Applications of Mathematics

Igor Vajda; E. van der Meulen

Global statistical information in exponential experiments and selection of exponential models

Applications of Mathematics, Vol. 43 (1998), No. 1, 23--51

Persistent URL: http://dml.cz/dmlcz/134373

43 (1998) APPLICATIONS OF MATHEMATICS No. 1, 23–51

GLOBAL STATISTICAL INFORMATION IN EXPONENTIAL EXPERIMENTS AND SELECTION OF EXPONENTIAL MODELS[1]

Igor Vajda, Praha, and E. van der Meulen, Leuven

(Received September 17, 1997)

Abstract. The concept of global statistical information in the classical statistical experiment with independent exponentially distributed samples is investigated. Explicit formulas are evaluated for common exponential families. It is shown that the generalized likelihood ratio test procedure of model selection can be replaced by a generalized information procedure. Simulations in a classical regression model are used to compare this procedure with that based on the Akaike criterion.

Keywords: exponential families, information divergence, Fisher information, global information, Akaike criterion, model selection

MSC 2000: Primary 62B10; Secondary 62B15, 62F05

1. Introduction

We consider the standard model of asymptotic statistics, i. e. a sequence of parametrized product probability spaces $(\mathscr{X}^n, \mathscr{A}^n, P_\theta^n\colon\ \theta \in \Theta)$ for $\Theta \subset \mathbb{R}^m$. In the family $(P_\theta\colon\ \theta \in \Theta)$ of distributions corresponding to the sample size $n = 1$ all distributions are supposed to be defined by densities $p_\theta = \mathrm{d}P_\theta/\mathrm{d}\lambda$.

This model describes a statistical experiment producing a sequence of data vectors $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ with observations X_k i. i. d. by P_{θ_0} where θ_0 is the true parameter from Θ. In regular models the Fisher formula specifies the amount of information I_{θ_0} contained in one observation from this experiment. As is well known, the Fisher information measures the sensitivity of the distributions P_θ to variations of parameter θ in the neighborhood of θ_0. This information is local in the sense that I_{θ_0} cannot be affected by modifications of the distribution P_θ outside open neighborhoods of θ_0.

[1] Supported by the Czech Academy of Sciences grant 175 402 and by GA CR grant 201/96/0415.

In this paper we are interested in global measures of information reflecting the structure of the whole family $(P_\theta\colon\ \theta \in \Theta)$. Intuitively we would like to express the amount of evidence per observation provided by the experiment in favour of the hypothesis that the true value is in a given subset $S \subset \Theta$. Similarly as in the local case, this amount of evidence is called briefly information about S. In order to distinguish this information from the previously mentioned local information, we call it global information.

Vajda (1997) introduced the *global information* $I(S) = I_{\theta_0}(S)$ as the difference of asymptotic expected maximum likelihoods achieved on S and its complement, i. e. he considered

$$I_{\theta_0}(S) = \lim_{n\to\infty} \mathsf{E} \sup_{\theta\in S} \frac{1}{n} \sum_{k=1}^{n} \ln p_\theta(X_k) - \lim_{n\to\infty} \mathsf{E} \sup_{\theta\in\Theta-S} \frac{1}{n} \sum_{k=1}^{n} \ln p_\theta(X_k).$$

He showed that if

$$\mathsf{E} \inf_{\theta\in\Theta} \ln p_\theta(X_1) > -\infty \quad \text{and} \quad \mathsf{E} \ln p_\theta(X_1) < \infty \quad \text{for all } \theta \in \Theta$$

then this information is well defined by the above formulas for all $\theta_0 \in \Theta$ and all nonvoid open or closed proper subsets $S \subset \Theta$. He also proved that then, moreover,

$$I_{\theta_0}(S) = \lim_{n} \frac{1}{n} \ln \frac{\sup\limits_{\theta\in S} \prod\limits_{k=1}^{n} p_\theta(X_k)}{\sup\limits_{\theta\in S^c} \prod\limits_{k=1}^{n} p_\theta(X_k)} \quad \text{a. s.}, \tag{1}$$

and that one has at one's disposal simple necessary and sufficient conditions for consistency of maximum likelihood estimators (MLE's) and generalized likelihood ratio tests (GLRT's) in cases when $I_{\theta_0}(S)$ is available for $\theta_0 \in \Theta$ and appropriate open or closed subsets $S \subset \Theta$.

In this paper we show by using (1) that in models with exponential densities one can obtain explicit formulas for $I_{\theta_0}(S)$ when θ_0 is arbitrary and S is a set of parameters with reasonably simple boundary ∂S. In Section 3 we prove that the global information is the minimal Kullback divergence $I(\theta_0, \theta)$ achieved by $\theta \in \partial S$, with the sign + or − depending on whether θ_0 is in S or not. In Section 4 we present formulas for the divergences $I(\theta_0, \theta)$ in all common exponential families. If $\Theta \subset \mathbb{R}$ then these formulas provide the global information simply by Euclidean projections of θ_0 on ∂S.

In Section 5 we show that in exponential models the generalized likelihood ratio test of a hypothesis $S \subset \Theta$ can be formulated as a global information test based on

the statistic $I_{\hat\theta_n}(S)$ where $\hat\theta_n$ is the MLE of θ_0. By using this, we show in Section 6 that the "bottom to top" strategy of model selection using the likelihood ratio tests, applied previously in special statistical models (cf. e. g. Pötscher (1983) and Bauer et al (1988)), can in the case of an arbitrary exponential model be based on the global information statistics $I_{\hat\theta_n}(S_1), \dots, I_{\hat\theta_n}(S_M)$ where subsets $S_1 \subset \dots \subset S_M = \Theta$ represent possible submodels. The global information selection criterion formulated at the end of Section 6 in fact differs from the likelihood ratio test criteria, and also from other familiar "information criteria" (cf. Akaike (1973), Schwartz (1978), Rissanen (1979), Sahamoto et al (1986), Nishii (1988), Speed and Yu (1993), Berlinet and Francq (1994), Rydén (1995), Vieu (1995)). We compare this criterion with the Akaike criterion by means of simulations in a classical nonlinear regression model. Note that the growing need for new information-theoretic methods of reduction of complexity of regression models has been stressed not long ago e. g. by E. Ronchetti in his talk at the 12th Prague Conference, cf. Ronchetti (1994).

2. Basic concepts and results

Exponential families of distributions on $\mathscr{X}$ are described by the densities

$$p_\theta(x) = a(\theta)\, b(x)\, \mathrm{e}^{T(x)\, Q(\theta)^t}$$

with respect to a dominating measure λ on $\mathscr{X}$ where $Q\colon \Theta \to \mathbb{R}^m$ is continuous and invertible, $T\colon \mathscr{X} \to \mathbb{R}^m$ is measurable, and ^{t} denotes the vector transpose (cf. Lehmann (1986)). These densities can be simplified by the reparametrization $Q(\theta) \to \theta$ and modification of the dominating measure and the density factor $b(x)$ into the form

$$p_\theta(x) = \frac{\mathrm{e}^{T(x)\, \theta^t}}{c(\theta)}, \tag{2}$$

where $c(\theta) = 1/a(Q^{-1}(\theta))$ for the new parameter $\theta \in \mathbb{R}^m$ called a *natural parameter*. The function $c(\theta)$ is then given by the simple formula

$$c(\theta) = \int_{\mathscr{X}} \mathrm{e}^{T(x)\, \theta^t}\, \mathrm{d}\lambda(x),$$

where λ is the new dominating measure.

By Hölder's inequality, $c(\theta)$ is convex (even logconvex, i. e. $\ln c(\theta)$ is convex). Therefore, the set

$$\left\{\theta \in \mathbb{R}^m\colon\ 0 < \int_{\mathscr{X}} \mathrm{e}^{T\, \theta^t}\, \mathrm{d}\lambda < \infty\right\}$$

is convex.

