KYBERNETIKA — VOLUME 33 (1997), NUMBER 5, PAGES 465 - 475

TESTING IN STATIONARY MODELS BASED ON DIVERGENCES OF OBSERVED AND THEORETICAL FREQUENCIES[1]

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

Goodness-of-fit tests for stationary distributions of dependent data are considered, based on f-divergences of observed and theoretical cell frequencies. Pearson's X_n^2 is a special version. A methodology is presented leading to asymptotically α-level variants of these tests, and also to the selection of most powerful versions. This methodology is illustrated on binary Markov data. Similar procedures have been previously established for independent data. The possibility to extend these procedures to dependent data is a new argument in favour of the f-divergence alternatives to the classical Pearson's X_n^2.

1. INTRODUCTION

Let $\boldsymbol{X} = (X_0, X_1, \ldots)$ be a stationary sequence of random variables taking on values in $\mathcal{X} \subset R$, and P the distribution of components $X_0, X_1, \ldots$ on $\mathcal{X}$. We consider the statistical test of the hypothesis $P = P_0$ based on observations $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ quantized by a fixed decomposition $\mathcal{D} = (D_1, \ldots, D_m)$ of $\mathcal{X}$. In other words, we consider the classical goodness-of-fit tests for vectors $\hat{p}_n = (\hat{p}_{n1}, \ldots, \hat{p}_{nm})$ of the observed cell frequencies

$$\hat{p}_{ni} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{D_i}(X_k)$$

and vectors $p = (p_1, \ldots, p_m)$ of the theoretical cell frequencies $p_i = P(D_i)$. The hypothesis P_0 is indicated by writing $p_0 = (p_{01}, \ldots, p_{0m})$ instead of $p = (p_1, \ldots, p_m)$ and it is assumed that all components of p_0 are nonzero.

The classical goodness-of-fit test for independent observations is based on the Pearson's statistic

$$X_n^2 = n \sum_{i=1}^{m} \frac{(\hat{p}_{ni} - p_{0i})^2}{p_{0i}}. \tag{1}$$

[1]This work was supported by grants DGICYT PB 93-0068, PB 93-0022 and GA AV ČR A 1075709.

Other common tests are based on the divergences between observed and theoretical frequencies $\hat{p}_n$ and p_0 introduced by Csiszár [3] and Ali and Silvey [1], defined for convex functions $f : (0, \infty) \to R$ by the formula

$$D_f(\hat{p}_n, p_0) = \sum_{i=1}^{m} p_{0i} f\left(\frac{\hat{p}_{ni}}{p_{0i}}\right),$$

where $f(0)$ (possibly infinite) is obtained by the continuous extension.

We restrict ourselves to the convex functions $f(t)$ with $f(1) = 0$, twice continuously differentiable in a neighbourhood of $t = 1$ with $f''(1) = 1$. We shall normalize the f-divergences into the form of statistics

$$T_n^f = 2n\, D_f(\hat{p}_n, p_0). \tag{2}$$

Obviously the Pearson's statistic (1) coincides with T_n^f for $f(t) = (t-1)^2/2$. The class

$$T_n^a = T_n^{f_a}, \quad a \in R, \tag{3}$$

of the so-called power divergence statistics, defined by $f_a(t) = (t^a - 1)/[a(a-1)]$ for $a \neq 0$, $a \neq 1$, and by

$$f_0(t) = -\ln t \quad \text{for } a = 0$$

and

$$f_1(t) = t \ln t \quad \text{for } a = 1,$$

has been introduced by Cressie and Read [4] (cf. also Read and Cressie [7]). This class satisfies the above considered assumptions and contains the best known goodness-of-fit test statistics. As an example of statistic not contained in this class but satisfying our assumptions one can take T_n^f for $f(t) = 2(1-t)/(1+t)$. Similar function has been used to define f-divergence of probability distributions in Rukhin [8]. Other examples can be found in Liese and Vajda [5].

We present a methodology for specification of critical values and powers of the tests based on (2) and (3) in the framework of statistical models with dependent data satisfying an asymptotic normality condition. We also describe a method of specification of the most powerful test, provided it exists. The general methodology is illustrated on stationary Markov models. The problem of most powerful test is solved numerically for the uniform hypothesis in the framework of binary Markov data models.

