KYBERNETIKA — VOLUME *34* (1998), NUMBER 3, PAGES 245 – 263

GLOBAL INFORMATION IN STATISTICAL EXPERIMENTS AND CONSISTENCY OF LIKELIHOOD–BASED ESTIMATES AND TESTS[1]

Igor Vajda

In the framework of standard model of asymptotic statistics we introduce a global information in the statistical experiment about the occurrence of the true parameter in a given set. Basic properties of this information are established, including relations to the Kullback and Fisher information. Its applicability in point estimation and testing statistical hypotheses is demonstrated.

1. INTRODUCTION

We consider the standard conceptual framework of asymptotic statistics, i. e. a statistical experiment consisting of a sequence of product probability spaces parametrized by $\theta \in \Theta \subset R^m$. Under some regularity the Fisher information $\mathcal{I}_{\theta_0}$ characterizes the amount of information provided by the experiment about the true parameter value $\theta_0 \in \Theta$. This information is local. As found by Kullback [8], Rao [16] and some others, $\mathcal{I}_{\theta_0}$ measures the local sensitivity of the sample distribution P_θ figuring in the experiment to small variations of parameter θ in the neighbourhood of θ_0. If $I(P_{\theta_0}, P_\theta)$ is the Kullback information (information divergence of P_{θ_0} and P_θ) then, asymptotically for $\theta \to \theta_0$,

$$I(P_{\theta_0}, P_\theta) = \frac{1}{2}\,(\theta - \theta_0)\,\mathcal{I}_{\theta_0}(\theta - \theta_0)^t + o(\|\theta - \theta_0\|^2).$$

If one arbitrarily modifies the distributions P_θ with θ outside an arbitrarily small neighborhood of θ_0 then $\mathcal{I}_{\theta_0}$ remains unchanged.

We are interested in the global information contained in the experiment about the true parameter. The first concept of global information in a statistical experiment has been proposed by Lindley [11] and developed later by several authors, see Rényi [17, 18]. This concept was based on the approach of Shannon (see Cover and Thomas [3]), where the information is the difference between prior and posterior uncertainties. De Groot [4, 5] extended this approach and considered the difference between

[1] Supported by the Grant Agency of the Academy of Sciences of the Czech Republic under Grant 175 402.

prior and posterior risks (for further developments and references see Torgersen [21]). Obviously, all definitions of this kind are restricted to Bayesian experiments.

In this paper we propose a global information applicable to the classical non-bayesian statistical experiments. It is an asymptotic characteristic of the experiments whose values are changed by an appropriate modification of any distribution P_θ figuring in the experiment.

We introduce the global information as a real valued function $I_{\theta_0}(S)$ defined for all open or closed sets $S \subset \Theta$ and all parameter values $\theta_0 \in \Theta$. The real number $I_{\theta_0}(S)$ characterizes an asymptotic likelihood of the event that the true parameter θ_0 belongs to S. We present formulas for evaluation of this information and clarify its relation to both the information divergence $I(P_{\theta_0}, P_\theta)$ of Kullback and to the local information $\mathcal{I}_{\theta_0}$ of Fisher.

We also study the applicability of the global information to the maximum likelihood estimates and generalized likelihood ratio tests. As shown in Vajda [23] and Liese and Vajda [10], an asymptotically maximum likelihood estimate in a contaminated experiment can easily be inconsistent (a new example illustrating this is presented in Section 3 below). Similar phenomenon can take place for the generalized likelihood ratio tests. Perlman [13], Pfanzagl [14, 15], Strasser [20], Vajda [23] and Liese and Vajda [10] considered necessary and sufficient conditions for consistency of all asymptotically maximum likelihood estimates. In this paper a new necessary and sufficient condition using the concept of global information is found. Similar conditions are obtained also for the consistency of generalized likelihood ratio tests.

2. GENERAL RESULTS

We consider a statistical experiment $((\mathcal{X}^n, \mathcal{A}^n, P_\theta^n : \theta \in \Theta), n = 1, 2 \ldots)$ where $(\mathcal{X}^n, \mathcal{A}^n, P_\theta^n)$ are products of a sample component probability space $(\mathcal{X}, \mathcal{A}, P_\theta)$ satisfying the identifiability condition $P_{\theta_1} \neq P_{\theta_2}$ for $\theta_1 \neq \theta_2$. The experiment provides random samples $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ defined by sample probability spaces $(\mathcal{X}^n, \mathcal{A}^n, P_{\theta_0}^n)$ where $\theta_0 \in \Theta$ is an unknown true parameter. We restrict ourselves to asymptotic properties of the experiment for $n \to \infty$.

We are interested in the amount of information $I_{\theta_0}(S)$ which the experiment provides asymptotically about the occurrence of the unknown parameter $\theta_0 \in \Theta$ in a given parameter set $S \subset \Theta$.

The attention is focused on experiments satisfying mild regularity conditions. The parameter space Θ is assumed to be a subset of the Euclidean space R^m and the distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ dominated by a σ-finite measure μ, with densities

$$p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$$

such that the function $(\theta, x) \to p_\theta(x)$ is measurable and the random function $\theta \to p_\theta(X_1)$ separable and a. s. continuous. Then, for every open or closed $S \subset \Theta$,

$$f_n(\boldsymbol{X}_n, S) = \inf_{\theta \in S} -\frac{1}{n} \sum_{k=1}^{n} \ln p_\theta(X_k), \quad \ln 0 = -\infty, \tag{1}$$

may be considered measurable in $\boldsymbol{X}_n$ (cf. Liese and Vajda [10]). We shall work with the random variables

$$f_n(S) = f_n(\boldsymbol{X}_n, S) \quad \text{and} \quad f_n(\theta) = f_n(\{\theta\}) = -\frac{1}{n}\sum_{k=1}^{n} \ln p_\theta(X_k) \tag{2}$$

for open sets S, their complements $S^c = \Theta - S$ and points $\theta \in \Theta$.

For every subset $S \subset \Theta$, the random sample $\boldsymbol{X}_n$ provides some evidence about whether the true parameter θ_0 belongs to S, i. e. some "likelihood" of the occurrence of θ_0 in S. This likelihood may vary with n. To avoid the dependence on n, we shall deal with the asymptotic likelihood for $n \to \infty$. Consider for all open or closed $S \subset \Theta$ the limits

$$\mathcal{H}_{\theta_0}(S) = \begin{cases} \lim_n \mathsf{E}\, f_n(S) & \text{if} \quad S \neq \emptyset \\ \infty & \text{if} \quad S = \emptyset. \end{cases} \tag{3}$$

The asymptotic likelihood of the event $\theta_0 \in S$ is proportional to the limits $-\mathcal{H}_{\theta_0}(S)$ and $\mathcal{H}_{\theta_0}(S^c)$. We put

$$I_{\theta_0}(S) = \mathcal{H}_{\theta_0}(S^c) - \mathcal{H}_{\theta_0}(S). \tag{4}$$

Definition. The expression (4) is the global information in the experiment about the event $\theta_0 \in S$ for open or closed $S \subset \Theta$ such that the right-hand side is well-defined in the extended real line by (3).

The following statement follows directly from the definition.

Theorem 1. The global information $I_{\theta_0}(S)$ is monotone in the sense that

$$I_{\theta_0}(S_1) \leq I_{\theta_0}(S_2) \tag{5}$$

for all $S_1 \subset S_2 \subset \Theta$ (including $S_1 = \emptyset$) such that the global information exist, and skew-symmetric in the sense that

$$I_{\theta_0}(S^c) = -I_{\theta_0}(S) \tag{6}$$

for all $S \subset \Theta$ (including $S = \emptyset$) such that the global information exist.