We assume that Θ is nonvoid open and consider the experiment with $\mathscr{P} = (P_\theta : \theta \in \Theta)$ where P_θ is defined by the density (2) w. r. t. a σ-finite measure λ. This experiment is regular in the sense of Brown (1986). Moreover, we assume that for different $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$ there is no real c with the property

$$\lambda(\{x \in \mathscr{X} : T(x)\,(\theta_1 - \theta_2)^t \neq c\}) = 0.$$

This assumption means that $\mathscr{P}$ is not overparametrized (the experiment is minimal in the sense of Brown (1986)). This implies in particular that all distributions in $\mathscr{P}$ are different and that $\ln c(\theta)$ has a positive definite Hessian matrix on Θ.

If we denote

$$d(\theta) = \ln c(\theta), \quad \hat{T}_n = \hat{T}_n(\boldsymbol{X}_n) = \frac{1}{n}\sum_{k=1}^{n} T(X_k),$$

then the function

$$f_n(\theta) = d(\theta) - \hat{T}_n\,\theta^t, \quad \text{i. e.} \quad f_n(\theta, \boldsymbol{X}_n) = d(\theta) - \hat{T}_n(\boldsymbol{X}_n)\,\theta^t \tag{3}$$

is convex and the expression behind $\lim\limits_n$ in (1) reduces to

$$\inf_{\theta \in S^c} f_n(\theta) - \inf_{\theta \in S} f_n(\theta),$$

where $S^c = \Theta - S$. We will be interested in the random variables $f_n(S) = f_n(S, \boldsymbol{X}_n)$ defined by

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

for various subsets $S \subset \Theta$. By means of these variables (1) can be rewritten into the form

$$I_{\theta_0}(S) = \lim_n (f_n(S^c) - f_n(S)) \quad \text{a. s. for } S \neq \emptyset,\ S \neq \Theta. \tag{4}$$

As is well known (cf. Brown (1986)), (3) is not only convex but also analytic in the variable $\theta \in \Theta$. We draw several useful consequences of this. First, (3) is continuous on Θ so that inf over S can be replaced by inf over a dense countable subset of S. This implies that $f_n(\boldsymbol{X}_n, S)$ is measurable in $\boldsymbol{X}_n$, i. e. $f_n(S)$ is a random variable for all sets S considered in (4). This formally justifies the definition (4).

The following assertion follows directly from (4). Note that we extend formulas (5), (6) in this assertion to all subsets $S \subset \Theta$ by the convention $I_{\theta_0}(\Theta) = \infty$, $I_{\theta_0}(\emptyset) = -\infty$.

Theorem 1. *If for some* $\emptyset \neq S \neq \Theta$ *there exist constants* $I_{\theta_0}(S)$ *and* $I_{\theta_0}(S^c)$ *satisfying* (4) *then*

$$I_{\theta_0}(S^c) = -I_{\theta_0}(S). \tag{5}$$

If for $\emptyset \neq S_1 \subset S_2 \neq \Theta$ *there exist constants* $I_{\theta_0}(S_1)$ *and* $I_{\theta_0}(S_2)$ *satisfying* (4) *then*

$$I_{\theta_0}(S_1) \leqslant I_{\theta_0}(S_2). \tag{6}$$

We will also need some facts concerning the *maximum likelihood estimator* (MLE)

$$\hat{\theta}_n = \hat{\theta}_n(\boldsymbol{X}_n) \stackrel{\triangle}{=} \arg\min\, f_n(\theta).$$

The maximum likelihood equation $\nabla f_n(\theta) = 0$ takes on the form

$$\tau(\hat{\theta}_n) = \hat{T}_n \quad \text{for} \quad \tau(\theta) = \nabla\, d(\theta) \tag{7}$$

where $\nabla = (\partial/\partial\theta_1), \ldots \partial/\partial\theta_m$ stands for the gradient. As follows from (2),

$$\tau(\theta_0) = \mathsf{E}\,\hat{T}_n = \mathsf{E}\,T. \tag{8}$$

The assumption of strict convexity of $d(\theta)$ implies that $\tau(\theta)$ is invertible and, by Theorem 3.6 and formula (2) on p. 145 in Brown (1986), there exist unique solutions $\hat{\theta}_n \in \Theta$ of equation (7).

Let us express the *Fisher information* $m \times m$ matrix by

$$\mathscr{I}_\theta = \nabla^t\, \tau(\theta) = \nabla^t\, \nabla d(\theta).$$

By what has been said above, $\mathscr{I}_\theta$ as the Hessian of $\ln c(\theta)$ must be positive definite on Θ, and its elements are obviously continuous on Θ. By applying the Taylor theorem we get from the maximum likelihood equation

$$\hat{\theta}_n = \theta_0 + \frac{\xi}{\sqrt{n}} + o_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{for } n \to \infty \tag{9}$$

where ξ is $N(0, \mathscr{I}_{\theta_0}^{-1})$.

Throughout this paper we denote for any set $S \subset \Theta$ by $\overline{S}$ and S^0 the relative closure and interior of S in the subspace Θ of $\mathbb{R}^m$, by S^c the above introduced relative complement and by

$$\partial S = \overline{S} - S^0$$

the relative boundary.

In the next lemma we consider the straight line $L = L(\theta_1, \theta_2)$ passing through two different points $\theta_1, \theta_2 \in \mathbb{R}^m$, i. e. we consider the linear subspace of $\mathbb{R}^m$ defined by

$$L = \{\theta_1(1-y) + \theta_2\, y\colon\ y \in \mathbb{R}\}.$$

We say that a subset $S \subset \Theta$ separates θ_1 and θ_2 if $\theta_1 \in S$ and $\theta_2 \in S^c$.

Lemma 1. *The set $L \cap \partial S$ is nonempty for every subset $S \subset \mathbb{R}^m$ which separates $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$.*

P r o o f. If θ_1 or θ_2 belongs to ∂S then the statement holds. Otherwise one of the points, say θ_1, belongs to the relative interior S^0 while θ_2 belongs to the relative complement $(\overline{S})^c$ of the relative closure $\overline{S}$, i. e.

$$L \cap S^0 \neq \emptyset \quad \text{and} \quad L \cap \overline{S} \neq L \cap \Theta. \tag{10}$$

Suppose now that $L \cap \partial S$ is empty, i. e. $L \cap (\overline{S} - S^0) = \emptyset$ or, equivalently,

$$L \cap S^0 = L \cap \overline{S}.$$

Since $L \cap S^0$ is relatively open and $A \stackrel{\triangle}{=} L \cap \overline{S}$ is relatively closed in the subspace $L \cap \Theta$ of $\mathbb{R}^m$, the last assumption implies that A is simultaneously relatively open and closed in $L \cap \Theta$. Therefore

$$A = \emptyset \quad \text{or} \quad A = L \cap \Theta,$$

which contradicts (10). □

By the strong law of large numbers,

$$\lim_n f_n(\theta) = f(\theta) \quad \text{for all } \theta \in \Theta, \tag{11}$$

where

$$f(\theta) = d(\theta) - (\mathsf{E}\, T)\, \theta^t = d(\theta) - \tau(\theta_0)\, \theta^t.$$

It follows from here for every nonvoid proper subset $S \subset \Theta$ that

$$\limsup_n f_n(S) \leqslant f(S) \quad \text{for} \quad f(S) \stackrel{\triangle}{=} \inf_S f(\theta).$$

We can prove a stronger result.

Lemma 2. *For every $S \subset \Theta$, $S \neq \emptyset$, $S \neq \Theta$, and the above defined $f(S)$,*

$$\lim_n f_n(S) = f(S).$$

P r o o f. (I) It suffices to prove

$$\liminf_n f_n(S) \geqslant f(S) - \varepsilon$$

for all S under consideration and all $\varepsilon > 0$. To this end we need the fact that the convergence in (11) is uniform on bounded sets, which follows from the relation

$$f_n(\theta) - f(\theta) = \theta(\hat{T}_n - \mathsf{E}\, T), \quad \theta \in \Theta.$$

In fact, for bounded S the assertion of Lemma 2 follows directly from here.