2. ASYMPTOTICALLY α–LEVEL f–DIVERGENCE TESTS

Under the assumptions about f considered in (2) it holds asymptotically, for $t \to 1$,

$$\begin{aligned} f(t) &= f(1) + f'(1)(t-1) + \frac{f''(1)}{2}(t-1)^2 + o((t-1)^2) \\ &= f'(1)(t-1) + \frac{1}{2}(t-1)^2 + o((t-1)^2). \end{aligned}$$

If under the hypothesis asymptotically, for $n \to \infty$,

$$\hat{p}_{ni} = p_{0i} + o_p(1) \qquad \text{for all} \quad 1 \le i \le m \tag{4}$$

then

$$\begin{aligned} D_f(\hat{p}_n, p_0) &= \frac{1}{2}\sum_{i=1}^{m} \frac{(\hat{p}_{ni} - p_{0i})^2}{p_{0i}} + o_p\left(\sum_{i=1}^{m} \frac{(\hat{p}_{ni} - p_{0i})^2}{p_{0i}}\right) \\ &= \frac{1}{2n} X_n^2 + o_p\left(\frac{X_n^2}{n}\right) \qquad \text{(cf.(1))}, \end{aligned}$$

so that by (2)

$$T_n^f = X_n^2(1 + o_p(1)). \tag{5}$$

We consider models satisfying the assumption

$$\sqrt{n}\left(\frac{\hat{p}_{n1} - p_{01}}{\sqrt{p_{01}}}, \ldots, \frac{\hat{p}_{nm} - p_{0m}}{\sqrt{p_{0m}}}\right) \longrightarrow N(0, V) \quad \text{in law.} \tag{6}$$

E. g., this is the case when the quantized process forms an irreducible aperiodic Markov chain (cf. Tavaré and Altham [9]). Under (6),

$$X_n^2 \longrightarrow Y \equiv \sum_{i=1}^{m} \rho_i Z_i^2 \quad \text{in law} \tag{7}$$

where ρ_i are eigenvalues of the matrix V and Z_i are independent $N(0,1)$. Hence we have the following

Theorem 1. If the model satisfies the regularity assumptions (4) and (6) and Q_α is the $(1-\alpha)$–quantile of Y defined by (7) then, for all f under consideration, the tests (T_n^f, Q_α) are asymptotically of size α.

Remark 1. The matrix V and, consequently, the eigenvalues ρ_i may not be specified uniquely by the null hypothesis P_0 (uniquely is specified only the marginal distribution of components X_i). If V depends continuously on the model parameters which remain free under P_0, and there exist consistent estimates of these parameters leading to the estimate V_n of the matrix V, then we can use the tests

$$(T_n^f, Q_{n\alpha}), \tag{8}$$

where $Q_{n\alpha}$ is the $(1-\alpha)$–quantile of

$$Y_n = \sum_{1=1}^{m} \rho_{ni} Z_i^2$$

and ρ_{ni} are eigenvalues of the matrix V_n. The continuity argument leads to the conclusion that ρ_{ni} consistently estimate ρ_i, i. e. $Q_{n\alpha}$ consistently estimates Q_α. As a result, all tests (8) are of the asymptotic size α.

Theorem 1 and Remark 1 assert that all test (8) are asymptotically equivalent from the point of view of the test size defined by

$$\alpha_n(P_0, f) = \Pr(T_n^f > Q_{n\alpha}|P_0).$$

Preferences between them have thus be based on other criteria, e. g. on the test powers

$$\pi_n(P, f) = \Pr(T_n^f > Q_{n\alpha}|P) \tag{9}$$

for $P \neq P_0$. In general, there is no universally preferable test in the class (8) (cf. e. g. p. 411 in Cressie and Read [4] in the particular case of independent observations). However, for special models with independent observations the most powerful test might exist (cf. e. g. Menéndez et al [6]).

The problem is what methodology leads to the selection of the most powerful test in the class (8). If the asymptotic size $0 < \alpha < 1$ is fixed, this problem reduces to the most preferable statistic T_n^f. The class (3) seems to be rich and interesting enough to justify the reduction of the original problem into the problem of most preferable statistic T_n^a. Indeed (cf. e. g. Read and Cressie [7]), T_n^2 is the Pearson's X_n^2, T_n^{-1} is the Neyman modified X_n^2 statistic and $T_n^{1/2}$ is the Freeman–Tukey statistic F_n^2. Further,

$$T_n^1 = 2n \sum_{i=1}^{n} \hat{p}_{0i} \ln \frac{\hat{p}_{0i}}{p_{0i}} \quad \text{and} \quad T_n^0 = 2n \sum_{i=1}^{n} p_{0i} \ln \frac{p_{0i}}{\hat{p}_{ni}}$$

