16a}$$

$$T_n^1 = 2n \sum_{i=1}^{m} \hat{p}_{ni} \ln \frac{\hat{p}_{ni}}{p_i(\theta_0)}, \tag{16b}$$

$$T_n^0 = 2n \sum_{i=1}^{m} p_i(\theta_0) \ln \frac{p_i(\theta_0)}{\hat{p}_{ni}}. \tag{16c}$$

We see that T_n^1 and T_n^0 are the likelihood ratio statistics, sometimes called G^2 and modified G^2. T_n^2 and T_n^{-1} are the Pearson X^2 and Neyman modified X^2, and $T_n^{1/2}$ is a Freeman–Tukey statistic. Thus the class of statistics (16) for $-6 \leq a \leq 6$ seems to be rich and interesting enough to be able to represent all convex functions in the statistical experimentation under consideration. A similar restriction has been recommended by Drost et al [5] on the basis of power considerations in the case of independent observations.

In [10] we also suggested Monte Carlo approximations to the test powers and sizes

$$\pi_n(\theta, a) = \pi_n(\theta, \phi_a) = \Pr(T_n^a > Q_{n,\alpha} | P(\theta)) \tag{17}$$

for a from a reasonable interval around 0, by the relative frequencies $\pi_{n,M}(\theta, a)$ of the event $T_n^a > Q_{n,\alpha}$ in M independent realizations. We proposed a method of choice of a leading to a best test statistic T_n^a, based on these approximations.

In the following two sections we extend Theorem 1 to composite hypotheses $\Theta_0 \subset \Theta$. The statistics of our interest will be for example the members of family (9) obtained from (16) by replacing the true probabilities $p_i(\theta_0)$ by their estimates $p_i(\widehat{\theta}_n^{(\phi_*)})$, in particular by the estimates obtained for $\phi_* = \phi_{a_*}$ given by (15). To this end we need at the first place appropriate results concerning estimators $\widehat{\theta}_n^{(\phi)}$, $\phi \in \Phi$. Therefore we start in the next section with the estimation problem.

4. ESTIMATION

In this section we consider the minimum ϕ–divergence estimators $\widehat{\theta}_n = \widehat{\theta}_n^{(\phi)}$ defined by (8). If $\phi(t) = t \ln t$ then $\widehat{\theta}_n$ is the MLE discussed above. Let us introduce the following regularity conditions.

(A1) $p(\theta)$ is continuously differentiable in the neighbourhood of θ_0 and

$$(p(\theta) - p(\theta_0))^t = J_0(\theta - \theta_0)^t + o(\|\theta - \theta_0\|) \quad \text{for } \theta \to \theta_0,$$

where $J_0 = J(\theta_0)$ is the Jacobian defined by

$$J(\theta)^t = \left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_s}\right)^t p(\theta).$$

(A2) $A_0^t A_0$ is positive definite for

$$A_0 = \operatorname{diag}\left(p_1(\theta_0)^{-1/2}, \ldots, p_m(\theta_0)^{-1/2}\right) J_0.$$

Hereafter we consider the matrix

$$B_0 = \operatorname{diag} p(\theta_0)^{-1/2} \Omega_0 \operatorname{diag} p(\theta_0)^{-1/2},$$

where Ω_0, defined at the end of Section 2, is the asymptotic covariance matrix of the asymptotically normal zero mean random vector

$$\sqrt{n}\,(\widehat{p}_{n1} - p_1(\theta_0), \ldots, \widehat{p}_{nm} - p_m(\theta_0))$$

(for the asymptotic normality see Billingsley [1] or (2.2) in Tavaré and Altham [15]), and diag $p(\theta_0)^{-1/2}$ denotes the same diagonal matrix as in the formula for A_0 above. Put for brevity

$$\Delta_0 = A_0(A_0^t A_0)^{-1}, \qquad \Sigma_0 = \Delta_0 A_0^t = A_0(A_0^t A_0)^{-1} A_0^t.$$

The following theorem summarizes the properties of minimum ϕ-divergence estimators of parameters of stationary distributions of Markov chains. It extends similar results for the maximum likelihood and other estimators with independent observations in Birch [2], Bishop, Fienberg and Holland [3], Read and Cressie [13] and Morales et al [12].