(II) For unbounded S we assume for simplicity the existence of points

$$\theta^* = \arg\min_S f(\theta) \quad \text{and} \quad \theta_n^* = \arg\min_S f_n(\theta).$$

Modification of the proof in the case that the infima of $f(\theta)$ and $f_n(\theta)$ are not attained on S will be obvious. Consider a sphere $S_r = S_r(\theta^*) \subset \mathbb{R}^m$ of radius $r > 0$ centered at θ^*. Put

$$A = \partial S_r = \{\theta \in \Theta\colon \|\theta\| = r\}.$$

Since S_r separates θ^* and any $\theta \in S_r^c$, Lemma 1 implies that for every $\theta \in S_r^c$ there exists $\theta_* \in A$ and $y \geqslant 1$ such that

$$\theta = \theta^*(1 - y) + \theta_*\, y.$$

Moreover, the convexity of f_n implies

$$f_n(\theta^*(1 - y) + \theta_*\, y) \geqslant f_n(\theta^*)\,(1 - y) + f_n(\theta_*)\, y.$$

Thus if $f_n(\theta^*) < \min_A f_n(\theta)$ then

$$f_n(\theta) \geqslant f_n(\theta^*)\,(1 - y) + f_n(\theta^*)\, y = f_n(\theta^*) \quad \text{for all } \theta \in S_r^c.$$

It follows from here that $\theta_n^* \in S_r^c$ implies $f_n(\theta^*) \geqslant \min_A f_n(\theta)$. Taking the limits on both sides and using the uniform convergence in (11) established in part (I), we get

$$f(\theta^*) \geqslant \min_A f(\theta).$$

But the last minimum strictly exceeds $f(\theta^*)$ due to the strict convexity of f at θ^*. Therefore, all but finitely many θ_n^* are a. s. in the sphere S_r. If however this happens then

$$\liminf_n f_n(\theta_n^*) \geqslant \lim_n \inf_{S_r \cap \Theta} f_n(\theta) = \inf_{S_r \cap \Theta} f(\theta),$$

where the second relation follows from the above mentioned uniform convergence in (11). The last infimum can be made greater than $f(\theta^*) - \varepsilon$ for any $\varepsilon > 0$ by taking the diameter sufficiently small. □

In the rest of the paper we consider the I-divergence of distributions P_{θ_0}, P_θ (information divergence, Kullback–Leibler number, see Kullback (1959) or Liese and Vajda (1987))

$$I(\theta_0; \theta) = \int p_{\theta_0}(x) \ln \frac{p_{\theta_0}(x)}{p_\theta(x)} \, \mathrm{d}\lambda(x).$$

It follows from (2) that for the exponential models under consideration

$$I(\theta_0; \theta) = f(\theta) - f(\theta_0) = d(\theta) - d(\theta_0) - \tau(\theta_0)\,(\theta - \theta_0)^t. \tag{12}$$

By combining (12) with Lemma 2 and (4) we obtain the following result.

Lemma 3. *For every parameter set S*

$$\begin{aligned} I_{\theta_0}(S) &= \inf_{\theta \in S^c} I(\theta_0; \theta) - \inf_{\theta \in S} I(\theta_0; \theta) \\ &= \begin{cases} \inf\limits_{\theta \in S^c} I(\theta_0; \theta) & \text{if } \theta_0 \in S \\ -\inf\limits_{\theta \in S} I(\theta_0; \theta) & \text{if } \theta_0 \in S^c, \end{cases} \end{aligned}$$

where $I(\theta_0; \theta)$ is given by (12) *and* $\inf \emptyset = \infty$.

The next assertion can obviously be applied in Lemma 3.

Lemma 4. *If $\emptyset \neq S \subset \Theta$ and $\theta_0 \in S^c$ then the boundary ∂S is nonempty and*

$$\inf_{\theta \in S} I(\theta_0; \theta) = \inf_{\theta \in \partial S} I(\theta_0; \theta).$$

P r o o f. Consider an arbitrary $\theta_* \in S$. It suffices to prove that there exists $\theta^* \in \partial S$ such that

$$I(\theta_0; \theta_*) \geqslant I(\theta_0; \theta^*).$$

Since S separates θ_0 and θ_*, Lemma 1 implies that the set $L = L(\theta_0, \theta_*)$ has a nonvoid intersection with the relative boundary ∂S. Thus there exist $\theta^* \in L \cap \partial S$ and $\eta \geqslant 1$ such that

$$\theta_* = \theta_0(1-\eta) + \theta^*\,\eta.$$

Further, since $I(\theta_0;\theta)$ is convex in the variable $\theta \in \Theta$,

$$\varphi(y) = I(\theta_0; \theta_0(1-y) + \theta^*\,y)$$

is convex in the domain $y \geqslant 0$ with $\varphi(0) = 0$. Therefore

$$I(\theta_0, \theta_*) = \varphi(\eta) \geqslant \varphi(0)\,(\eta - 1) + \varphi(1)\,\eta \geqslant \varphi(1) = I(\theta_0, \theta^*).$$

□

Now we can summarize the previous auxiliary results as follows.

Theorem 2. *For every subset $S \subset \Theta$ different from $\emptyset$ and Θ, the global information $I_{\theta_0}(S)$ is equal to*

$$\begin{aligned} \pm \inf_{\theta\in\partial S} I(\theta_0;\theta) &= \pm \big[\inf_{\partial S} f(\theta) - f(\theta_0)\big] \\ &= \pm \big[\inf_{\theta\in\partial S} (d(\theta) - \tau(\theta_0)\,\theta^t) - (d(\theta_0) - \tau(\theta_0)\,{\theta_0}^t)\big], \end{aligned}$$

where the sign $+$ takes place if $\theta_0 \in S$ and $-$ in the opposite case, and ∂S is nonempty.

P r o o f. Clear from Lemmas 3 and 4. □

Note that the boundary ∂S coincides with the boundary $\partial(S^c)$ of the complement. Thus the global information $I_{\theta_0}(S)$ is zero if and only if $I_{\theta_0}(S^c)$ is zero and this is equivalent to $\theta_0 \in \partial S$. If θ_0 is in the interior S^0 then the global information is positive, and if θ_0 is in the interior $(S^c)^0$ of the complement then the global information is negative.

C o n c l u s i o n s. The global information $I_{\theta_0}(S)$ characterizes the likelihood as to whether the unknown parameter θ_0 is in the set S. Within its range $[-\infty, +\infty]$, this information respects the intuitively appealing monotonicity and skew-symmetry rules of Theorem 1. It can be computed by means of Theorem 2.

3. Explicit formulas

In this section we evaluate the I-divergence (12) for the common exponential models. For univariate parameters these formulas enable us to evaluate explicitly the global information by employing the following result.

Theorem 3. *If $\Theta \subset \mathbb{R}$ and*

$$\theta_1 = \arg\min_{\partial S} |\theta_0 - \theta| \tag{13}$$

minimizes the distance between θ_0 and the boundary ∂S then

$$I_{\theta_0}(S) = \begin{cases} I(\theta_0;\theta_1) & \text{if } \theta_0 \in S \\ -I(\theta_0;\theta_1) & \text{if } \theta_0 \in S^c. \end{cases} \tag{14}$$

P r o o f. If θ is real-valued then the convexity of $\varphi(\theta) = I(\theta_0;\theta)$ implies that

$$\psi(\theta) = \frac{\varphi(\theta) - \varphi(\theta_0)}{\theta - \theta_0} = \frac{\varphi(\theta)}{\theta - \theta_0}$$

is increasing in the domain $\Theta - \{\theta_0\}$. Thus $\varphi(\theta)/(\theta - \theta_0)$ is increasing in the domain $\theta > \theta_0$ and $\varphi(\theta)/(\theta_0 - \theta)$ decreasing in the domain $\theta < \theta_0$. In other words, the functions

$$\varphi(\theta) = I(\theta_0;\theta) \quad \text{and} \quad |\theta - \theta_0|$$

are isotone on the whole domain Θ. Consequently, the minimization of Euclidean distance means the minimization of the I-divergence and vice versa. This together with Theorem 2 implies (14). □

Note that some formulas are listed below also for bivariate $\theta = (\vartheta_1, \vartheta_2)$. In the multivariate case, $I(\theta_0;\theta)$ usually, but not always, increases with the Euclidean distance $\|\theta - \theta_0\|$. Therefore in this case Theorem 3 with $|\theta_0 - \theta|$ in (13) replaced by $\|\theta_0 - \theta\|$ need not be true, and the formulas given below have to be inserted into Theorem 2. If the boundary ∂S is defined by means of a differentiable function, then these formulas can be combined with the Lagrange multipliers method.