are the loglikelihood ratio statistic G_n^2 and the modified G_n^2 statistic. Thus the circle of candidates T_n^a may be restricted by an interval of a's arround the "center of symmetry" $a = 1/2$, large enough to contain the above listed important particular cases. Within this interval, the most preferable a can be obtained by a statistical experimentation, similar to what has been called small sample studies on p. 143 of Read and Cressie [7]. It consists of the approximation of $\pi_n(P, a) \equiv \pi_n(P, f_a)$ given by (9) by the relative frequency $\pi_{n,N}(P, a)$ of the event $T_n^a > D_{n,\alpha}$ in a large number N of simulated realizations of $\boldsymbol{X}_n$. The simulations should be carried out for $P = P_0$ and for several sufficiently representative alternatives $P \neq P_0$. The most preferable value a_* of the parameter a is defined by the condition that, for all selected $P \neq P_0$, $\pi_{n,N}(P, a_*) = \max \pi_{n,N}(P, a)$, where the maximum extends over the subdomain of a's where $|\alpha - \pi_{n,N}(P_0, a)|$ achieves minimum or nearly minimum values.

3. APPLICATIONS TO REVERSIBLE MARKOV MODELS

In this section we apply the above considered tests to irreducible aperiodic Markov chains $\boldsymbol{X} = (X_1, X_2, \ldots)$ with states $1, \ldots, m$ and stochastic $m \times m$ matrix $\boldsymbol{P}$ of transition probabilities. The decomposition may be defined by $D_i = \{i\}$. Then the regularity assumptions (4) and (6) hold (cf. e. g. Billingsley [2]). Since the distributions P in this case coincide with the vectors p, we can replace P and P_0 by p and p_0. Hence we consider the hypothesis $p = p_0$ about the stationary distribution of the chain matrix $\boldsymbol{P}$. Since no states are transient, the hypothesis satisfies the condition that all components of p_0 are nonzero.

In the model under consideration the goodness–of–fit tests of the hypothesis p_0 based on the statistic X_n^2 have been considered previously by Tavaré and Altham [9]. They established relation (7) and for reversible chains they found a simplified representation of the eigenvalues ρ_i. Namely, they proved the following version of (7):

$$X_n^2 \longrightarrow Y \equiv \sum_{i=1}^{m-1} \frac{1+\lambda_i}{1-\lambda_i} Z_i^2 \qquad \text{in law,} \tag{10}$$

where $\lambda_1, \ldots, \lambda_{m-1}$ are the nonunit eigenvalues of the chain matrix $\boldsymbol{P}$. Combining this with our previous result (5) we obtain the following assertion.

Theorem 2. If the model is an irreductible, aperiodic and reversible Markov chain and Q_α is an $(1-\alpha)$–quantile of the random variable Y figuring in (10) then, for all f under consideration, the tests (T_n^f, Q_α) are asymptotically of size α.

Remark 2. Except very special cases, the matrix $\boldsymbol{P}$ is not uniquely defined by the condition $p_0 = p_0\boldsymbol{P}$. But the relative frequencies

$$\hat{p}_n(i,j) = \frac{\sum_{k=2}^n \mathbf{1}_{\{(i,j)\}}(X_{k-1}, X_k)}{\sum_{k=2}^n \mathbf{1}_{\{i\}}(X_{k-1})} \tag{11}$$

consistently estimate the transition probabilities $p(i,j)$ of the matrix $\boldsymbol{P}$ (cf. Billingsley [2]). Since the eigenvalues λ_i considered in (10) are continuous functions of the elements $p(i,j)$ of $\boldsymbol{P}$, the substitution $\hat{p}_n(i,j) = p(i\ j)$ in these functions leads to consistent estimates λ_{ni} of λ_i. Denote by Y_n the random variable defined by (10) with λ_i replaced by λ_{ni} and by $Q_{n\alpha}$ the corresponding $(1-\alpha)$–quantile. Then one can argue similarly as in Remark 1 that the conclusion of Theorem 2 holds with Q_α replaced by $Q_{n\alpha}$.