Theorem 2. Let ϕ satisfy the assumptions considered in (9) and let (A1), (A2) hold. Then the minimum ϕ–divergence estimator $\widehat{\theta}_n$ satisfies the following asymptotic relations:

$$\widehat{\theta}_n \to \theta_0 \text{ a. s.}, \tag{18}$$

$$\widehat{\theta}_n = \theta_0 + (\widehat{p}_n - p(\theta_0)) \operatorname{diag} p(\theta_0)^{-1/2} \Delta_0 (1 + o_p(1)), \tag{19}$$

$$n^{1/2}(\widehat{\theta}_n - \theta_0) \to N\left(0, \Delta_0^t B_0 \Delta_0\right) \text{ in law}, \tag{20}$$

$$p(\widehat{\theta}_n) = p(\theta_0) + (\widehat{p}_n - p(\theta_0)) \operatorname{diag} p(\theta_0)^{-1/2} \Sigma_0 \operatorname{diag} p(\theta_0)^{1/2} (1 + o_p(1)), \tag{21}$$

$$n^{1/2}(p(\widehat{\theta}_n) - p(\theta_0)) \to N\left(0, \operatorname{diag} p(\theta_0)^{1/2} \Sigma_0 B_0 \Sigma_0 \operatorname{diag} p(\theta_0)^{1/2}\right) \text{ in law}. \tag{22}$$

Proof. (I) By the strong law of large numbers holding for the chains under consideration (cf. Billingsley [1]) $\widehat{p}_n \to p(\theta_0)$ a.s., so that also $D_\phi(\widehat{p}_n, p(\theta_0)) \to 0$ a.s. Further, by the definition of $\widehat{\theta}_n$,

$$0 \le D_\phi(\widehat{p}_n, p(\widehat{\theta}_n)) \le D_\phi(\widehat{p}_n, p(\theta_0))$$

which implies $D_\phi(\widehat{p}_n, p(\widehat{\theta_n})) \to 0$ a.s. Hence, by Proposition 9.49 in Liese and Vajda [8],

$$\sum_{i=1}^{m} |\widehat{p}_{ni} - p_i(\widehat{\theta}_n)| \to 0 \quad \text{a.s.}$$

But

$$|p_i(\theta_0) - p_i(\widehat{\theta}_n)| \le |p_i(\theta_0) - \widehat{p}_{ni}| + |\widehat{p}_{ni} - p_i(\widehat{\theta}_n)|$$

so that the above convergences imply

$$\sum_{i=1}^{m} |p_i(\theta_0) - p_i(\widehat{\theta}_n)| \to 0 \quad \text{a.s.},$$

or briefly $p(\widehat{\theta}_n) \to p(\theta_0)$ a.s. By the assumed continuity of the mapping $p \mapsto \theta(p)$, this is equivalent to (18).

(II) Let us consider in the neighbourhood of θ_0 the function

$$\Psi(p, \theta) = \nabla D_\phi(p, p(\theta)) = \psi(p, \theta) J(\theta)$$

where $\psi(p, \theta) = (\psi_1(p, \theta), \ldots, \psi_m(p, \theta))$ has the components

$$\psi_i(p, \theta) = \phi\left(\frac{p_i}{p_i(\theta)}\right) - \frac{p_i}{p_i(\theta)} \phi'\left(\frac{p_i}{p_i(\theta)}\right), \qquad p = (p_1, \ldots, p_m) \in \Pi_m.$$

By taking into account the asymptotic normality of $n^{1/2}(\widehat{p}_n - p(\theta_0))$ one obtains from the Taylor theorem

$$\Psi(\widehat{p}_n, \widehat{\theta}_n) - \Psi(p(\theta_0), \widehat{\theta}_n) = \sum_{i=1}^{m} \left(\frac{\partial \Psi(p, \widehat{\theta}_n)}{\partial p_i}\right)_{p=p(\theta_0)} (\widehat{p}_{ni} - p_i(\theta_0)) + o_p(n^{-1/2}).$$