Binomial model with a natural parameter $\theta \in \mathbb{R}$. Here λ is supported by $\mathscr{X} = \{0, 1, \ldots, n\}$. For every $x \in \mathscr{X}$

$$\lambda(x) = \binom{n}{x}$$

and

$$p_\theta(x) = \frac{\mathrm{e}^{\theta x}}{(1+\mathrm{e}^\theta)^n} = p^x(1-p)^{n-x} \quad \text{for } p = \frac{\mathrm{e}^\theta}{1+\mathrm{e}^\theta}.$$

From (12) we obtain

$$I(\theta_0;\theta) = n\left[\frac{(\theta_0-\theta)\,\mathrm{e}^{\theta_0}}{1+\mathrm{e}^{\theta_0}} + \ln\frac{1+\mathrm{e}^\theta}{1+\mathrm{e}^{\theta_0}}\right]$$

and

$$I(p_0;p) = n\left[p_0\,\ln\frac{p_0}{p} + (1-p_0)\ln\frac{1-p_0}{1-p}\right].$$

Poisson model with a natural parameter $\theta \in \mathbb{R}$. The support of λ is $\mathscr{X} = \{0,1\ldots\}$ and for every $x \in \mathscr{X}$ we have

$$\lambda(x) = \frac{1}{x!}$$

and

$$p_\theta(x) = \frac{\mathrm{e}^{\theta x}}{\mathrm{e}^{\mathrm{e}^\theta}} = \frac{\tau^x}{\mathrm{e}^\tau} \quad \text{for } \tau = \mathrm{e}^\theta.$$

By (12) we infer

$$I(\theta_0;\theta) = (\theta_0-\theta)\,\mathrm{e}^{\theta_0} + \mathrm{e}^\theta - \mathrm{e}^{\theta_0}$$

and

$$I(\tau_0;\tau) = \tau_0\left(\frac{\tau}{\tau_0} - 1 - \ln\frac{\tau}{\tau_0}\right).$$

Geometric model with a natural parameter $\theta > 0$. Here λ is counting with the same support $\mathscr{X}$ as above. For every $x \in \mathscr{X}$ we have

$$p_\theta(x) = (1-\mathrm{e}^{-\theta})\,\mathrm{e}^{-\theta x} = (1-p)\,p^x \quad \text{for } p = \mathrm{e}^{-\theta}.$$

It follows from (12) that

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

and

$$I(p_0;p) = \left(p_0\ln\frac{p_0}{p} + (1-p_0)\ln\frac{1-p_0}{1-p}\right)\Big/\,(1-p_0).$$

Negative binomial model with a natural parameter $\theta > 0$. The space $\mathscr{X}$ is as before and, for every $x \in \mathscr{X}$,

$$\lambda(x) = \binom{r+x-1}{x}, \qquad p_\theta(x) = (1-\mathrm{e}^{-\theta})^r \, \mathrm{e}^{-\theta x},$$

where $r > 0$ is given in advance and fixed. The common parameter p is the same as in the previous example. We obtain from (12) that

$$I(\theta_0;\theta) = r\left(\frac{\theta-\theta_0}{\mathrm{e}^{\theta_0}-1} + \ln\frac{1-\mathrm{e}^{-\theta_0}}{1-\mathrm{e}^{-\theta}}\right)$$

and

$$I(p_0;p) = r\left(p_0 \ln\frac{p_0}{p} + (1-p_0)\ln\frac{1-p_0}{1-p}\right)\Big/(1-p_0).$$

Now we turn to continuous exponential models. The simplest of these models is given for $\theta > 0$ by the density

$$p_\theta(x) = \theta \, \mathrm{e}^{-\theta x}$$

with respect to the restriction λ of the Lebesgue measure on $\mathscr{X} = (0,\infty)$. This is the common *exponential distribution*. Here (12) implies

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

Normal model with natural parameters $(\vartheta_1,\vartheta_2) \in \mathbb{R}\times(0,\infty)$. The dominating measure λ is the Lebesgue one on $\mathscr{X} = \mathbb{R}$ and, for every $x \in \mathbb{R}$,

$$p_{\theta_1,\theta_2}(x) = \frac{\mathrm{e}^{\vartheta_1 x - \vartheta_2 x^2}}{\sqrt{\frac{\pi}{\vartheta_2}}\mathrm{e}^{\frac{1}{4}\vartheta_1^2/\vartheta_2}} = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } \mu = \frac{\vartheta_1}{2\vartheta_2},\ \sigma = \frac{1}{\sqrt{2\vartheta_2}}.$$

By (12),

$$\begin{aligned} I((\vartheta_{01};\vartheta_{02});(\vartheta_1,\vartheta_2)) = {} & \frac{\vartheta_{01}(\vartheta_{01}-\vartheta_1)}{2\vartheta_{02}} + \left[\left(\frac{\vartheta_{01}}{2\vartheta_{02}}\right) - \frac{1}{2\vartheta_{02}}\right](\vartheta_{02}-\vartheta_2) \\ & + \frac{1}{2}\ln\frac{\vartheta_{02}}{\vartheta_2} + \frac{\vartheta_1^2}{4\vartheta_2} - \frac{\vartheta_{01}^2}{4\vartheta_{02}} \end{aligned}$$

so that

$$I((\mu_0,\sigma_0);(\mu,\sigma)) = \frac{1}{2}\left[\frac{(\mu-\mu_0)^2}{\sigma^2} + \frac{\sigma_0^2}{\sigma^2} - 1 - \ln\frac{\sigma_0^2}{\sigma^2}\right].$$

Here $\inf\limits_{\sigma>0} I((\mu_0,\sigma_0);(\mu,\sigma))$ is attained at $\sigma^2=(\mu-\mu_0)^2+\sigma_0^2$ and

$$\inf_{\sigma>0} I((\mu_0,\sigma_0);(\mu,\sigma)) = \frac{1}{2}\ln\left(1+\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right). \tag{15}$$

R e m a r k 1. It follows from the invariance of the I-divergence with respect to sufficient transformations that the I-divergence in the model $(P_\theta T^{-1}\colon\ \theta\in\Theta)$ for the one-to-one mapping $T(x)$, $x\in\mathscr{X}$, coincides with that in the model $(P_\theta\colon\ \theta\in\Theta)$. This is illustrated by the following model.

Lognormal model with natural parameters $(\vartheta_1,\vartheta_2)\in\mathbb{R}\times(0,\infty)$. The dominating measure λ is concentrated on $S=(0,\infty)$ where it has the density $1/x$ with respect to the Lebesgue measure. For every $x\in S$,

$$p_{\vartheta_1,\vartheta_2}(x)=\frac{\mathrm{e}^{\vartheta_1\log x-\vartheta_2 x^2}}{\sqrt{\frac{\pi}{\vartheta_2}}\mathrm{e}^{\frac{1}{4}\vartheta_1^2/\vartheta_2}}=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(\log x-\mu)^2}{2\sigma^2}}\quad\text{for }\mu=\frac{\vartheta_1}{2\vartheta_2},\ \sigma=\frac{1}{\sqrt{2\vartheta_2}}.$$

Here $I((\vartheta_{01};\vartheta_{02});(\vartheta_1,\vartheta_2))$ and $I((\mu_0,\sigma_0);(\mu,\sigma))$ are the same as in the normal model.