Example. We present evaluation of the most preferable test $(T_n^a, Q_{n\alpha})$ in the sense specified above for the uniform hypothesis $p_0 = (1/2, 1/2)$ in the binary Markov model satisfying the assumptions of Theorem 2. Obviously,

$$\begin{pmatrix} 1-\beta & \beta \\ \gamma & 1-\gamma \end{pmatrix}: \quad 0 < \beta, \gamma \leq 1, \quad \beta + \gamma < 2,$$

is the class of possible Markov matrices $\boldsymbol{P}$ and

$$\begin{pmatrix} 1-\beta & \beta \\ p_1\beta/p_2 & 1-p_1\beta/p_2 \end{pmatrix}: \quad 0 < \beta \leq \min\left\{1, \frac{p_2}{p_1}\right\}, \quad \beta < 1, \tag{12}$$

is the subclass satisfying the condition $(p_1, p_2) = (p_1, p_2)\boldsymbol{P}$. In particular,

$$\begin{pmatrix} 1-\beta & \beta \\ \beta & 1-\beta \end{pmatrix}: \quad 0 < \beta < 1,$$

is the subclass under the uniform hypothesis p_0. The nonunit eigenvalue of this matrix is $1-2\beta$ so that we get from (10), (11)

$$Y = \frac{1-\beta}{\beta} Z^2, \qquad Y_n = \frac{\hat{p}_n(1,1)+\hat{p}_n(2,2)}{2-\hat{p}_n(1,1)-\hat{p}_n(2,2)} Z^2.$$

Hence

$$Q_{n\alpha} = \frac{2-\hat{p}_n(1,1)-\hat{p}_n(2,2)}{\hat{p}_n(1,1)+\hat{p}_n(2,2)} \chi_1^2(1-\alpha).$$

Let us choose the most preferable statistics T_n^a by using the relative frequencies $\pi_{n,10^4}(p, f_a) = \pi_{n,10^4}(\theta, a)$ evaluated for $20 \le n \le 50$, $-4 \le a \le 8$ and all

$$p = (p_1, p_2) \equiv (\theta, 1-\theta), \quad 0 < \theta \le \beta/(\beta+1),$$

where the domain of θ depends on $0 < \beta < 1$ in matrix (12). For our experiment we have chosen $\beta \in \{1/4, 1/2, 3/4\}$. Figures 1 a, b present the corresponding behavior of $\pi_{20,10^4}(\theta, a)$, $\pi_{50,10^4}(\theta, a)$ on the interval $0 < \theta < 2/3$ for $\beta = 1/2$ and selected values of a. From the point of view of our task, these results are representative enough also for $\beta = 1/4$ and $\beta = 3/4$. To illustrate this conclusion, we present $\pi_{50,10^4}(\theta, a)$ for $\beta = 1/4$ in Figure 2 a and $\pi_{20,10^4}(\theta, a)$ for $\beta = 3/4$ in Figure 2 b.

We see from the figures that the rate of convergence to the designed test size α depends on the test and also on the association between data represented by the matrix (12). Differences between the test powers depend too, they seem to grow with the degree of association.

In accordance with the criterion formulated at the end of Section 2, we can conclude from the presented figures that in testing the uniformity of stationary Markov distributions, the most preferable among (3) seems to be the Neyman modified X_n^2 statistic T_n^{-1} represented in Figures 1 and 2 by $+$. The power of the statistics T_n^{-4} and T_n^8 represented by $\blacksquare$ and $\times$ exceeds that of T_n^{-1} but the α–levels $\pi_{n,10^4}(0.5, -4)$ and $\pi_{n,10^4}(0.5, 8)$ are too far from the designed size $\alpha = 0.05$. On the other hand, the minimum of absolute deviation $|0.05 - \pi_{n,10^4}(0.5, a)|$ is achieved at $a = 2$ but $\pi_{n,10^4}(0.5, -1)$ practically coincides with $\pi_{n,10^4}(0.5, 2)$, and $\pi_{n,10^{-4}}(\theta, -1)$ essentially exceeds $\pi_{n,10^{-4}}(\theta, 2)$ for $\theta \neq 0.5$.

It is interesting (cf. the results about powers of the tests based on the f–divergence statistics considered in (3), summarized in the case of independent observations in Read and Cressie [7]), that in the case of independent data the Neyman modified X_n^2 was never found optimal in the stated sense. Also quite sharp differences in power clearly visible in the presented figures are interesting. They in fact represent strong arguments in favour of f–divergence alternatives to the classical Pearson's X_n^2, similar to those collected by Read and Cressie [7] for models with independent observations.

Our results demonstrate that the investigation of powers of tests based on f–divergence statistics leads even in models with dependent data to nontrivial results. In this sense they motivate the need to extend this study to composite hypotheses about marginal distributions of stationary data sources.