Proof. If S_1, S_2 satisfy the monotonicity assumed in (5) then (1), (2) imply $\mathsf{E}\, f_n(S_1) \geq \mathsf{E}\, f_n(S_2)$ and $\mathsf{E}\, f_n(S_1^c) \leq \mathsf{E}\, f_n(S_2^c)$ for nonvoid S_1 and S_2^c. Then $\mathcal{H}_{\theta_0}(S_1) \geq \mathcal{H}_{\theta_0}(S_2)$ and $\mathcal{H}_{\theta_0}(S_1^c) \leq \mathcal{H}_{\theta_0}(S_2^c)$ and these inequalities remain true also in the void case. Thus (5) follows from (4). The relation (6) follows directly from Definition. □

In the next theorem we show that the definition of global information is applicable to all open or closed sets S under a mild regularity of the experiment. The theorem is based on the following simple lemma.

Lemma 1. Let $\emptyset \neq S \subset \Theta$ be open or closed with

$$-\infty < \inf_n \mathsf{E}\, f_n(S) \leq \sup_n \mathsf{E}\, f_n(S) < \infty. \tag{7}$$

Then (3) holds and the limit $\mathcal{H}_{\theta_0}(S)$ satisfies the relation

$$\mathcal{H}_{\theta_0}(S) = \sup_n \mathsf{E}\, f_n(S).$$

Proof. Consider random vectors $\boldsymbol{Y}_k = (Y_1, \dots, Y_k)$ and $\boldsymbol{Z}_n = (Z_1, \dots, Z_n)$ defined for arbitrary k, n by $\boldsymbol{X}_{k+n} = (\boldsymbol{Y}_k, \boldsymbol{Z}_n)$ and a subset S satisfying the assumptions. It follows from (1)

$$(k+n)\, f_{k+n}(\boldsymbol{X}_{k+n}, S) \geq k\, f_k(\boldsymbol{Y}_k, S) + n\, f_n(\boldsymbol{Z}_n, S) \tag{8}$$

and from the i. i. d. property of the components of $\boldsymbol{X}_k$, $\boldsymbol{X}_n$ and $\boldsymbol{X}_{k+n}$

$$\mathsf{E}\, f_k(\boldsymbol{Y}_k, S) = \mathsf{E}\, f_k(\boldsymbol{X}_k, S) \quad \text{and} \quad \mathsf{E}\, f_n(\boldsymbol{Z}_n, S) = \mathsf{E}\, f_n(\boldsymbol{X}_n, S).$$

Therefore it holds for all k, $n \geq 1$

$$(k+n)\, \mathsf{E}\, f_{k+n}(S) \geq k\, \mathsf{E}\, f_k(S) + n\, \mathsf{E}\, f_n(S).$$

By a well-known lemma of mathematical analysis (cf. e. g. Lemma 2 on p. 112 of Gallager [6]), every bounded sequence with this property is convergent, i. e. in our case (3) holds, and the limit $\mathcal{H}_{\theta_0}(S)$ fulfils the stated relation. □

Theorem 2. Let $-\infty < \mathsf{E}\, f_1(\Theta) \leq \mathsf{E}\, f_1(\theta) < \infty$ for all $\theta \in \Theta$. Then the condition (7) of Lemma 1 holds for all nonvoid open or closed subsets $S \subset \Theta$. Consequently the global information $I_{\theta_0}(S)$ is well-defined by (3) and (4) for all open or closed subsets $S \subset \Theta$. If $S_1 \subset S_2$ are such subsets and $S_1 \neq \emptyset$, $S_2 \neq \Theta$ then the relation (5) can be precised as follows

$$-\infty = I_{\theta_0}(\emptyset) < I_{\theta_0}(S_1) \leq I_{\theta_0}(S_2) < I_{\theta_0}(\Theta) = \infty. \tag{9}$$

Proof. Clear from Lemma 1 and Theorem 1. □

The next assertion extends the result of Theorem 2.

Theorem 3. Under the assumptions of Theorem 2 it holds for all S considered there

$$\lim_n f_n(S) = \mathcal{H}_{\theta_0}(S) \quad \text{a. s.} \tag{10}$$

so that

$$I_{\theta_0}(S) = \lim_n [f_n(S^c) - f_n(S)] \quad \text{a. s.} \tag{11}$$

Proof. Consider an arbitrary natural r, and define for all $n \geq r$ and S under consideration

$$f_n^{(r)}(S) = \binom{n}{r}^{-1} \sum_{\kappa_r} \inf_{\theta \in S} -\frac{1}{r} \sum_{j=1}^{r} \ln p_\theta(X_{k_j})$$

where the summation extends over all $\kappa_r = \{k_1, \dots, k_r\} \subset \{1, \dots, n\}$. According to Berk [2] and Perlman [13], the sequence $(f_n^{(r)}(S) : n = r, r+1, \dots)$ forms a reversed martingale. Since

$$\mathsf{E}\, f_n^{(r)}(S) = \mathsf{E}\, f_r(S)$$

and the reversed martingale is ergodic, the convergence theorem for reversed martingales implies $f_n^{(r)}(S) \to \mathsf{E}\, f_r(S)$ a. s. Further, for every $\theta \in S$ it holds $f_n(\theta) = f_n^{(r)}(\theta) \geq f_n^{(r)}(S)$ so that $f_n(S) \geq f_n^{(r)}(S)$ and, consequently,

$$\text{liminf}_n\, f_n(S) \geq \mathsf{E}\, f_r(S) \quad \text{a. s.}$$

On the other hand, by Lemma 1, the limit relation (3) holds. Taking the limit for $r \to \infty$ we obtain from (3) and from the last relation

$$\text{liminf}_n\, f_n(S) \geq \mathcal{H}_{\theta_0}(S) \quad \text{a. s.} \tag{12}$$

Since

$$f_n(S) \geq \frac{1}{n} \sum_{k=1}^{n} f_1(X_k, S) \geq \frac{1}{n} \sum_{k=1}^{n} f_1(X_k, \Theta)$$

where $\mathsf{E} f_1(X_k, \Theta) = \mathsf{E} f_1(\Theta)$ is assumed to be finite, the sequence $f_n(S)$ has an integrable minorant. Therefore, by the Fatou–Lebesgue theorem (cf. pp. 125 and 162 in Loéve [12]),

$$\mathsf{E}\, \text{liminf}_n\, f_n(S) \leq \mathcal{H}_{\theta_0}(S).$$

Consequently, the inequality in (12) cannot be strict with a positive probability, i. e.

$$\text{liminf}_n\, f_n(S) = \mathcal{H}_{\theta_0}(S) \quad \text{a. s.} \tag{13}$$

Finally, if $\theta_* \in S$ then $f_n(S) \leq f_n(\theta_*)$. As the strong law of large numbers implies $f_n(\theta_*) \to H_{\theta_0}(\theta_*)$ a. s., it holds

$$\text{limsup}_n\, f_n(S) \leq H_{\theta_0}(\theta_*) < \infty \quad \text{a. s.}$$

Therefore all but finitely many terms of the sequence $f_n(S)$ are a. s. bounded. By the same method as used in the proof of Lemma 1, it follows from here and from (8) that the sequence $f_n(S)$ is a. s. convergent. Hence (13) is equivalent to (10). □

It follows from (11) that the global information $I_{\theta_0}(S)$ is the a. s. limit of the difference $T_{S,n} - T_{S^c,n}$ of the generalized likelihood ratio test statistics defined by the formula

$$T_{S,n} = \frac{1}{n} \log \frac{\sup\limits_{\theta \in S} \prod\limits_{k=1}^{n} p_\theta(X_k)}{\sup\limits_{\theta \in \Theta} \prod\limits_{k=1}^{n} p_\theta(X_k)} = f_n(\Theta) - f_n(S).$$

$T_{S,n}$ or $T_{S^c,n}$ is the generalized likelihood ratio test statistic in testing the hypothesis "$\theta_0 \in S$" against the alternative "$\theta_0 \in S^c$", or testing the hypothesis "$\theta_0 \in S^c$" against the alternative "$\theta_0 \in S$", respectively. This provides an alternative motivation of our Definition.