Gamma model with natural parameters $(\vartheta_1,\vartheta_2)\in(-1,\infty)\times(0,\infty)$. Here λ is the Lebesgue measure restricted to the same $\mathscr{X}$ as above and, for every $x\in S$,

$$p_{\vartheta_1,\vartheta_2}(x)=\frac{\mathrm{e}^{\vartheta_1\log x-\vartheta_2 x}}{\Gamma(\vartheta_1+1)/\vartheta_2^{\vartheta_1+1}}=\frac{\tau^r}{\Gamma(n)}x^{n-1}\mathrm{e}^{-\tau x}\quad\text{for }n=\vartheta_1+1,\ \tau=\vartheta_2.$$

It follows from (12) that

$$I((\theta_1,\vartheta_{02});(\vartheta_1,\vartheta_2))=(\vartheta_1+1)\left(\frac{\vartheta_2}{\vartheta_{02}}-1-\ln\frac{\vartheta_2}{\vartheta_{02}}\right)$$

and

$$I((n,\tau_0);(n,\tau))=n\left(\frac{\tau}{\tau_0}-1-\ln\frac{\tau}{\tau_0}\right).$$

Beta model with natural parameters $\vartheta_1,\ \vartheta_2>-1$. Here λ is the restriction of the Lebesgue measure on the interval $\mathscr{X}=(0,1)$ and, for every $x\in\mathscr{X}$,

$$p_{\vartheta_1,\vartheta_2}(x)=\frac{\mathrm{e}^{\vartheta_1\log x+\vartheta_2\log(1-x)}}{B(\vartheta_1+1,\vartheta_2+1)}=\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\quad\text{for }a=\vartheta_1+1,\ b=\vartheta_2+1.$$

We obtain from (12)

$$I((a_0,b_0);(a,b)) = \ln\frac{B(a,b)}{B(a_0,b_0)} + B_1'(a_0,b_0)\,(a_0-a) + B_2'(a_0,b_0)\,(b_0-b)$$

where $B_1'(x,y) = \partial B(x,y)/\partial x$ and $B_2'(x,y) = \partial B(x,y)/\partial y$. Consider the so-called digamma function defined by

$$\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\ln\Gamma(x), \quad x>0,$$

and satisfying the relation

$$\psi(x)-\psi(y) = (x-y)\sum_{j=0}^{\infty}\frac{1}{(j+x)\,(j+y)}, \quad x,\,y>0$$

(see Spanier and Oldham (1987)). By virtue of the formula $B(x,y) = \Gamma(x)\,\Gamma(y)/r(x+y)$ and the function

$$\varphi(x,y) = y\sum_{j=0}^{\infty}\frac{1}{(j+x)\,(j+x+y)},$$

the I-divergence of two Beta distributions can be expressed as

$$\ln\frac{B(a,b)}{B(a_0,b_0)} - B(a_0,b_0)\left[\varphi(a_0,b_0)\,(a_0-a) + \varphi(b_0,a_0)\,(b_0-b)\right].$$

The I-divergences in the next three models can be obtained from the formula for the Gamma model with $n=\frac{1}{2},\ 1$ and $\frac{3}{2}$, by using Remark 1 for $T(x)=\sqrt{x}$.

Modular model with natural parameter $\theta>0$. The dominating measure is the Lebesgue one on $\mathscr{X}=(0,\infty)$ and

$$p_\theta(x) = \frac{\mathrm{e}^{-\theta\,x^2/2}}{\sqrt{\frac{\pi}{2\theta}}}$$

and

$$I(\theta_0;\theta) = \frac{1}{2}\left(\frac{\theta}{\theta_0} - 1 - \ln\frac{\theta}{\theta_0}\right).$$

Rayleigh model with natural parameter $\theta>0$. Here the dominating λ has density x with respect to the Lebesgue measure on the same $\mathscr{X}$ as above,

$$p_\theta(x) = \theta\,\mathrm{e}^{-\theta\,x^2/2}$$

and

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

Maxwell model with natural parameter $\theta > 0$. Here the density of the dominating measure is x^2 on $\mathscr{X} = (0,\infty)$,

$$p_\theta(x) = \frac{\mathrm{e}^{-\theta\,x^2/2}}{\sqrt{\frac{\pi}{2\theta^3}}}$$

and

$$I(\theta_0;\theta) = \frac{3}{2}\left(\frac{\theta}{\theta_0} - 1 - \ln\frac{\theta}{\theta_0}\right).$$

4. Testing of hypotheses

In this section we consider a hypothesis and an alternative

$$\mathscr{H}\colon\ \theta_0 \in S \quad \text{and} \quad \mathscr{A}\colon\ \theta_0 \in S^c$$

being tested on the basis of data $\boldsymbol{X}_n$ in the general exponential model (2). The set S of parameters is supposed to be arbitrary, different from $\emptyset$ and Θ. By Lemma 4, the boundary ∂S is then nonempty.

We investigate the *global information test* (GIT) based on the global information statistic

$$\Gamma_n = I_{\hat\theta_n}(S) = \begin{cases} \inf\limits_{\theta\in\partial S} I(\hat\theta_n,\theta) & \text{if } \hat\theta_n \in S \\ -\inf\limits_{\theta\in\partial S} I(\hat\theta_n,\theta) & \text{otherwise,} \end{cases}$$

where $\hat\theta_n = \hat\theta(\boldsymbol{X}_n)$ is the MLE defined by (7), which is rejecting $\mathscr{H}$ if and only if

$$\Gamma_n \leqslant \varepsilon_n. \tag{16}$$

Further, we consider the *generalized likelihood ratio test* (GLRT) based on the statistic

$$\Lambda_n = \frac{\sup\limits_{S} \mathrm{e}^{n(\hat T_n\,\theta^t - d(\theta))}}{\sup\limits_{\Theta} \mathrm{e}^{n(\hat T_n\,\theta^t - d(\theta))}} = \left(\frac{\sup\limits_{S} \mathrm{e}^{\hat T\theta^t - d(\theta)}}{\sup\limits_{\Theta} \mathrm{e}^{\hat T_n\theta^t - d(\theta)}}\right)^n,$$

where $\hat T_n = \hat T_n(\boldsymbol{X}_n)$ is the same as in (3), which is rejecting $\mathscr{H}$ if and only if

$$\Lambda_n \leqslant \lambda_n. \tag{17}$$

Unless otherwise stated, ε_n and λ_n are arbitrary real sequences.