(Received September 19, 1996.)

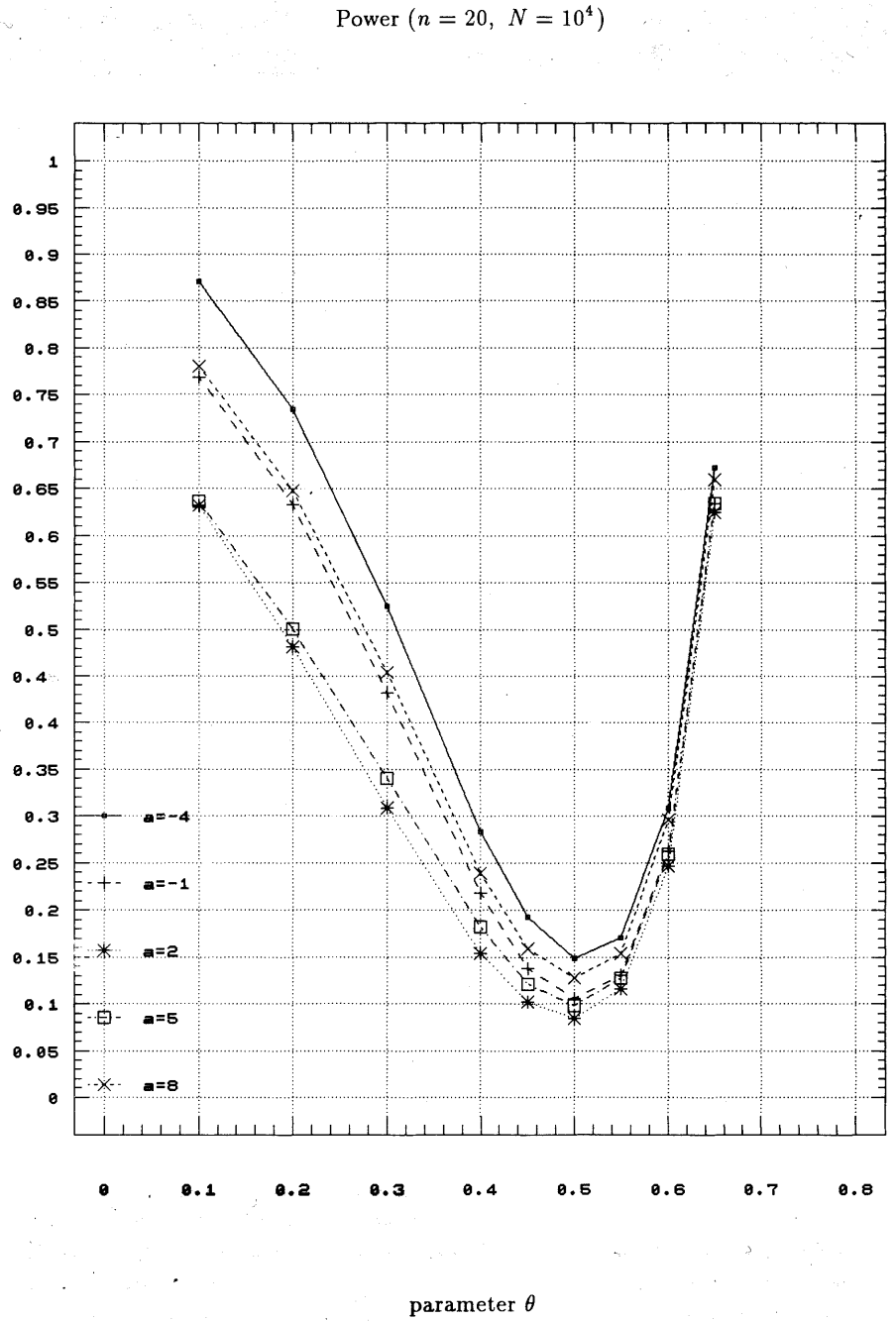

Fig. 1 a. $\pi_{20,10^4}(\theta, a)$ as a function of θ for selected a and $\beta = 1/2$.