But

$$\frac{\partial \Psi(p, \theta)}{\partial p_i} = \frac{-p_i}{p_i(\theta)^2} \phi''\left(\frac{p_i}{p_i(\theta)}\right) \nabla p_i(\theta)$$

so that

$$\begin{aligned} \Psi(\widehat{p}_n, \widehat{\theta}_n) - \Psi(p(\theta_0), \widehat{\theta}_n) &= -\phi''(1) \sum_{i=1}^{m} \frac{\nabla p_i(\theta_0)}{p_i(\theta_0)} (\widehat{p}_{ni} - p_i(\theta_0)) + o_p(n^{-1/2}) \\ &= -\phi''(1) (\widehat{p}_n - p(\theta_0)) \operatorname{diag} p(\theta_0)^{-1/2} A_0 + o_p(n^{-1/2}). \end{aligned}$$

It follows from the definition of $\widehat{\theta}_n$ that $\Psi(\widehat{p}_n, \widehat{\theta}_n) = 0$. Therefore

$$\Psi(p(\theta_0), \widehat{\theta}_n) = \phi''(1) (\widehat{p}_n - p(\theta_0)) \operatorname{diag} p(\theta_0)^{-1/2} A_0 + o_p(n^{-1/2}).$$

On the other hand, we obtain in a similar way as above

$$\psi_i(p(\theta_0), \widehat{\theta}_n) - \psi_i(p(\theta_0), \theta_0) = \phi''(1) \frac{\nabla p_i(\theta_0)\,(\widehat{\theta}_n - \theta_0)^t}{p_i(\theta_0)} (1 + o_p(1)),$$

i. e.

$$\psi(p(\theta_0), \widehat{\theta}_n) - \psi(p(\theta_0), \theta_0) = \phi''(1)\,(\widehat{\theta}_n - \theta_0)\,A_0^t \mathrm{diag}\; p(\theta_0)^{-1/2} (1 + o_p(1)).$$

Multiplying both sides by $J(\widehat{\theta}_n)$ we obtain

$$\Psi(p(\theta_0), \widehat{\theta}_n) - \psi(p(\theta_0), \theta_0) J(\widehat{\theta}_n) = \phi''(1)(\widehat{\theta}_n - \theta_0) A_0^t \mathrm{diag}\; p(\theta_0)^{-1/2} J(\widehat{\theta}_n)\,(1 + o_p(1)).$$

Since $\mathbf{1}J(\theta) = 0$ for all θ under consideration and $\psi(p(\theta_0), \theta_0) = -\phi'(1)\mathbf{1}$, it holds $\psi(p(\theta_0), \theta_0) J(\widehat{\theta}_n) = 0$. This together with (18) implies that the last formula is equivalent to

$$\begin{aligned} \Psi(p(\theta_0), \widehat{\theta}_n) &= \phi''(1)(\widehat{\theta}_n - \theta_0) A_0^t \mathrm{diag}\; p(\theta_0)^{-1/2} J(\theta_0)\,(1 + o_p(1)) \\ &= \phi''(1)(\widehat{\theta}_n - \theta_0) A_0^t A_0 (1 + o_p(1)). \end{aligned}$$

From here and the former formula for $\Psi(p(\theta_0), \widehat{\theta}_n)$, we obtain

$$(\widehat{\theta}_n - \theta_0) A_0^t A_0 = (\widehat{p}_n - p(\theta_0)) \mathrm{diag}\; p(\theta_0)^{-1/2} A_0 (1 + o_p(1)).$$

Since $A_0^t A_0$ is positive definite by (A2), this implies (19).

(III) The convergence (20) follows directly from the definitions of Ω_0, B_0 and Δ_0 and from (19). Further, by employing the Taylor theorem as in (II) and using (19) and (20), one obtains (21). The convergence in (22) follows directly from (21) and from the definition of Ω_0, B_0 and Σ_0. □

Remark 2. The matrix Ω_0, and consequently the matrices B_0, Δ_0 and Σ_0 figuring in Theorem 2, are known only if $P(\theta_0) \in \boldsymbol{P}_{\theta_0}$ is specified. If this is not the case and the values of these matrices are needed to obtain confidence intervals or critical regions of statistical tests, then we can estimate the matrices B_0, Δ_0 and Σ_0 consistently by replacing the unknown elements $p_{ij}(\theta_0)$ of $P(\theta_0)$ in Ω_0 by their estimates $\widehat{p}_{nij}$ as in Remark 1 of the previous section.