Lemma 5. *For every S under consideration,*

$$\Lambda_n = \begin{cases} 1 & \text{if } \hat\theta_n \in S \\ \mathrm{e}^{n\,\Gamma_n} & \text{if } \hat\theta_n \in S^c. \end{cases}$$

P r o o f. By definition, (7) and (11) we have

$$\begin{aligned} (\Lambda_n)^{1/n} &= \frac{\sup\limits_{S} \mathrm{e}^{\tau(\hat\theta_n)\,\theta^t - d(\theta)}}{\sup\limits_{\Theta} \mathrm{e}^{\tau(\hat\theta_n)\,\theta^t - d(\theta)}} \\ &= \frac{\sup\limits_{S} \mathrm{e}^{\tau(\hat\theta_n)\,\theta^t - d(\theta)}}{\mathrm{e}^{\tau(\hat\theta_n)\,\hat\theta_n^t - d(\hat\theta_n)}} \\ &= \sup_{S} \mathrm{e}^{f(\hat\theta_n) - f(\theta)} = \sup_{S} \mathrm{e}^{-I(\hat\theta_n;\theta)} \\ &= \mathrm{e}^{-\inf\limits_{S} I(\hat\theta_n;\theta)}. \end{aligned}$$

The desired equality follows from the fact that $I(\hat\theta_n;\hat\theta_n) = 0$ minimizes $I(\hat\theta_n;\theta)$ on S and that, by Lemma 2,

$$-\inf_{S} I(\hat\theta_n;\theta) = I_{\hat\theta_n}(S) \quad \text{if } \hat\theta_n \in S^c.$$

□

R e m a r k 2. This lemma implies that the global information statistic Γ_n contains in some sense more information about the unknown parameter θ_0 than the GLR Λ_n. Namely, $\Lambda_n = \Psi_n(\Gamma_n)$ where

$$\Psi_n(y) = \mathrm{e}^{n\,y}\mathbf{1}_{(-\infty,0)}(y) + \mathbf{1}_{[0,\infty)}(y), \quad y \in \mathbb{R},$$

while Γ_n cannot be obtained from Λ_n when $\Lambda_n = 1$, i. e. the nonnegative values of Γ_n cannot be reconstructed from the statistic Λ_n.

The next result implies that GIT can always be at least as good as the well known GLRT. We see that ε_n given by (18) is from the subset $[-\infty, 0)\cup\{\infty\}$ of the extended real line. In fact, the critical values $\varepsilon_n \in [0,\infty]$ are of a limited practical importance since they lead to unpleasant behaviour of probabilities $\mathsf{P}(\Gamma_n \leqslant \varepsilon_n)$ for θ_0 close to the boundary of S. This is reflected in Theorem 5 below, where only negative ε_n are allowed. Therefore the advantages of GIT over GLRT mentioned in Remark 2 and

Theorem 4 are interesting rather from the theoretical than from the practical point of view.

Theorem 4. (i) *If*

$$\varepsilon_n = -\infty\, \mathbf{1}_{(-\infty,0]}(\lambda_n) + \mathbf{1}_{(0,1)}(\lambda_n)\, \frac{1}{n}\, \ln \lambda_n + \infty\, \mathbf{1}_{[1,\infty)}(\lambda_n) \tag{18}$$

then GIT and GLRT coincide in the sense that the event (17) *is equivalent to* (16).

(ii) *If* $\varepsilon_n \in [0, \infty)$ *then GLRT cannot coincide with GIT, i. e.* (17) *is for no* $-\infty \leqslant \lambda_n \leqslant \infty$ *equivalent with* (16).

P r o o f. If $\lambda_n \in (0, 1)$ then (18) implies

$$\varepsilon_n = \frac{1}{n} \ln \lambda_n \quad \text{or} \quad \lambda_n = \mathrm{e}^{n\,\varepsilon_n}$$

and it follows from Lemma 5 that (16) holds, if and only if (17) takes place, i. e. in this case (i) holds, too. If $\varepsilon_n \in [0, \infty)$ then (17) is not equivalent (16) in the models with $\mathsf{P}(\Gamma_n > 0) > 0$, and such models obviously exist. Thus (ii) holds as well. □

The following fact is well known for many particular exponential models and common hypotheses S, like intervals or rectangles. GLRT's of one-sided or two-sided hypotheses about the mean of normal model, with known or unknown variances, are perhaps the best known examples.

Corollary. *The generalized likelihood ratio statistic is, for every* S *under consideration and for the MLE* $\hat{\theta}_n$, *given by the formula*

$$\begin{aligned} \Lambda_n &= \exp\{-n \inf_{\theta \in \partial S} I(\hat{\theta}_n; \theta)\} \\ &= \exp\{-n\big[\inf_{\partial S} \left(d(\theta) - \tau(\hat{\theta}_n)\, \theta^t\right) - \left(d(\hat{\theta}_n) - \tau(\hat{\theta}_n)\, \hat{\theta}_n^t\right)\big]\} \end{aligned}$$

unless $\hat{\theta}_n \in S$, *in which case* $\Lambda_n = 1$.

P r o o f. Clear from Theorem 2 and Lemma 5. □

The next result is not principially new, either. It follows by the equivalence in part (i) of Theorem 4 from the familiar consistency of GLRT (cf. e. g. Lehmann (1986)). What might be of interest is the full generality of the model and hypothesis and a relatively simple proof. Recall also that the consistency of tests is considered in this paper in a slightly stronger form than in Lehmann (1986). By the consistency of GIT we mean that the asymptotic power of the test (16) is 1, i. e.

$$\lim_n \pi_n = 1 \quad \text{if } \theta_0 \in S^c \tag{19}$$

for

$$\pi_n = \mathsf{P}(\Gamma_n \leqslant \varepsilon_n),$$

and the asymptotic size is zero, i. e.

$$\lim_n \pi_n = 0 \quad \text{if } \theta_0 \in S. \tag{20}$$

In Lehmann (1986), the limit condition of (20) is replaced by $\limsup_n \pi_n \leqslant \alpha$ for a given $0 < \alpha < 1$.

Theorem 5. *If $\varepsilon_n \uparrow 0$ and $n\,\varepsilon_n \to -\infty$ as $n \to \infty$, then the GIT is consistent for all hypotheses $\mathscr{H}$ with S relatively closed in the subspace $\Theta \subset \mathbb{R}^m$. More precisely, $\varepsilon_n \uparrow 0$ implies* (19) *and $n\,\varepsilon_n \to -\infty$ implies* (20).

P r o o f. In the trivial cases $S = \emptyset$ or $S = \Theta$ the statement is obviously true. Let us therefore suppose $S \neq \emptyset$ and $S \neq \Theta$. (I) The strong law of large numbers together with (7) and (8) implies the strong consistency of the MLE $\hat{\theta}_n$. Further, by Theorem 2 we have in the notation of Lemma 1

$$|I_{\hat{\theta}_n}(S) - I_{\theta_0}(S)| \leqslant |f_n(\partial S) - f(\partial S)| + |f_n(\hat{\theta}_n) - f(\theta_0)|.$$

Lemma 1 and the strong consistency of $\hat{\theta}_n$ now imply

$$\lim_n I_{\hat{\theta}_n}(S) = I_{\theta_0}(S) \quad \text{a. s.} \tag{21}$$

(II) Since S^c is relatively open, $\theta_0 \in S^c$ implies $I_{\theta_0}(S) < 0$ so that (19) follows for all $\varepsilon_n = o(1)$ from (21).

(III) By the Bayes formula we obtain

$$\begin{aligned}\pi_n = {} & \mathsf{P}\left(I_{\hat{\theta}_n}(S) \leqslant \varepsilon_n \,|\, \hat{\theta}_n \in S\right) \mathsf{P}(\hat{\theta}_n \in S) \\ & + \mathsf{P}\left(I_{\hat{\theta}_n}(S) \leqslant \varepsilon_n \,|\, \hat{\theta}_n \in S^c\right) \mathsf{P}(\hat{\theta}_n \in S^c),\end{aligned}$$

where $I_{\hat{\theta}_n}(S) \geqslant 0$ for $\hat{\theta}_n \in S$ so that the first probability is zero and

$$\pi_n = \mathsf{P}\left(I_{\hat{\theta}_n}(S) \leqslant \varepsilon_n \,|\, \hat{\theta}_n \in S^c\right) \mathsf{P}(\hat{\theta}_n \in S^c).$$