In the sequel we consider under the assumptions of Theorem 2 the following modified asymptotic representation

$$I_{\theta_0}(S) = \lim\nolimits_n (g_n(S^c) - g_n(S)) \quad \text{a. s.}, \tag{14}$$

where

$$g_n(S) = \inf_{\theta \in S} g_n(\theta)$$

for

$$g_n(\theta) = g_n(\boldsymbol{X}_n, \theta) = \frac{1}{n} \sum_{k=1}^{n} \ln \frac{p_{\theta_0}(X_k)}{p_\theta(X_k)}, \quad \theta \in \Theta.$$

Note that the difference $g_n(S^c) - g_n(S) = T_{S,n} - T_{S^c,n}$ has been used already by Kullback [8] as an operational characteristics of the generalized likelihood ratio tests.

By the strong law of large numbers, under the assumptions of Theorem 2 it holds for every $\theta \in \Theta$

$$\lim\nolimits_n g_n(\theta) = I(\theta_0; \theta) \quad \text{a. s.}, \tag{15}$$

where

$$I(\theta_0, \theta) = \int p_{\theta_0} \ln \frac{p_{\theta_0}}{p_\theta} \, \mathrm{d}\mu \tag{16}$$

is the Kullback's I-divergence of models p_{θ_0} and p_θ. We shall be interested in conditions on the family $\{P_\theta : \theta \in \Theta\}$ and open or closed sets $S \subset \Theta$ under which one can interchange the lim and inf in (14), i. e. to establish the relations

$$\lim\nolimits_n \inf\nolimits_S g_n(\theta) = \inf\nolimits_S \lim\nolimits_n g_n(\theta) = \inf\nolimits_S I(\theta_0; \theta) \quad \text{a. s.} \quad \text{(cf. (15))}. \tag{17}$$

Such conditions are important since, by inserting (17) in (14), one obtains the following simple formula for the global information

$$\begin{aligned} I_{\theta_0}(S) &= \inf_{\theta \subset S^c} I(\theta_0; \theta) - \inf_{\theta \in S} I(\theta_0; \theta) \\ &= \begin{cases} \inf\limits_{\theta \in S^c} I(\theta_0; \theta) & \text{if} \quad \theta_0 \in S \\ -\inf\limits_{\theta \in S} I(\theta_0; \theta) & \text{if} \quad \theta_0 \notin S. \end{cases} \end{aligned} \tag{18}$$

Our aim is to justify this formula by finding sufficient conditions for (17). The following two lemmas are obvious.

Lemma 2. If the assumptions of Theorem 2 are satisfied and the convergence in (15) is uniform on Θ then (18) holds for all open or closed $S \subset \Theta$.

Remark 1. The experiments with the convergence in (15) uniform on Θ are relatively rare. Such experiments can be obtained e. g. from the experiments with the locally uniform convergence in (15), by restricting the parameter space to bounded closed subsets $\Theta_* \subset \Theta$. As well known, if Θ is open and convex and $g_n(\theta)$ are a. s. convex on Θ then (cf. e. g. Theorem 10.8 of Rockafellar [19]), the convergence in (15) is locally uniform on Θ and $I(\theta_0;\theta)$ is convex in the variable $\theta \in \Theta$. Typical (e. g. exponential) statistical experiments fulfil these conditions.

Remark 2. The compact parameter sets Θ_* figuring in Remark 1 can be employed also in the framework of general unrestricted models. Namely, let S be an open set containing the true parameter θ_0 and contained in a compact Θ_* from the interior of Θ. Then in typical examples of statistical experiments the relative complements

$$S_*^c = S^c \cap \Theta_* = \Theta_* - S \tag{19}$$

satisfy the relation

$$\lim_n \inf_{S^c} g_n(\theta) = \lim_n \inf_{S_*^c} g_n(\theta) \quad \text{a. s.} \tag{20}$$

Since the closures $\overline{S}$ and $\overline{S_*^c}$ are compacts contained in the interior of Θ, Remark 1 implies that the convergence in (15) is uniform on S and S_*^c in all models with the convergence in (15) locally uniform on Θ.

Lemma 3. If the assumptions of Theorem 2 hold and the convergence in (15) is locally uniform on Θ then

$$I_{\theta_0}(S) = \inf_{\theta \in S_*^c} I(\theta_0;\theta) \tag{21}$$

for all S and S_*^c considered in (19) and satisfying (20).

Next we formulate a stronger result for open balls

$$S_r(\theta_0) = \{\theta \in R^m : \|\theta - \theta_0\| < r\}, \quad r > 0,$$

with surfaces

$$\mathcal{S}_r(\theta_0) = \{\theta : \|\theta - \theta_0\| = r\}.$$

Theorem 4. If the assumptions of Theorem 2 hold, Θ is open and convex, and $g_n(\theta)$ are a. s. convex on Θ, then for every open ball $S_r(\theta_0)$ with the closure $\overline{S}_r(\theta_0)$ contained in Θ

$$I_{\theta_0}(S_r(\theta_0)) = \inf_{\theta \in \mathcal{S}_r(\theta_0)} I(\theta_0;\theta). \tag{22}$$

Proof. It suffices to prove that (20) holds for $S = S_r(\theta_0)$ and $\Theta_* = \overline{S}_r(\theta_0)$. Since Θ is assumed to be open, $\overline{S}_r(\theta_0)$ is contained in the interior of Θ. Further, for S and Θ_* under consideration

$$S_*^c = \Theta_* - S = \mathcal{S}_r(\theta_0).$$

Hence (20) will be proved if for every random sequence $\theta_n = \theta_n(\boldsymbol{X}_n) \in R^m$ with $\|\theta_n - \theta_0\| \geq r$ and for

$$\hat{\theta}_n = \theta_0 + \frac{r(\theta_n - \theta_0)}{\|\theta_n - \theta_0\|} \in \mathcal{S}_r(\theta_0)$$

we prove the relation

$$\operatorname{liminf}_n (g_n(\theta_n) - g_n(\hat{\theta}_n)) \geq 0 \quad \text{a.s.} \tag{23}$$

To this end fix n and consider the segment $\theta(t) = \theta_0 + r\,t(\theta_n - \theta_0) \in R^m$ for $0 < t < 1/r$. The assumed convexity of $g_n(\theta)$ together with the identity $g_n(\theta_0) = 0$ implies

$$g_n(\theta(t)) \leq (1 - r\,t)\,g_n(\theta_0) + r\,t\,g_n(\theta_n) = r\,t\,g_n(\theta_n).$$