Example 4. Let us consider the binary version of the model of Example 2 with $\theta \in \Theta = (0, 1)$,

$$P(\theta) = \begin{pmatrix} 1-\theta & \theta \\ 1 & 0 \end{pmatrix} \in \boldsymbol{P} \quad \text{and} \quad p(\theta) = \left(\frac{1}{1+\theta}, \frac{\theta}{1+\theta} \right).$$

We shall estimate a true parameter $\theta_0 \in (0, 1)$. We get

$$
\begin{aligned}
J(\theta_0)^t &= \frac{d}{d\theta}p(\theta_0) = \frac{1}{(1+\theta_0)^2}(-1,1),\\
D_0 &= \operatorname{diag} p(\theta_0) = \frac{1}{1+\theta_0}\begin{pmatrix} 1 & 0\\ 0 & \theta_0\end{pmatrix},\\
A_0^t &= J(\theta_0)^t D^{-1/2} = \frac{1}{(1+\theta_0)^{3/2}}(-1,\theta_0^{-1/2}),\\
A_0^t A_0 &= \frac{1}{\theta_0(1+\theta_0)^2},\\
\Delta_0^t &= (A_0^t A_0)^{-1}A_0^t = [\theta_0(1+\theta_0)]^{1/2}(-\theta_0^{1/2},1),
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_0 &= A_0(A_0^tA_0)^{-1}A_0^t = A_0\Delta_0^t = \frac{\theta_0^{1/2}}{1+\theta_0}\begin{pmatrix}\theta_0^{1/2} & -1\\ -1 & \theta_0^{-1/2}\end{pmatrix},\\
C_0^{-1} &= I - P_0 + \mathbf{1}^t p(\theta_0) = \begin{pmatrix}1&0\\0&1\end{pmatrix} - \begin{pmatrix}1-\theta_0 & \theta_0\\ 1 & 0\end{pmatrix} + \frac{1}{1+\theta_0}\begin{pmatrix}1&\theta_0\\1&\theta_0\end{pmatrix}\\
&= \frac{1}{1+\theta_0}\begin{pmatrix}\theta_0^2+\theta_0+1 & -\theta_0^2\\ -\theta_0 & 1+2\theta_0\end{pmatrix},\\
\Omega_0 &= D_0C_0 + C_0^tD_0 - D_0 - p(\theta_0)^tp(\theta_0) = \frac{\theta_0(1-\theta_0)}{(1+\theta_0)^3}\begin{pmatrix}1&-1\\-1&1\end{pmatrix}
\end{aligned}
$$

and

$$
B_0 = D_0^{-1/2}\Omega_0 D_0^{-1/2} = \frac{1-\theta_0}{(1+\theta_0)^2}\begin{pmatrix}\theta_0 & -\theta_0^{1/2}\\ -\theta_0^{1/2} & 1\end{pmatrix}.
$$

The asymptotic variance of $n^{1/2}(\widehat{\theta}_n - \theta_0)$ is $\Delta_0^t B_0 \Delta_0 = \theta_0(1-\theta_0^2)$. The asymptotic variance–covariance matrix of $n^{1/2}(p(\widehat{\theta}_n) - p(\theta_0))$ is

$$
D_0^{1/2}\Sigma_0 B_0 \Sigma_0 D_0^{1/2} = \frac{\theta_0(1-\theta_0)}{(1+\theta_0)^3}\begin{pmatrix}1&-1\\-1&1\end{pmatrix}.
$$

The asymptotic variance–covariance matrix of $n^{1/2}(\widehat{p}_n - p(\theta_0))$ is Ω_0, which coincides in this example with $D_0^{1/2}\Sigma_0 B_0 \Sigma_0 D_0^{1/2}$. The minimum ϕ–divergence estimator is

$$
\widehat{\theta}_n = \operatorname*{argmin}_{0<\theta<1} D_\phi(\widehat{p}_n, p(\theta)) = \operatorname*{argmin}_{0<\theta<1} \frac{1}{1+\theta}\left\{\phi(1+\theta)\widehat{p}_{n1} + \theta\phi\left(\frac{\theta+1}{\theta}\widehat{p}_{n2}\right)\right\}.
$$

For the class of functions ϕ_a defined in (15), we have

$$
\widehat{\theta}_n = \operatorname*{argmin}_{0<\theta<1} (1+\theta)^{a-1}\{\widehat{p}_{n1}^a + \theta^{1-a}\widehat{p}_{n2}^a\} = \frac{\widehat{p}_{n2}}{\widehat{p}_{n1}} \quad \text{if} \quad 0 < \frac{\widehat{p}_{n2}}{\widehat{p}_{n1}} < 1,
$$

which coincides with the estimator obtained by the method of moments. Thus we did not obtain a new estimator but, on the other hand, this result indicates that the minimum ϕ–divergence estimators are in general not bad.