(IV) Let us suppose $\theta_0 \in S$. If θ_0 is in the interior S^0 then the consistency of $\hat{\theta}_n$ implies $\lim_n \mathsf{P}(\hat{\theta}_n \in S^c) = 0$ so that (20) holds. Let us therefore suppose $\theta_0 \in \partial S$. If $\hat{\theta}_n \in S$ then $I_{\hat{\theta}_n}(S) \geqslant 0$ and

$$\mathsf{P}\left(I_{\hat{\theta}_n}(S) \leqslant \varepsilon_n\right) = 0.$$

If $\hat{\theta}_n \in S^c$ then $I_{\hat{\theta}_n}(S) = -\inf_{\theta \in \partial S} I(\hat{\theta}_n; \theta) \geqslant -I(\hat{\theta}_n; \theta_0)$ and

$$\begin{aligned} \mathsf{P}\left(I_{\hat{\theta}_n}(S) \leqslant \varepsilon_n\right) &\leqslant \mathsf{P}\left(-I(\hat{\theta}_n; \theta_0) \leqslant \varepsilon_n\right) \\ &= \mathsf{P}\left(I(\hat{\theta}_n; \theta_0) \geqslant -\varepsilon_n\right). \end{aligned}$$

By applying the Taylor theorem to (11) we get for all θ_1, $\theta_2 \in \Theta$

$$I(\theta_1; \theta_2) = \frac{1}{2}(\theta_1 - \theta_2)\, \mathscr{I}_\theta (\theta_1 - \theta_2)^t,$$

where $\mathscr{I}_\theta$ is the Fisher information introduced in Section 2 and θ is a point between θ_1 and θ_2. By inserting (9) in the last formula we get for $n \to \infty$

$$\begin{aligned} I(\hat{\theta}_n; \theta_0) &= \frac{1}{2n}\, \xi\, \mathscr{I}_{\theta_0}\, \xi^t + o_p\left(\frac{1}{n}\right) \\ &= \frac{1}{2n}\left(\mathscr{I}_{\theta_0}^{\frac{1}{2}}\, \xi^t\right)^t \left(\mathscr{I}_{\theta_0}^{\frac{1}{2}}\, \xi^t\right) + o_p\left(\frac{1}{n}\right) \\ &= \frac{\eta_m}{2n} + o_p\left(\frac{1}{n}\right), \end{aligned}$$

where η_m is chi-square distributed with m degrees of freedom. We see that if $n\,\varepsilon_n \to -\infty$ then

$$\lim_n \mathsf{P}(I(\hat{\theta}_n; \theta_0) < -\varepsilon_n) = 1.$$

□

It follows from the last proof that if

$$\varepsilon_n = -\frac{\chi^2_m(1-\alpha)}{2n},$$

where $\chi^2_m(1-\alpha)$ is the $(1-\alpha)$-quantile of χ^2_m, then the GIT is *asymptotically α-level* in the sense that $\theta_0 \in S$ implies $\limsup_n \pi_n \leqslant \alpha$. If the term $o_p(1/n)$ is of stochastic order $1/n$ uniformly for all $\theta_0 \in S$ (in fact, for all $\theta_0 \in \partial S$ only) then the GIT is *asymptotically uniformly α-level* in the sense

$$\limsup_n \sup_{\theta_0 \in S} \mathsf{P}_{\theta_0}\left(I_{\hat{\theta}_n} \leqslant \varepsilon_n\right) \leqslant \alpha,$$

where P_{θ_0} denotes the dependence of P on the true parameter $\theta_0 \in \Theta$. By Theorem 6, this test is then also *unbiased* in the sense that $\theta_0 \in S^c$ implies

$$\liminf_n \left[\pi_n - \sup_{\theta_0 \in S} \pi_n\right] \geqslant 0.$$

E x a m p l e 1. Let us consider in the normal model with unknowns μ_0 and σ_0 the hypothesis $\mathscr{H}\colon\ \mu_0 \leqslant c$. Here the MLE $\hat{\theta}_n = (\hat{\mu}_n, \hat{\sigma}_n)$ is given by the formulas

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2$$

and ∂S is the halfline $\{\mu, \sigma\colon\ \mu = c \text{ and } \sigma > 0\}$. By (15),

$$\inf_{(\mu,\sigma)\in\partial S} I((\hat{\mu}_n, \hat{\sigma}_n); (\mu, \sigma)) = \frac{1}{2}\ln\left(1 + \frac{(\hat{\mu}_n - c)^2}{\hat{\sigma}_n^2}\right).$$

Thus $\mathscr{H}$ is rejected by the GIT with a negative ε_n if and only if

$$\hat{\mu}_n > c \quad \text{and} \quad \frac{1}{2}\ln\left(1 + \frac{(\hat{\mu}_n - c)^2}{\hat{\sigma}_n^2}\right) \geqslant -\varepsilon_n \qquad \text{(cf. (16))},$$

i. e.

$$\hat{\mu}_n > c \quad \text{and} \quad \left|\frac{\sqrt{n}\,(\hat{\mu}_n - c)}{s_n}\right| \geqslant K_n$$

where

$$s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2} \quad \text{and} \quad K_n = \sqrt{(n-1)\,(\mathrm{e}^{-2\varepsilon_n} - 1)}.$$

Since $\sqrt{n}\,(\hat{\mu}_n - c)/s_n \leqslant T_n$ for every $\mu_0 \leqslant c$ where $T_n \stackrel{\triangle}{=} \sqrt{n}\,(\hat{\mu}_n - \mu_0)/s_n$ is t-distributed with $n-1$ degreees of freedom, the choice of the $(1-\alpha)$-quantile

$$K_n = t_{n-1}(1-\alpha)$$

guarantees for every n the test size α, and the asymptotic power 1. The GIT rejection rule becomes

$$\sqrt{n}\,(\hat{\mu}_n - c) \geqslant s_n\, t_{n-1}(1-\alpha) \quad \text{for all } 0 < \alpha < 1/2,$$

and the test is (nonasymptotically) uniformly α-level and unbiased.

5. Selection of models

As is well known, all dominated statistical models with densities regular in the Dynkin sense can be approximated by an exponential model of sufficiently high dimension m. In this section we apply the results of Section 4 in order to obtain statistical rules enabling us to reduce the dimension m of exponential models, or the scope of the parameter space Θ, or simultaneously both. In order to avoid confusion with terminology developed in other areas of statistics, we speak in this section about decision rules instead of selection rules.

For $j = 1, \ldots, M$ let us consider a sequence of hypotheses

$$\mathscr{H}_j\colon\ \theta_0 \in S_j, \quad \text{where} \quad S_1 \subset S_2 \subset \ldots \subset S_M$$

are subsets of Θ relatively closed in Θ. Put $S_0 = \emptyset$ and $S_{M+1} = \Theta$. We are interested in measurable decision rules $\Phi\colon\ \mathscr{X}^n \to \{1, \ldots, M+1\}$ where $\Phi(\boldsymbol{X}_n) = j$ decides on the submodel $(p_\theta\colon\ \theta \in S_j)$. For every $\theta_0 \in \Theta$ the decision $\Phi_n = \Phi(\boldsymbol{X}_n)$ is errorless if and only if

$$\Phi_n = \min\{j\colon\ \theta_0 \in S_j\} \equiv \sum_{j=0}^{M} 1_{S_j^c}(\theta_0).$$

Thus the probability of error is given by the formula

$$P(\Phi_n, \theta_0) = \mathsf{P}\left(\Phi_n \neq \sum_{j=0}^{M} 1_{S_j^c}(\theta_0)\right) \tag{22}$$

for all rules Φ and true parameters $\theta_0 \in \Theta$.

A decision rule Φ is said to be *consistent* if

$$\lim_n P(\Phi_n, \theta_0) = 0 \quad \text{for all } \theta_0 \in \Theta.$$

Consistent decision rules for selection of reduced models have been studied under special assumptions by several authors, cf. Bauer et al (1988), Nishii (1988) and others cited in Section 1. In this section we show that the selection criteria based on the GLRT's applied subsequently to the hypotheses $\mathscr{H}_1, \ldots, \mathscr{H}_M$, with critical values λ_n determined by the penalty terms, can in the case of exponential models be interpreted, and also practically realized, as information criteria using the global information statistics.