The unique point $\theta(t)$ of the segment belonging to $\mathcal{S}_r(\theta_0)$ corresponds to $t_n = 1/\|\theta_n - \theta_0\| \leq 1/r$, i.e. $\theta(t_n) = \hat{\theta}_n$. It follows from here that $g_n(\hat{\theta}_n) \leq r\,t_n\,g_n(\theta_n) \leq g_n(\theta_n)$, i.e. $g_n(\theta_n) - g_n(\hat{\theta}_n) \geq 0$, provided $g_n(\hat{\theta}_n) \geq 0$. Thus (23) follows from the relation

$$\operatorname{liminf}_n g_n(\hat{\theta}_n) > 0 \quad \text{a.s.} \tag{24}$$

By Remark 1, $I(\theta_0;\theta)$ is continuous on Θ and positive on $\Theta - \{\theta_0\}$. Therefore its infimum on the compact $\mathcal{S}_r(\theta_0)$ is positive. Also the convergence in (15) is locally uniform on Θ and thus uniform on all compact subsets of Θ. It follows from here

$$\operatorname{limsup}_n \left|\inf_{\mathcal{S}_r(\theta_0)} g_n(\theta) - \inf_{\mathcal{S}_r(\theta_0)} I(\theta_0;\theta)\right| \leq \lim_n \sup_{\mathcal{S}_r(\theta_0)} |g_n(\theta) - I(\theta_0;\theta)| = 0 \quad \text{a.s.}$$

which implies (24). □

By means of Theorem 4 one can clarify relation between the global information $I_{\theta_0}(S)$ and the Bahadur exact slopes (see Bahadur [1]). We do not go into details here. In the sequel we clarify the relation to the local Fisher information $\mathcal{I}_{\theta_0}$ mentioned in the Introduction. Suppose that θ_0 is from the interior of Θ and consider an experiment for which the gradient

$$\nabla \ln p_{\theta_0} = \left(\frac{\partial}{\partial \theta_1} \ln p_\theta, \ldots, \frac{\partial}{\partial \theta_m} \ln p_\theta\right)_{\theta=\theta_0}$$

exists and the Fisher information matrix

$$\mathcal{I}_{\theta_0} = \int (\nabla \ln p_{\theta_0})^t\,(\nabla \ln p_{\theta_0})\,p_{\theta_0}\,\mathrm{d}\mu$$

is positive definite. Let the experiment satisfy also the assumptions of Theorem 4 and the asymptotic relation

$$I(\theta_0;\theta) = \frac{1}{2}(\theta - \theta_0)\,\mathcal{I}_{\theta_0}(\theta - \theta_0)^t + o(\|\theta - \theta_0\|^2)$$

for $\theta \to \theta_0$.

Theorem 5. Under the above considered assumptions it holds asymptotically, for $r \downarrow 0$,

$$I_{\theta_0}(S_r(\theta_0)) = \frac{1}{2} r^2 \omega_0 \, \mathcal{I}_{\theta_0} \omega_0^t + o(r^2), \tag{25}$$

where ω_0 minimizes the quadratic form $\omega \mathcal{I}_{\theta_0} \omega^t$ on the surface $\mathcal{S}_1 = \overline{S}_1(0) - S_1(0)$ of the unit sphere centered at 0, i. e. $\lambda(\theta_0) = \omega_0 \mathcal{I}_{\theta_0} \omega_0^t$ is the smallest eigenvalue of $\mathcal{I}_{\theta_0}$.

Proof. All spheres $S = S_r(\theta_0)$ with sufficiently small $r > 0$ satisfy the assumptions of Theorem 4. By Remark 2, $I(\theta_0, \theta)$ is convex and consequently continuous in the variable $\theta \in \Theta$. By (22) it holds for all sufficiently small $r > 0$

$$I_{\theta_0}(S_r(\theta_0)) = \inf_{\omega \in S_1} I(\theta_0; \theta_0 + r\omega) = I(\theta_0; \theta_0 + r\omega_r)$$

where the minimizing $\omega_r \in \mathcal{S}_1$ exists. Further, asymptotically for $r \downarrow 0$

$$I(\theta_0, \, \theta_0 + r\omega) = \frac{1}{2} r^2 \omega \mathcal{I}_{\theta_0} \omega^t + o(r^2), \quad \omega \in \mathcal{S}_1.$$

Since the last asymptotic formula holds uniformly for all ω from the compact $\mathcal{S}_1$, the points ω_r tend to the above defined ω_0. Consequently $\omega_r \mathcal{I}_{\theta_0} \omega_r^t$ tends to $\omega_0 \mathcal{I}_{\theta_0} \omega_0^t$ and (25) follows from the relation

$$I_{\theta_0}(S_r(\theta_0)) = \frac{1}{2} r^2 \omega_r \mathcal{I}_{\theta_0} \omega_r^t + o(r^2). \qquad \square$$

Let us note that Theorem 4 can be extended to arbitrary bounded closed or open sets S with the closure $\overline{S}$ contained in Θ. The spherical surface $\mathcal{S}_r(\theta_0)$ figuring implicitly in (22) is in this case replaced by the boundary $\mathcal{S} = \overline{S} - S^0$ where S^0 denotes the interior of S.

3. EXAMPLES

First we illustrate the general theory by experiments with discrete and continuous sample spaces $\mathcal{X}$.

Example 1 (Bernoulli experiment). Consider the experiment defined by

$$\mathcal{X} = \{0,1\} \quad \text{and} \quad P_\theta(\{x\}) = p_\theta(x) = \theta^x (1-\theta)^{1-x} \quad \text{for} \quad x \in \{0,1\}, \ \theta \in \Theta = (0,1).$$

Here the sample $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ is i. i. d. by the Bernoulli law $B(\theta)$ with $\theta = \theta_0 \in (0,1)$. It holds

$$\begin{aligned} f_n(\theta) &= -\frac{1}{n} \sum_{k=1}^n \ln \theta^{X_k} (1-\theta)^{1-X_k} \\ &= -\frac{1}{n} \left[\sum_{k=1}^n \mathbf{1}_{\{1\}}(X_k) \ln \theta + \left(n - \sum_{k=1}^n \mathbf{1}_{\{1\}}(X_k) \right) \ln(1-\theta) \right] \\ &= \theta_n \ln \frac{1}{\theta} + (1-\theta_n) \ln \frac{1}{1-\theta}, \end{aligned}$$

where

$$\theta_n = \theta_n(\boldsymbol{X}_n) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{1\}}(X_k).$$

By definition,

$$f_n(S) = \inf_{\theta \in S}\left[\theta_n \ln\frac{1}{\theta} + (1-\theta_n)\ln\frac{1}{1-\theta}\right]$$

for every subset $S \subset (0,1)$ and

$$I(\theta_0;\theta) = \theta_0 \ln\frac{\theta_0}{\theta} + (1-\theta_0)\ln\frac{1-\theta_0}{1-\theta}$$

for every $\theta \in (0,1)$. Since the functions $1/\theta$ and $1/(1-\theta)$ are logarithmically convex in the domain $\theta \in (0,1)$, the functions $f_n(\theta)$ and $I(\theta_0;\theta)$ are convex in the same domain. Further, the Fisher information is given by the formula

$$\mathcal{I}_{\theta_0} = \frac{1}{\theta_0(1-\theta_0)}$$

and, for $r \downarrow 0$,

$$I(\theta_0;\theta_0 \pm r) = \frac{r^2}{2}\mathcal{I}_{\theta_0} + o(r^2).$$

We see that the assumptions of Theorem 2, Remark 1 and Theorem 5 are satisfied. Therefore, by Theorem 2 the global information $I_{\theta_0}(S)$ exists for all subsets $S \subset (0,1)$. By Theorems 1 and 2, this information satisfies the relation (6) and in the case $\emptyset \neq S_1 \subset S_2 \neq \Theta$ also (9). By Theorem 3, it satisfies the relations

$$I_{\theta_0}(S) = \lim\nolimits_n \left[\inf\nolimits_{S^c} I(\theta_n,\theta) - \inf\nolimits_S I(\theta_n,\theta)\right] \quad \text{a.s.}$$