5. TESTING COMPOSITE HYPOTHESES

In this section we consider statistical tests of composite hypothesis $\Theta_0 \subset \Theta$ introduced in Section 1 using the divergence statistics (9). The assumptions (A1) and (A2) of Section 3 are supposed to be fulfilled. Moreover, both convex functions ϕ and ϕ_* figuring in (9) are supposed to verify the assumptions imposed on ϕ in (9). The regularity assumptions concerning ϕ_* allow to extend the properties established in Theorem 2 of Section 4 to the estimator $\widehat{\theta}_n = \widehat{\theta}_n^{(\phi_*)}$ figuring in (9).

We consider the matrices Ω_0 introduced in Section 2 and A_0 and $\Sigma_0 = A_0(A_0^t A_0)^{-1} A_0^t$ introduced in Section 4 and we put for brevity

$$\begin{aligned} W_+ &= \operatorname{diag} p(\theta_0)^{1/2} \Sigma_0 \operatorname{diag} p(\theta_0)^{-1/2}, \\ W_- &= \operatorname{diag} p(\theta_0)^{-1/2} \Sigma_0 \operatorname{diag} p(\theta_0)^{1/2}, \end{aligned}$$

and

$$L_0 = \operatorname{diag} p(\theta_0)^{-1/2} [I - W_+] \Omega_0 [I - W_-] \operatorname{diag} p(\theta_0)^{-1/2}.$$

Theorem 3. Under the above considered assumptions all statistics (9) satisfy the asymptotic relation

$$T_n \to \sum_{i=1}^{m} \rho_i Z_i^2 \qquad \text{in law,} \tag{23}$$

where Z_i are independent $N(0,1)$ and ρ_i are the eigenvalues of the matrix L_0.

Proof. By (21),

$$p(\widehat{\theta}_n) = p(\theta_0) + (\widehat{p}_n - p(\theta_0)) W_- + o_p(n^{-1/2}).$$

Therefore

$$\widehat{p}_n - p(\widehat{\theta}_n) = (\widehat{p}_n - p(\theta_0)) (I - W_-) + o_p(n^{1/2}).$$

It follows from here and from the relation

$$n^{1/2}(\widehat{p}_n - p(\theta_0)) \to N(0, \Omega_0) \qquad \text{in law (cf. Section 3)}$$

that

$$n^{1/2}(\widehat{p}_n - p(\widehat{\theta}_n)) \to N(0, (I - W_-)^t \Omega_0 (I - W_-)) \qquad \text{in law.}$$

Since $(I - W_-)^t = I - W_+$, it follows from here

$$n^{1/2}(\widehat{p}_n - p(\widehat{\theta}_n)) \operatorname{diag} p(\theta_0)^{-1/2} \to N(0, L_0) \qquad \text{in law.}$$

Further, it follows from there

$$(\widehat{p}_n - p(\widehat{\theta}_n)) \operatorname{diag} p(\widehat{\theta}_n)^{-1/2} = (\widehat{p}_n - p(\widehat{\theta}_n)) \operatorname{diag} p(\theta_0)^{-1/2} + o_p(n^{-1/2})$$

so that also

$$n^{1/2}(\widehat{p}_n - p(\widehat{\theta}_n)) \operatorname{diag} p(\theta_n)^{-1/2} \to N(0, L_0) \qquad \text{in law.}$$