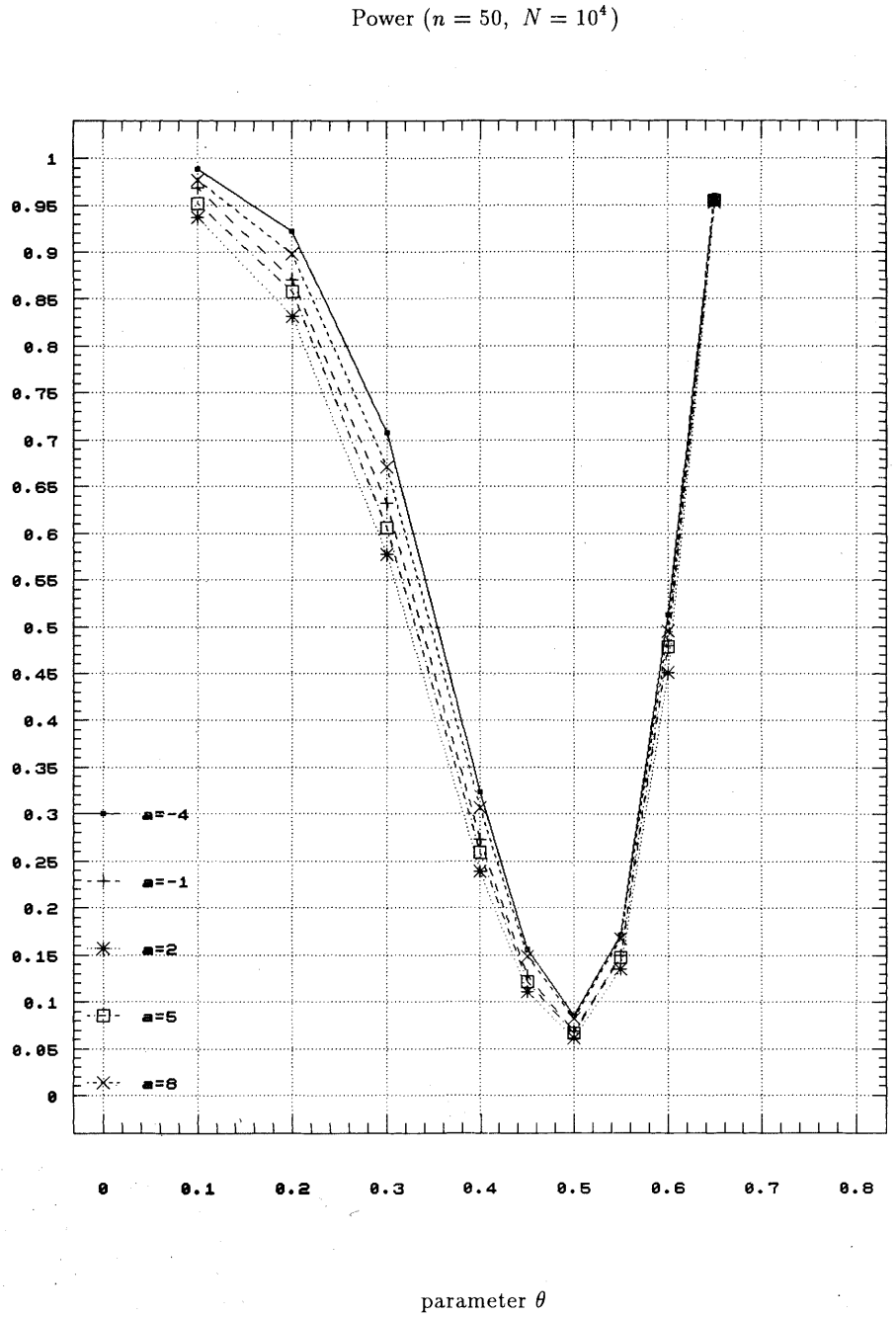

Fig. 1 b. $\pi_{50,10^4}(\theta, a)$ as a function of θ for selected a and $\beta = 1/2$.