By Theorem 4 it holds for all $0 < r < \min\{\theta_0, 1-\theta_0\}$

$$I_{\theta_0}((\theta_0 - r,\, \theta_0 + r)) = \min\{I(\theta_0;\theta_0 - r),\, I(\theta_0;\theta_0 + r)\} = \begin{cases} I(\theta_0,\theta_0 + r) & \text{if} \quad \theta_0 \le \frac{1}{2} \\ I(\theta_0,\theta_0 - r) & \text{if} \quad \theta_0 > \frac{1}{2} \end{cases}$$

By Theorem 5 it holds for $r \downarrow 0$

$$I_{\theta_0}((\theta_0 - r,\, \theta_0 + r)) = \frac{r^2}{2\theta_0(1-\theta_0)} + o(r^2).$$

Example 2 (Normal experiment). Let the experiment be defined for some $\sigma > 0$ by

$$\mathcal{X} = R \quad \text{and} \quad \frac{\mathrm{d}P_\theta((-\infty,x))}{\mathrm{d}x} = p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\theta)^2}{2\sigma^2}} \qquad \text{for } x \in R,\ \theta \in \Theta \subset R.$$

Then the sample $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ is i. i. d. by the normal law $N(\theta, \sigma^2)$ with $\theta = \theta_0 \in R$. It holds

$$\begin{aligned} f_n(\theta) &= \frac{\ln(2\pi\sigma^2)}{2} + \frac{1}{2n\sigma^2}\sum_{k=1}^{n}(X_k - \theta)^2 \\ &= \frac{\ln(2\pi\sigma^2)}{2} + \frac{Y_n}{2n} + \frac{(\overline{\boldsymbol{X}}_n - \theta)^2}{2\sigma^2} \\ &= f_n(\theta_0) + \frac{(\theta_0 - \theta)^2}{2\sigma^2} + \frac{(\overline{\boldsymbol{X}}_n - \theta_0)\,(\theta_0 - \theta)}{\sigma^2}, \end{aligned}$$

where

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \sim N(\theta_0, \sigma^2/n) \quad \text{and} \quad Y_n = \frac{1}{\sigma^2}\sum_{k=1}^{n}(X_k - \overline{\boldsymbol{X}}_n)^2 \sim \chi^2_{n-1}.$$

Further,

$$f_n(S) = \frac{\ln(2\pi\sigma^2)}{2} + \frac{Y_n}{2n} + \frac{1}{2\sigma^2}\inf_{\theta \in S}(\overline{\boldsymbol{X}}_n - \theta)^2$$

for every $S \subset R$ and

$$I(\theta_0; \theta) = \frac{(\theta_0 - \theta)^2}{2\sigma^2}$$

for every $\theta \in R$. Obviously,

$$f_n(S) \geq \frac{\ln(2\pi\sigma^2)}{2} > -\infty$$

and the functions $f_n(\theta)$ and $I(\theta_0; \theta)$ are convex in the domain $\theta \in R$. Finally, the Fisher information is constant,

$$\mathcal{I}_{\theta_0} = \frac{1}{\sigma^2},$$

and for every $r > 0$

$$I(\theta_0, \theta_0 \pm r) = \frac{r^2}{2}\mathcal{I}_{\theta_0}.$$

The assumptions of Theorem 2, Remark 2 and Theorem 5 are satisfied. Analogously as in the previous example, we obtain from Theorems 1 and 2 that the global information $I_{\theta_0}(S)$ exists for all open or closed sets $S \subset R$ and satisfies the relations (6) and (9). From Theorem 3 we get in this case

$$I_{\theta_0}(S) = \frac{1}{2}\left(\inf\nolimits_{S^c}\|\theta_0 - \theta\|^2 - \inf\nolimits_{S}\|\theta_0 - \theta\|^2\right)^2$$

and from Theorem 4

$$I_{\theta_0}((\theta_0 - r,\, \theta_0 + r)) = \frac{r^2}{2\sigma^2} = \frac{r^2}{2}\mathcal{I}_{\theta_0} \quad \text{for all } r > 0.$$

We see that the relation to Fisher's information is more concrete than in the general case considered in Theorem 5.

In the next example assumptions of Theorem 2 are not satisfied but, nevertheless, the global information $I_{\theta_0}(S)$ can be evaluated for all bounded neighborhoods S of θ_0.

Example 3 (Contaminated errorless observations). Let us consider the experiment with $\mathcal{X} = \{1, 2, \ldots\}$ and

$$\frac{P_\theta(\{x\})}{\mu_x} = p_\theta(x) = \varepsilon + (1-\varepsilon)\,\frac{\mathbf{1}_{\{\theta\}}(x)}{\mu_x} \quad \text{for } x = 1, 2, \ldots,$$

where $0 < \varepsilon < 1/2$, $\mu_x > 0$, $\mu_1 + \mu_2 + \cdots = 1$, and $\theta \in \Theta = \{1, 2, \ldots\}$. As before, the sample $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ is i. i. d. by $P^n_{\theta_0}$. In accordance with the robust statistics (see e. g. Huber [7]), the sample components can be described by the formula

$$X_k = (1 - \epsilon_k)\,\theta_0 + \epsilon_k\, Z_k,$$

where ϵ_k are the Bernoulli trials with a fixed parameter ε and the random variables Z_k are mutually and also on $\epsilon_1, \epsilon_2, \ldots$ independent, identically distributed by μ. Thus this experiment describes an errorless observation of the true parameter θ_0 contaminated at the level ε by data from a source distributed by μ. The alternative data will be called a noise.

It holds

$$f_n(\theta) = -\frac{1}{n}\sum_{k=1}^{n} \ln\left(\varepsilon + (1-\varepsilon)\,\frac{\mathbf{1}_{\{\theta\}}(X_k)}{\mu_{X_k}}\right).$$

Let us consider the entropy function

$$\begin{aligned} H_{\theta_0}(\theta) &= -\int p_{\theta_0} \ln p_\theta \,\mathrm{d}\mu \\ &= -\sum_{x=1}^{\infty} \left(\varepsilon\mu_x + (1-\varepsilon)\,\mathbf{1}_{\{\theta_0\}}(x)\right) \ln\left(\varepsilon + (1-\varepsilon)\,\frac{\mathbf{1}_{\{\theta\}}(x)}{\mu_x}\right) \\ &= \begin{cases} \ln\frac{1}{\varepsilon} - \varepsilon\mu_\theta \ln\left(1 + \frac{1-\varepsilon}{\varepsilon\mu_\theta}\right) & \text{if} \quad \theta \neq \theta_0 \\ \ln\frac{1}{\varepsilon} - (\varepsilon\mu_{\theta_0} + 1 - \varepsilon) \ln\left(1 + \frac{1-\varepsilon}{\varepsilon\mu_{\theta_0}}\right) & \text{if} \quad \theta = \theta_0. \end{cases} \end{aligned}$$

Therefore if $\theta \neq \theta_0$ then

$$I(\theta_0; \theta) = H_{\theta_0}(\theta) - H_{\theta_0}(\theta_0) = \varepsilon\mu_\theta \ln \frac{\varepsilon\mu_\theta}{\varepsilon\mu_\theta + 1 - \varepsilon} + (\varepsilon\mu_{\theta_0} + 1 - \varepsilon) \ln \frac{\varepsilon\mu_{\theta_0} + 1 - \varepsilon}{\varepsilon\mu_{\theta_0}}$$

is positive and bounded above by $(\varepsilon\mu_{\theta_0} + 1 - \varepsilon)\,(1-\varepsilon)/\varepsilon\mu_{\theta_0}$.