Finally, since $U_n = n^{1/2}(\widehat{p}_n - p(\widehat{\theta}_n))$ diag $p(\widehat{\theta}_n)^{-1/2}$ satisfies the relation $U_n U_n^t = X^2(\widehat{p}_n, p(\widehat{\theta}_n))$ where $X^2(\widehat{p}_n, p(\widehat{\theta}_n))$ is defined in accordance with Section 2, the last relation implies

$$X^2(\widehat{p}_n, p(\widehat{\theta}_n)) \to \sum_{i=1}^{m} \rho_i Z_i^2 \qquad \text{in law}$$

for ρ_i and Z_i considered in Theorem 3. The desired relation (23) follows from here and from the fact that under (18) it holds $\widehat{p}_{ni}/p_i(\widehat{\theta}_n) = 1 + o_p(1)$ so that, for the statistics (9), (14) can be extended into the form

$$T_n = X^2(\widehat{p}_n, p(\widehat{\theta}_n))(1 + o_p(1)). \qquad \square$$

Remark 3. Theorem 3 leads to the family of asymptotically α-level tests (12) for the eigenvalues $\rho_1, \ldots \rho_m$ figuring in (23). These eigenvalues depend not only on the unknown chain transition matrix $P(\theta_0) = (p_{ij}(\theta_0))$, but also on the unknown stationary distribution $p(\theta_0)$. Replacing the matrix by the consistent estimate $\widehat{P}_n = (\widehat{p}_{nij})$ defined in Remark 1 (cf. also Remark 2) and $p(\theta_0)$ by the consistent estimate $\widehat{p}_n$ defined by (6), we obtain an estimate $\widehat{L}_n$ of the matrix L_0. Similarly as in Remark 1, we can argue that the eigenvalues $\rho_{n1}, \ldots, \rho_{nm}$ of $\widehat{L}_n$ are consistent estimates of the eigenvalues figuring in (23) and in the formula

$$Q_\alpha = Q_\alpha(\rho_1, \ldots, \rho_m)$$

for critical values of the tests (12). Therefore the empirical $(1 - \alpha)$-quantile

$$Q_{n\alpha} = Q_\alpha(\rho_{n1}, \ldots, \rho_{nm}) \tag{24}$$

tends in probability to Q_α.

Corollary 3. All tests in the family (12) with T_n given in (9) and $Q_{n\alpha}$ given by (24) are asymptotically α-level for the composite hypothesis Θ_0 under consideration.

We demonstrate practical applicability of Theorem 3 and Corollary 3 by two examples, which at the same time illustrate practical advantages as well as disadvantages of the testing method proposed in this section.

Example 5. Let us consider a composite hypothesis $\Theta_0 = (a, b) \subset (0, 1)$ in the model of Example 4. It follows from the results of Example 4 that

$$I - W_+ = \frac{1}{1 + \theta_0} \begin{pmatrix} 1 & 1 \\ \theta_0 & \theta_0 \end{pmatrix}, \quad I - W_- = (I - W_+)^t$$

and

$$L_0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

with both eigenvalues $\rho_1 = \rho_2 = 0$. By employing the results of Example 4 we see that the statement of Theorem 3 is in this case true. Further, the $(1-\alpha)$-quantile $Q_\alpha(\rho_1) = 0$ and all tests of Corollary 3 are asymptotically 0-level. Hence the statement of Corollary 3 is true too. Of course the practical significance of the tests (12) is in this case doubtful as their powers tend exponentially to zero.

Example 6. Let us consider the ternary version of the model of Example 2, with $\theta = (\beta, \gamma) \in \Theta = (0,1)^2$,

$$P(\beta,\gamma) = \begin{pmatrix} 1-\beta & \beta & 0 \\ 1-\gamma & 0 & \gamma \\ 1 & 0 & 0 \end{pmatrix} \in \boldsymbol{P}_{\beta,\gamma}$$

and with the stationary distribution $p(\beta,\gamma) = (1,\beta,\beta\gamma)/(1+\beta+\beta\gamma)$. Let the composite hypothesis be

$$\Theta_0 = \{(\beta,\gamma) \in (0,1)^2 : \gamma = \beta,\ 1/2 \le \beta < 1\}$$

and consider a true parameter $\theta_0 = (\beta_0, \beta_0)$ with $0 < \beta_0 < 1$. Here, of the eigenvalues ρ_1, ρ_2 and ρ_3 of the matrix L_0, only $\rho_1 = \rho(\beta_0)$ is nonzero. The values of $\rho(\beta_0)$ are given for various β_0 in Table 1.

Table 1. The nonzero eigenvalue of L_0 for the transition matrix $P(\beta_0, \beta_0)$.