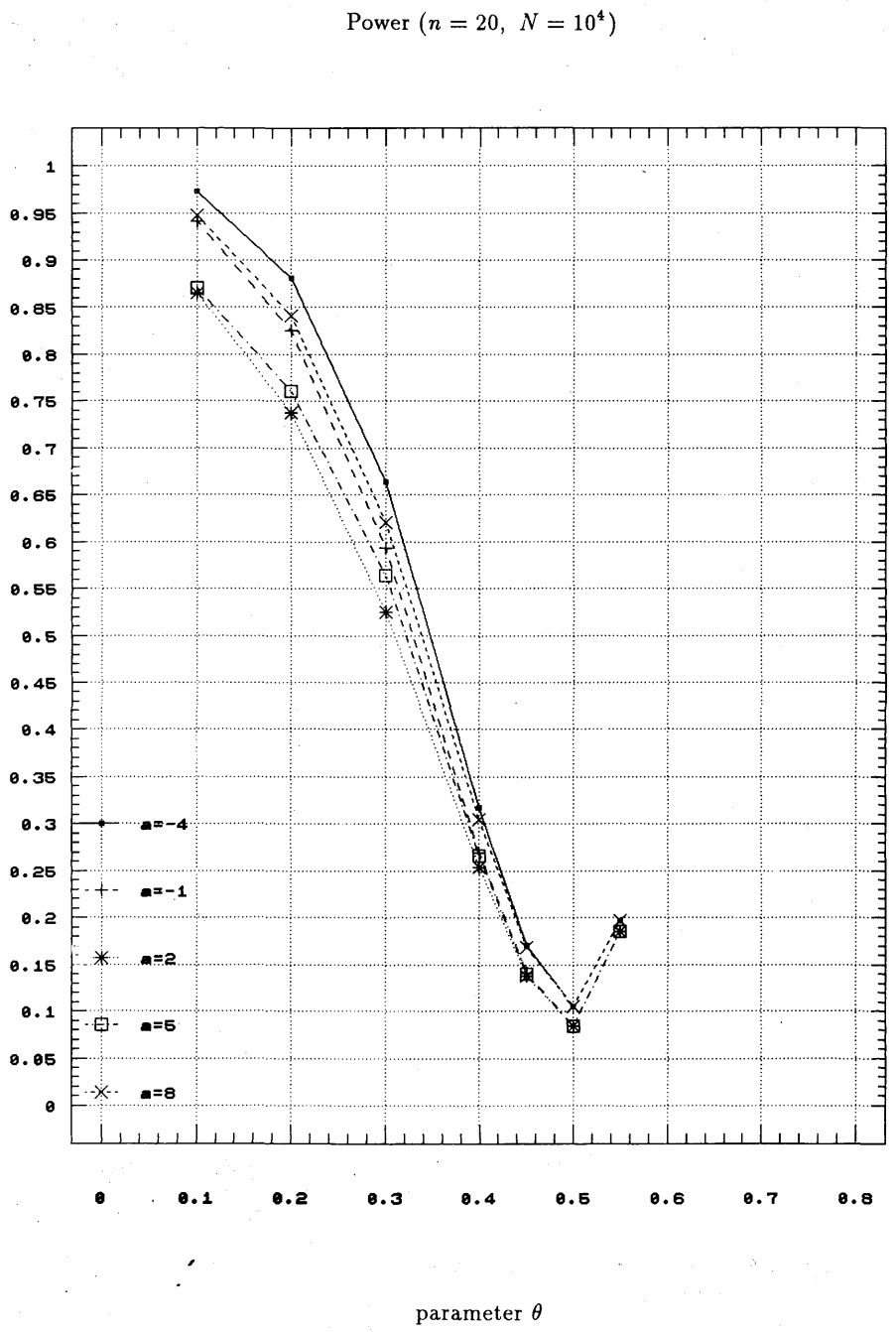

Fig. 2 a. $\pi_{20,10^4}(\theta, a)$ as a function of θ for selected a and $\beta = 1/4$.