We are interested in open neighbourhoods $S = S_r(\theta_0) \cap \Theta$ of θ_0 or, more generally, in bounded subsets

$$S = \{j, j+1, \ldots, k\} \subset \Theta, \quad 1 \leq j \leq \theta_0 \leq k.$$

For such S

$$m(S) = \min\nolimits_S \frac{1}{\mu_\theta} > 0 \quad \text{and} \quad M(S) = \max\nolimits_S \frac{1}{\mu_\theta} < \infty.$$

The obvious relation

$$\ln(\varepsilon + (1-\varepsilon)\, m(S)) \le f_n(S) \le \ln(\varepsilon + (1-\varepsilon)\, M(S))$$

implies that $\mathsf{E}\, f_n(S)$ is uniformly bounded for all n. Therefore, by Lemma 1, $\lim_n \mathsf{E}\, f_n(S) = \mathcal{H}_\theta(S)$ exists and is finite. In this situation, by Definition, $I_{\theta_0}(S) = -\infty$ follows from the relation

$$\mathcal{H}_{\theta_0}(S^c) = \lim\nolimits_n \mathsf{E}\, f_n(S^c) = -\infty. \tag{26}$$

In the rest of this section we investigate in more detail the statistical experiments of Example 3. We shall prove (26) for experiments with infinite entropy of noise. Such a noise is unpleasant as in the resulting statistical experiments all information $I_{\theta_0}(S)$ about bounded parameter sets attain the minimum possible value $I_{\theta_0}(\emptyset)$ (and the information about complements $I_{\theta_0}(S^c)$ attain the maximum possible value $I_{\theta_0}(\Theta)$). Thus the noise is not permitting a reasonable localization of the unknown value θ_0 in the given parameter space. We shall see in Section 5 that in such situations the maximum likelihood estimator is inconsistent.

Let us first notice that (26) follows from the relation

$$\lim\nolimits_n \mathsf{E}\, f_n(\boldsymbol{X}_{(n)}) = -\infty \quad \text{for } \boldsymbol{X}_{(n)} = \max\{X_1, \ldots, X_n\}. \tag{27}$$

To this end it suffices to take into account for $k = \max S$ the inequality $f_n(S^c) \le f_n((k,\infty))$ and relations

$$f_n((k,\infty)) \le \begin{cases} f_n(\boldsymbol{X}_{(n)}) & \text{if} \quad \boldsymbol{X}_{(n)} > k \\ f_n(k+1) \le \ln\varepsilon & \text{if} \quad \boldsymbol{X}_{(n)} \le k, \end{cases}$$

and for every $k \ge \theta_0$ the relations

$$\begin{aligned} \mathsf{P}(\boldsymbol{X}_{(n)} \le k) &= \left(\sum_{x=1}^{k} P_{\theta_0}(x)\right)^n = \left(\varepsilon \sum_{x=1}^{k} \mu_x + 1 - \varepsilon\right)^n \\ &= (\varepsilon\, F(k) + 1 - \varepsilon)^n = [1 - \varepsilon(1 - F(k))]^n, \end{aligned} \tag{28}$$

where $F(k)$ is the distribution function of noise.

Let for $T > 0$ there exist $k_0 = k_0(T,n)$ such that

$$\mu_x < e^{-nT} \mu_{k_0} \quad \text{for all } x > k_0. \tag{29}$$

Obviously, for all T large enough we obtain $k_0 > \theta_0$. Further, the assumed inequality $\varepsilon < 1 - \varepsilon$ implies for every $\boldsymbol{X}_n$

$$\begin{aligned} \frac{1}{n}\sum_{k=1}^{n} \ln p_{\boldsymbol{X}_{(n)}}(X_k) &= \frac{1}{n}\sum_{k=1}^{n} \begin{cases} \ln\varepsilon & \text{if} \quad X_k \ne \boldsymbol{X}_{(n)} \\ \ln\left(\varepsilon + \dfrac{1-\varepsilon}{\mu_{\boldsymbol{X}_{(n)}}}\right) & \text{if} \quad X_k = \boldsymbol{X}_{(n)} \end{cases} \\ &\ge \frac{n-1}{n}\ln\varepsilon + \frac{1}{n}\ln\left(\varepsilon + \frac{1-\varepsilon}{\mu_{\boldsymbol{X}_{(n)}}}\right) \\ &\ge \ln\varepsilon - \frac{1}{n}\ln\mu_{\boldsymbol{X}_{(n)}}, \end{aligned}$$

i. e.,

$$f_n(\boldsymbol{X}_{(n)}) \le -\ln\varepsilon + \frac{1}{n}\ln\mu_{\boldsymbol{X}_{(n)}}.$$

It follows from here

$$\begin{aligned}
\mathsf{E}\, f_n(\boldsymbol{X}_{(n)}) &\le -\ln\varepsilon + \frac{1}{n}\sum_{x=1}^{\infty}\mathsf{P}(\boldsymbol{X}_{(n)} = x)\ln\mu_x \\
&\le -\ln\varepsilon + \frac{1}{n}\sum_{x=k_0+1}^{\infty}\mathsf{P}(\boldsymbol{X}_{(n)} = x)\ln\mu_x \\
&\le -\ln\varepsilon + \frac{1}{n}\mathsf{P}(\boldsymbol{X}_{(n)} > k_0)\ln e^{-nT} \quad \text{(cf. (29))} \\
&= -\ln\varepsilon - T\,\mathsf{P}(\boldsymbol{X}_{(n)} > k_0).
\end{aligned}$$

Lemma 4. If the noise distribution function $F(k)$ satisfies the conditions (29) and

$$\lim_{T\to\infty}\lim_{n\to\infty} n[1 - F(k_0(T,n))] = \infty \tag{30}$$

then (27) and, consequently, (26) hold.

Proof. By (28),

$$\mathsf{P}(\boldsymbol{X}_{(n)} > k_0) = 1 - [1 - \varepsilon(1 - F(k_0))]^n$$

so that under (30)

$$\lim_{n\to\infty}\mathsf{P}(\boldsymbol{X}_{(n)} > k_0(T,n)) = 1 - e^{-\varepsilon\varphi(T)},$$

where $\lim_{T\to\infty}\varphi(T) = \infty$. The desired relation (27) follows from here and from the inequality preceding Lemma 4. □

The following result indicates that (29) and (30) represent a heavy tail condition on the noise distribution μ.

Lemma 5. The conditions (29) and (30) hold only if the entropy

$$H(\mu) = -\sum_{k=1}^{\infty}\mu_k\ln\mu_k$$

is infinite.

Proof. It follows from (29)

$$\begin{aligned}
H(\mu) &\ge -\sum_{k=1}^{k_0}\mu_n\ln\mu_n - \sum_{k=k_0+1}^{\infty}\mu_k\ln\left(e^{-n\,T}\mu_{k_0}\right) \\
&\ge (1 - F(k_0))\,n\,T.
\end{aligned}$$

Hence (30) implies $H(\mu) = \infty$. □

Example 4 (Light tailed noise). Let us consider the geometric noise $\mu_x = (1-\beta)\,\beta^{x-1}$ for $x = 1, 2, \ldots$ and $0 < \beta < 1$. The entropy

$$H(\mu) = \ln \frac{1}{1-\beta} + \frac{\beta}{1-\beta} \ln \frac{1}{\beta}$$

is finite. In this case (29) fails to hold.