β_0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\rho(\beta_0)$	11.78	3.2	1.37	0.7	0.38	0.2	0.09	0.03	0.008

Therefore, under the chain transition matrix $P(\beta_0, \beta_0)$ given above, the asymptotic distribution of all statistics (9) is $\rho(\beta_0)\chi_1^2$. Consequently

$$Q_\alpha(\rho_1,\rho_2,\rho_3) = \rho(\beta_0)\chi_1^2(1-\alpha),$$

where $\chi^2(\alpha)$ denotes the α-quantile of the random variable χ_1^2. Let us denote by $\widehat{\rho}_{n1}$, $\widehat{\rho}_{n2}$ and $\widehat{\rho}_{n3}$ the eigenvalues of the estimate $\widehat{P}_n$ of $P(\beta_0, \beta_0)$ considered in Corollary 3. Then

$$Q_{n\alpha} = Q_\alpha(\widehat{\rho}_{n1}, \widehat{\rho}_{n2}, \widehat{\rho}_{n3})$$

under the hypothesis tends to $\rho(\beta_0)\chi_1^2(1-\alpha)$ with $1/2 \le \beta_0 < 1$. By using the maximal value $\rho(\beta_0) = 0.38$ from Table 1, we obtain a family of tests $\mathcal{T} = \{(T_n^\phi, 0.38\chi_1^2(1-\alpha)) : \phi \in \Phi\}$ which are asymptotically α-level for the composite hypotheses under consideration.

(Received June 17, 1998.)

REFERENCES

[1] P. Billingsley: Statistical methods in Markov chains. Ann. Math. Statist. *32* (1961), 12–40.
[2] M. W. Birch: A new proof of the Pearson–Fisher Theorem. Ann. Math. Statist. *35* (1964), 817–824.
[3] Y. M. M. Bishop, S. E. Fienberg and P. W. Holland: Discrete Multivariate Analysis. Theory and Practice. The MIT Press, Cambridge, Massachusetts 1975.
[4] N. Cressie and T. R. C. Read: Multinomial goodness of fit tests. J. Royal Statist. Soc., Ser. B *46* (1984), 440–464.
[5] F. C. Drost, W. C. M. Kallenberg, D. S. Moore and J. Oosterhoff: Power approximations to multinomial tests of fit. J. Amer. Statist. Assoc. *84* (1989), 130–141.
[6] L. J. Glesser and D. S. Moore: The effect of dependence on chi-squared and empiric distribution tests of fit. Ann. Statist. *11* (1983), 1100-1108.
[7] L. J. Glesser and D. S. Moore: The effect of positive dependence on chi-squared tests for categorical data. J. Royal Statis. Soc., Ser. B *47* (1983), 459–465.
[8] F. Liese and I. Vajda: Convex Statistical Distances. Teubner, Leipzig 1987.
[9] M. L. Menéndez, D. Morales, L. Pardo and I. Vajda: Divergence–based estimation and testing of statistical models of classification. J. Multivariate Anal. *54* (1996), 329–354.
[10] M. L. Menéndez, D. Morales, L. Pardo and I. Vajda: Testing in stationary models based on f–divergences of observed and theoretical frequencies. Kybernetika *33* (1997), 465–475.
[11] D. S. Moore: The effect of dependence on chi-squared tests of fit. Ann. Statist. *10* (1982), 1163–1171.
[12] D. Morales, L. Pardo and I. Vajda: Asymptotic divergence of estimates of discrete distributions. J. Statist. Plann. Inference *48* (1995), 347–369.
[13] T. R. C. Read and N. A. C. Cressie: Goodness–of–Fit Statistics for Discrete Multivariate Data. Springer, Berlin 1988.
[14] M. Salicrú, D. Morales, M. L. Menéndez and L. Pardo: On the applications of divergence type measures in testing statistical hypotheses. J. Multivariate Anal. *51* (1994), 372–391.
[15] S. Tavaré and P. M. E. Altham: Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics. Biometrika *70* (1983), 139–144.

María Luisa Menéndez, Department of Applied Mathematics, Technical University of Madrid, E28040 Madrid. Spain.
e-mail: mmenende@aq.upm.es

Domingo Morales, Department of Statistics and Applied Mathematics, Miguel Hernández University of Elche, E03206 Elche. Spain.
e-mail: d.morales@umh.es

Leandro Pardo, Department of Statistics and Operations Research, Complutense University of Madrid, E28040 Madrid. Spain.
e-mail: leandro_pardo@mat.ucm.es

Igor Vajda, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: vajda@utia.cas.cz