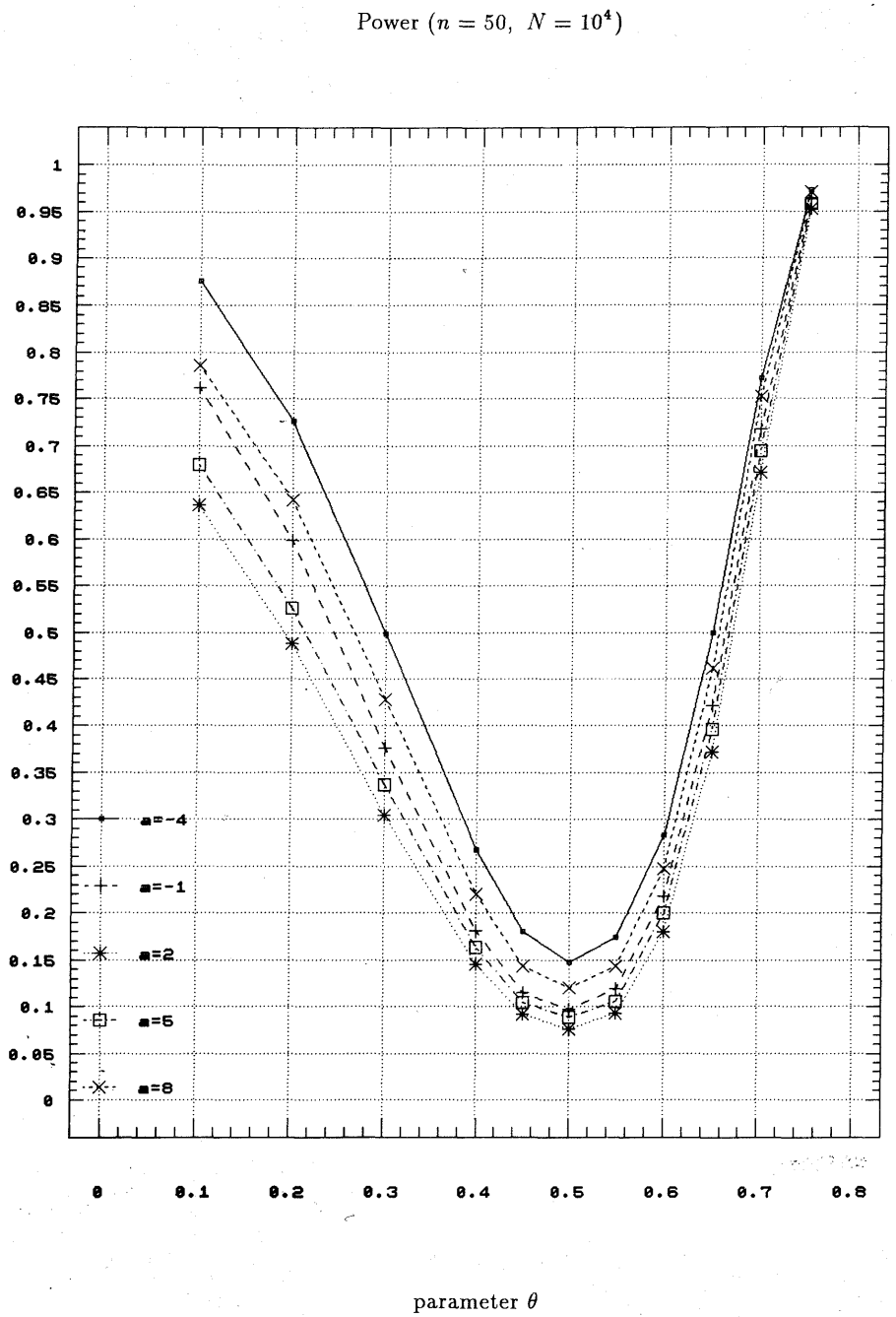

Fig. 2 b. $\pi_{50,10^4}(\theta, a)$ as a function of θ for selected a and $\beta = 3/4$.

REFERENCES

[1] S. M. and S. D. Silvey:A general class of coefficients of divergence of one distribution from another. J. Roy. Statist. Soc. *286* (1996), 131–142.

[2] P. Billingsley: Statistical methods in Markov chains. Ann. Math. Statist. *32* (1961), 12–40.

[3] I. Csiszár: Information type measures of difference of probability distributions and indirect observations. Stud. Sci. Mathem. Hungaricae *2* (1967), 299–318.

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

[5] F. Liese and I. Vajda: Convex Statistical Distances. Teubner, Leipzig 1987.

[6] M. L. Menéndez, D. Morales, L. Pardo and I. Vajda: Divergence-based estimation and testing of statistical models of classification. J. Multivariate Anal. *54* (1995), 329–354.

[7] T. R. C. Read and N. A. C. Cressie: Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer-Verlag, Berlin 1988.

[8] A. L. Rukhin: Optimal estimator of the mixture parameter by the method of moments and information affinity. In: Trans. 12 th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes (J. Á. Víšek and P. Lachout, eds.), Institute of Information Theory and Automation, Czech Academy of Sciences, Prague 1994, pp. 214–219.

[9] 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.

Prof. María Luisa Menéndez, Department of Applied Mathematics, Technical University of Madrid, E-28040 Madrid. Spain.

Prof. Domingo Morales and Prof. Leandro Pardo, Department of Statistics and Operations Research, Complutense University of Madrid, E-28040 Madrid. Spain.

Ing. Igor Vajda, DrSc., Ústav teorie informace a automatizace AV ČR (Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic), Pod vodárenskou věží 4, 18208 Praha 8. Czech Republic.