Example 5 (Heavy tailed noise). Consider the logarithmic noise distributed by

$$F(k) = 1 - \frac{1}{\ln^{\beta}(e+k)}, \quad k = 0, 1, 2, \ldots$$

where $\beta > 0$. Here

$$\mu_x = F(x) - F(x-1) = \beta \int_{e+x-1}^{e+x} \frac{\mathrm{d}y}{y \ln^{1+\beta} y}, \quad x = 1, 2, \ldots \tag{31}$$

If $\beta > 1$, i. e. if the tails of logarithmic noise are not heavy enough, then $H(\mu)$ is finite so that, by Lemma 5, the conditions of Lemma 4 do not hold. Therefore we shall restrict ourselves to $0 < \beta \leq 1$. For these β the assumptions of Lemma 4 hold. Indeed, the concavity of $\Phi(t) = -t \ln t$ implies for every x

$$\begin{aligned} \Phi(\mu_x) &= \beta\,\Phi\left(\int_{e+x-1}^{e+x} \frac{\mathrm{d}y}{y \ln^{1+\beta} y}\right) + \mu_x \ln \frac{1}{\beta} \\ &\geq \beta \int_{e+x-1}^{e+x} \Phi\left(\frac{1}{y \ln^{1+\beta} y}\right) \mathrm{d}y + \mu_x \ln \frac{1}{\beta}. \end{aligned}$$

Therefore

$$H(\mu) \geq \beta \int_0^{\infty} \Phi\left(\frac{1}{y \ln^{1+\beta} y}\right) \mathrm{d}y + \ln \frac{1}{\beta}.$$

The integral equals

$$\begin{aligned} \int_e^{\infty} \frac{\ln(y \ln^{1+\beta} y)}{y \ln^{1+\beta} y}\, \mathrm{d}y &= \int_e^{\infty} \frac{\mathrm{d}y}{y \ln^{\beta} y} + (1+\beta) \int_e^{\infty} \frac{\ln \ln y}{y \ln^{1+\beta} y}\, \mathrm{d}y \\ &= \frac{1}{1-\beta} \left[\ln^{1-\beta} y\right]_e^{\infty} + \frac{1+\beta}{\beta^2} \end{aligned}$$

so that the necessary condition $H(\mu) = \infty$ for (29) and (30) is satisfied. Further, (31) implies for every x

$$\frac{\beta}{(e+x-1) \ln^{1+\beta}(e+x-1)} \leq \mu_x \leq \frac{\beta}{(e+x) \ln^{1+\beta}(e+x)}$$

so that (29) holds for some $k_0 = k_0(T, n) \leq (e^{Tn} - e)$. It follows from here

$$1 - F(k_0(T, n)) \geq 1 - F(e^{Tn} - e) = \frac{1}{(Tn)^{\beta}}$$

which implies (30).

4. CONSISTENCY OF GENERALIZED LIKELIHOOD RATIO TESTS

Throughout this section we consider a statistical experiment of the above considered type satisfying the assumptions of Theorem 2. A null hypothesis $\mathcal{H}_0$ will be represented by an open or closed subset $\emptyset \neq S \subset \Theta$. This hypothesis is assumed to be tested against an alternative $\mathcal{H}_1$ represented by an open or closed subset $\emptyset \neq S^* \subset S^c = \Theta - S$.

The test is a sequence of measurable mappings $\tau_n = \tau_n(\boldsymbol{X}_n) \in \{0, 1\}$ where $\tau_n = 1$ means that $\mathcal{H}_0$ is rejected. Every test τ_n can be characterized by a family of random sequences

$$\pi_n(\theta_0) = \mathsf{P}(\tau_n = 1), \quad \theta_0 \in \Theta.$$

Members of the families

$$(\alpha_n(\theta) = \pi_n(\theta) : \theta \in S) \quad \text{and} \quad (\beta_n(\theta) = 1 - \pi_n(\theta) : \theta \in S^*)$$

are called first and second kind errors respectively, and $(\pi_n(\theta) : \theta \in S^*)$ is a power function. The test is consistent if

$$\lim_n \pi_n(\theta_0) = 0 \quad \text{for } \theta_0 \in S \tag{32}$$

and

$$\lim_n \pi_n(\theta_0) = 1 \quad \text{for } \theta_0 \in S^*. \tag{33}$$

In the Bayes theory the consistency leads to the asymptotically vanishing average errors

$$e_n = \int_\Theta [\mathbf{1}_S(\theta)\, \alpha_n(\theta) + \mathbf{1}_{S^*}(\theta)\, \beta_n(\theta)]\, \mathrm{d}W(\theta)$$

taken with respect to an arbitrary prior distribution W. The Neyman–Pearson theory is interested in families of asymptotically ε-level tests $(\tau_n^{(\varepsilon)} : 0 < \varepsilon < 1)$, i.e. in the tests $\tau_n^{(\varepsilon)}$ satisfying the condition

$$\operatorname{limsup}_n \pi_n^{(\varepsilon)}(\theta_0) \leq \varepsilon \quad \text{for every } \theta_0 \in S \text{ and } 0 < \varepsilon < 1.$$

Such a family is said to be consistent if (33) holds for all $0 < \varepsilon < 1$ with π_n replaced by $\pi_n^{(\varepsilon)}$.

The consistency of a test in the sense of (32), (33) implies that the family of identical tests $(\tau_n^{(\varepsilon)} = \tau_n : 0 < \varepsilon < 1)$ satisfies the assumptions of Neyman–Pearson theory and is consistent in the sense considered there. If, conversely, $(\tau_n^{(\varepsilon)} : 0 < \varepsilon < 1)$ is a consistent Neyman–Pearson family then under mild restrictions there exists a sequence $\varepsilon_n \downarrow 0$ such that (32) and (33) hold for the test $\tau_n = \tau_n^{(\varepsilon_n)}$. Thus the concept of consistency represented by (32) and (33) is relevant in the Bayes as well as Neyman–Pearson testing theory (cf. Strasser [21] and Lehman [9]).

A generalized likelihood ratio test (GLRT) of $\mathcal{H}_0 \equiv S$ is described by a sequence of pairs (T_n, t_n) where $t_n \in R$ and

$$T_n = T_n(\boldsymbol{X}_n) = \frac{1}{n} \ln \frac{\sup_S \prod_{i=1}^n p_\theta(X_i)}{\sup_\Theta \prod_{i=1}^n p_\theta(X_i)}$$

or, equivalently,

$$T_n = f_n(\Theta) - f_n(S).$$

It is defined by $\tau_n = \mathbf{1}_{(-\infty,t_n)}(T_n)$, i. e. $\mathcal{H}_0$ is rejected if and only if $T_n < t_n$. We see that the test statistic T_n is fixed so that the test and its universal characteristics

$$\pi_n(\theta_0) = \mathsf{P}(f_n(\Theta) - f_n(S) < t_n) \tag{34}$$

depend solely on the critical value t_n.

The consistency of GLRT's can be characterized by means of global statistical information as follows.

Theorem 6. Let the hypothesis S satisfy a. s. the relations $\lim_n f_n(S) = \mathcal{H}_{\theta_0}(S)$ and $\lim_n f_n(S^c) = \mathcal{H}_\theta(S^c)$. If there exist $\theta \in S$ and $\theta^* \in S^*$ such that

$$I_\theta(S) < 0 \quad \text{and} \quad I_{\theta^*}(S) > 0 \tag{35}$$

then no GLRT is consistent. If for all $\theta \in S$ and $\theta^* \in S^*$

$$I_\theta(S) > 0 \quad \text{and} \quad I_{\theta^*}(S) < 0$$

then every GLRT with $\lim_n t_n = 0$ is consistent.

Proof. By assumptions,

$$f_n(\Theta) = \min\{f_n(S),\, f_n(S^c)\} \to \min\{\mathcal{H}_{\theta_0}(S), \mathcal{H}_{\theta_0}(S^c)\} \quad \text{a. s.}$$

Hence

$$f_n(\Theta) - f_n(S) \to J_{\theta_0}(S) \quad \text{a. s.},$$

where

$$J_\theta(S) = \min\{0,\, \mathcal{H}_\theta(S^c) - \mathcal{H}_\theta(S)\} = \min\{0, I_\theta(S)\}.$$

Therefore (34) implies that $\limsup_n t_n \le J_{\theta_0}(S)$ is necessary for $\pi_n(\theta_0) \to 0$ and $\liminf_n t_n \ge J_{\theta_0}(S)$ is necessary for $\pi_n(\theta_0) \to 1$. It is clear from here and from (32), (33) that (35) contradicts the consistency of any GLRT. The sufficiency of the condition formulated in the theorem follows from the fact that $\limsup_n t_n < J_{\theta_0}(S)$ is sufficient for $\pi_n(\theta_0) \to 0$ and $\liminf_n t_n > J_{\theta_0}(S)$ is sufficient for $\pi_n(\theta_0) \to 1$. □

Example 6. In the normal experiment of Example 2 we get for $S = (a, b)$, $S^* = (-\infty, a) \cup (b, \infty)$ and $\theta_0 \in S$, $\theta_1 \in S^*$

$$I_{\theta_0}(S) = I_{\theta_0}((\theta_0 - r_0, \theta_0 + r_0)) = \frac{r_0^2}{2\sigma^2}$$

and

$$I_{\theta_1}(S) = -I_{\theta_1}(S^c) = -\min\{I_{\theta_1}((-\infty, a)),\, I_{\theta_1}((b, \infty))\} = -\frac{r_1^2}{2\sigma^2},$$

where $r_i = \min\{|\theta_i - a|,\, |\theta_i - b|\}$, $i = 0, 1$, are positive. The GLRT $\tau_n \equiv (T_n, 0)$ rejects $\mathcal{H}_0$ if and only if $\overline{\boldsymbol{X}}_n \notin S$ (cf. the formula for $T_n = f_n(\Theta) - f_n(S)$ which follows from Example 2). By Theorem 6, this test is consistent.

5. CONSISTENCY OF APPROXIMATELY MAXIMUM LIKELIHOOD ESTIMATES

In the framework of general statistical experiment we consider point estimators $\hat{\theta}_n = \theta_n(\boldsymbol{X}_n)$, i.e. sequences of measurable mappings $\mathcal{X}^n \to \Theta$. An estimator is said to be strongly consistent if $\hat{\theta}_n \to \theta_0$ a.s., i.e.,

$$\lim_n \mathbf{1}_{S_r(\theta_0)}(\hat{\theta}_n) = 1 \quad \text{a.s. for all } r > 0. \tag{36}$$

We may assume without loss of generality that $f_n(\theta)$ are uniformly bounded on Θ. Indeed, $f_n(\theta)$ can be replaced by $\tilde{f}_n(\theta) = \varphi \circ f_n(\theta)$ where $\varphi(x) = x/(1+|x|)$.

We are interested in the maximum likelihood estimators (MLE's) defined by the condition $f_n(\hat{\theta}_n) = f_n(\Theta)$ a.s. or, more generally, in the approximate MLE's (briefly, AMLE's) defined by the condition $f_n(\hat{\theta}_n) - f_n(\Theta) \to 0$ a.s.

Denote $\epsilon_n = f_n(\hat{\theta}_n) - f_n(\Theta)$. We shall need the obvious relations

$$f_n(S^c) - f_n(S) > \epsilon_n \Rightarrow \hat{\theta}_n \in S$$

and

$$f_n(S^c) - f_n(S) < -\epsilon_n \Rightarrow \hat{\theta}_n \notin S$$

for subsets $S \subset \Theta$ different from $\emptyset$ and Θ. If the left-hand side tends a.s. to $I_{\theta_0}(S)$ then $I_{\theta_0}(S_r(\theta_0)) > 0$ implies that (36) holds and $I_{\theta_0}(S_r(\theta_0)) < 0$ implies that (36) fails to hold. The following result follows directly from here.

Theorem 7. Let for all balls $S = S_r(\theta_0)$ contained in Θ the global information $I_{\theta_0}(S)$ exist and satisfy the relation $\lim_n(f_n(S^c) - f_n(S)) = I_{\theta_0}(S)$. If $I_{\theta_0}(S) > 0$ for all balls with sufficiently small $r > 0$ then all AMLE's are strongly consistent. If $I_{\theta_0}(S) < 0$ for one of these balls then no AMLE is strongly consistent.

This result is an alternative to the results on AMLE's obtained by the authors cited at the end of Section 1. The condition of existence of global information imposes stronger restriction on the model than assumed in their theorems. On the other hand, in models where the global information exists it provides considerably simpler characterization of consistency.

Example 7. In the model of errorless observations contaminated by a heavy tailed noise studied in Example 5 the assumptions of Theorem 7 hold. For the "open sphere" $S_1(\theta_0) = \{\theta_0\}$ it was proved $I_{\theta_0}(\{\theta_0\}) = -\infty$. Therefore Theorem 7 implies that no AMLE is strongly consistent in this model. In fact, it is easy to see from the proof of Lemma 4 that no AMLE is in this case consistent, even in the ordinary (non-strong) sense.

(Received September 12, 1997.)

REFERENCES

[1] R. R. Bahadur: Some Limit Theorems in Statistics. SIAM, Philadelphia 1971.

[2] R. H. Berk: Limiting behavior of posterior distributions when the model is incorrect. Ann. Math. Statist. *37* (1966), 51–58.

[3] T. M. Cover and J. B. Thomas: Elements of Information Theory. Wiley, New York 1991.

[4] M H. De Groot: Uncertainty, information and sequential experiments. Ann. Math. Statist. *33* (1962), 404–419.

[5] M. H. De Groot: Optimal Statistical Decisions. McGraw Hill, New York 1970.

[6] R. C. Gallager: Information Theory and Reliable Communication. Wiley, New York 1968.

[7] P. J. Huber: Robust Statistics. Wiley, New York 1981.

[8] S. Kullback: Information Theory and Statistics. Wiley, New York 1959.

[9] E. L. Lehman: Testing Statistical Hypotheses. Second edition. Wiley, New York 1986.

[10] F. Liese and I. Vajda: Necessary and sufficient conditions for consistency of generalized M-estimates. Metrika *42* (1995), 93–114.

[11] D. V. Lindley: On the measure of the information provided by an experiment. Ann. Math. Statist. *27* (1956), 986–1005.

[12] M. Loéve: Probability Theory. Wiley, New York 1963.

[13] M. D. Perlman: On the strong consistency of approximate maximum likelihood estimators. In: Proc. VIth Berkeley Symp. Prob. Math. Statist., 1972, pp. 263–281.

[14] J. Pfanzagl: On the measurability and consistency of minimum contrast estimators. Metrika *14* (1969), 249–272.

[15] J. Pfanzagl: Parametric Statistical Theory. Walter de Guyter, Berlin 1994.

[16] C. R. Rao: Linear Statistical Inference and its Applications. Wiley, New York 1965.

[17] A: Rényi: On the amount of information concerning an unknown parameter in a sequence of observations. Publ. Math. Inst. Hungar. Acad. Sci., Sec. A *9* (1964), 617–625.

[18] A. Rényi: Statistics and information theory. Stud. Scient. Math. Hungar. *2* (1967), 249–256.

[19] R. T. Rockafellar: Convex Analysis. Princeton Univ. Press, Princeton, N. J. 1970.

[20] H. Strasser: Consistency of maximum likelihood and Bayes estimates. Ann. Statist. *9* (1981), 1107–1113.

[21] A. Strasser: Mathematical Theory of Statistics. DeGruyter, Berlin 1985.

[22] E. Torgersen: Comparison of Statistical Experiments. Cambridge Univ. Press, Cambridge 1991.

[23] I. Vajda: Conditions equivalent to consistency of approximate MLE's for stochastic processes. Stochastic Process. Appl. *56* (1995), 35–56.

Ing. Igor Vajda, DrSc., 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