Let

$$\Gamma_{n,j} = I_{\hat{\theta}_n}(S_j), \quad j = 0, \ldots, M$$

be the global information statistics introduced in the previous section, satisfying the obvious inequalities

$$-\infty \equiv \Gamma_{n,0} < \Gamma_{n,1} \leqslant \ldots \leqslant \Gamma_{n,M+1} \equiv \infty.$$

We will be interested in the *global information rules* (GIR's), defined by means of real valued sequences ε_n as follows:

$$\Phi_n = \max\{j\colon\ \Gamma_{n,j} \leqslant \varepsilon_n\} + 1 \equiv \min\{j\colon\ \Gamma_{n,j} > \varepsilon_n\}. \tag{23}$$

This means that the GIR decides on the first of the hypotheses $\mathscr{H}_0, \mathscr{H}_1, \ldots, \mathscr{H}_{M+1}$ which is not rejected by the GIT rule (16). Since $\mathscr{H}_0$ is always and $\mathscr{H}_{M+1}$ never rejected, the GIR is well defined by (23).

Theorem 6. *If ε_n satisfies the assumptions of Theorem* 5 *then the GIR* (23) *is consistent.*

P r o o f. Let $\theta_0 \in \Theta$ be arbitrary, define

$$k = \sum_{j=0}^{M} 1_{S_j^c}(\theta_0)$$

and consider the events

$$A_n = \{\Gamma_{n,k-1} > \varepsilon_n\},$$
$$B_n = \begin{cases} \{\Gamma_{n,k} \leqslant \varepsilon_n\} \text{if } k < M \\ \emptyset \text{if } k = M+1. \end{cases}$$

The order in the statistics $\Gamma_{n,j}$ implies that the events A_n, B_n are disjoint and, for θ_0 and Φ_n under consideration,

$$P(\Phi_n, \theta_0) = \mathsf{P}(A_n) + \mathsf{P}(B_n). \tag{24}$$

If π_n is defined as in Section 4 then there exists $1 \leqslant j < k$ such that $1 - \mathsf{P}(A_n)$ is the power π_n of the GIT (16) for $\mathscr{H} = \mathscr{H}_j$, i. e.

$$\mathsf{P}(A_n) = 1 - \pi_n \quad \text{for } S = S_j,\ \theta_0 \in S_j^c,$$

and $\mathsf{P}(B_n)$ is the size of the GIT (16) for $\mathscr{H} = \mathscr{H}_k$, i. e.

$$\mathsf{P}(B_n) = \pi_n \quad \text{for } S = S_k,\ \theta_0 \in S_k.$$

By Theorem 5, it follows from here that the sum (24) tends to zero as $n \to \infty$. □

E x a m p l e 2. In the model of Example 1 let $M = 1$ and $S_1 = \{\mu, \sigma\colon\ \mu \leqslant c\}$. Let us first consider $\theta_0 \in S_1$. Then $\mathsf{P}(A_n) = 0$ and it follows from Example 1 that

$$\mathsf{P}(B_n) \leqslant \mathsf{P}\left(\left|\frac{\sqrt{n}(\hat{\mu}_n - \mu_0)}{s_n}\right| \geqslant \sqrt{(\mathrm{e}^{-2\varepsilon_n} - 1)\,(n-1)}\right).$$

Since $\mathrm{e}^{-2\varepsilon_n} - 1 = 2|\varepsilon_n| + o(\varepsilon_n)$, $\lim\limits_n n\,|\varepsilon_n| \to \infty$ assumed in Theorem 5 implies that also the limit of (24) is zero. If $\theta_0 \in S_1^c$, i. e. $\mu_0 > c$, then $\mathsf{P}(B_n) = 0$ and $\mathsf{P}(A_n) \leqslant \mathsf{P}(\hat{\theta}_n \in S_1) = \mathsf{P}(\hat{\mu}_n \leqslant c)$, where the last probability tends to zero by the law of large numbers. Thus again the limit of (24) is zero. If the problem is reformulated for $S_1 = \{\mu, \sigma\colon\ \mu = 0,\ \sigma > 0\}$ or $S_1 = \{\mu, \sigma\colon\ \mu = \sigma,\ \sigma > 0\}$ then we have a problem of dimension reduction. Here again the decision rule (23) is consistent.

Let us now consider $M = m - 1$ and let S_j be the subspace $\{\theta = (\vartheta_1, \ldots, \vartheta_2) \in \Theta\colon\ \vartheta_{j+1} = \ldots = \vartheta_m = 0\}$ so that the reduced model $(p_\theta\colon\ \theta \in S_j)$ is of dimension j, $1 \leqslant j < m$. If there exists $1 \leqslant j < m$ such that the true parameter values θ_0 are restricted to the subspace S_j (i. e. if $\theta_0 \in S_{m-1}$), then the rule described by Theorem 6 leads to the simplest possible reduced model with the asymptotically negligible error.

But even if the hypothesis $\mathscr{H}_{m-1}\colon\ \theta_0 \in S_{m-1}$ is false, the viewpoint of model simplicity leads to the need to consider the reduced models of dimensions $1 \leqslant j < m$. A submodel $(P_\theta\colon\ \theta \in S_k)$ is acceptable if the true value θ_0 is in a close neighborhood of S_k (the hypothesis $\mathscr{H}_k^*$) and if all similarly defined hypotheses $\mathscr{H}_j^*$, $1 \leqslant j < k$, are false.

We adopt this approach, with the global informations $I_{\theta_0}(S_j)$ serving as measures of proximity of θ_0 and S_j. Therefore we consider for a small $\tau > 0$ the hypotheses

$$\mathscr{H}_j\colon\ I_{\theta_0}(S_j) \leqslant -\tau \quad \text{for } 1 \leqslant j \leqslant m-1.$$

These hypotheses coincide with the above considered $\mathscr{H}_j$ for S_j replaced by

$$S_j^* = \{\theta \in \Theta\colon\ I_\theta(S_j) \leqslant -\tau\}.$$

Further, (12) implies for every θ_1, θ_2 and θ

$$I(\theta_1; \theta) - I(\theta_2; \theta) = I(\theta_1; \theta_2) + (\tau(\theta_2) - \tau(\theta_1))\,(\theta - \theta_2)^t$$

so that $\|\theta_1 - \theta_2\| \to 0$ implies $\|\mathrm{argmin}_S\, I(\theta_1; \theta) - \mathrm{argmin}_S\, I(\theta_2; \theta)\| \to 0$ for convex sets $S = S_j = \partial S_j$. It follows from here that the sets S_j^* are relatively closed in Θ.

Thus Theorem 6 implies that the criterion of dimension reduction defined by (23) with $\Gamma_{n,j}$ replaced by

$$\Gamma^*_{n,j} = I_{\hat{\theta}_n}(S^*_j)$$

is asymptotically errorless provided that ε_n figuring in (23) satisfy the assumptions of Theorem 5.

An open problem in the just formulated criterion is the numerical evaluation of $I_{\hat{\theta}_n}(S^*_j)$. Since the boundaries ∂S^*_j are not as simple as ∂S_j,

$$\inf_{\partial S^*_j} I(\hat{\theta}_n;\theta) = \pm I_{\hat{\theta}_n}(S^*_j)$$

are less easily evaluated than

$$\inf_{\partial S_j} I(\hat{\theta}_n;\theta) = -I_{\hat{\theta}_n}(S_j) \quad \text{a. s. for } \theta_0 \in S^c_{m-1}.$$

Therefore we formulate an alternative criterion based just on the statistics $I_{\hat{\theta}_n}(S_j)$.

In the rest of the paper we use the notation

$$I_j = |I_{\theta_0}(S_j)|, \quad \Delta I_j = I_j - I_{j+1}, \quad \varphi(j) = \frac{\Delta I_j}{I_j} \tag{25}$$

and

$$\Gamma_{n,j} = |I_{\hat{\theta}_n}(S_j)|, \quad \Delta \Gamma_{n,j} = \Gamma_{n,j} - \Gamma_{n,j+1}, \quad \psi(j) = \frac{\Delta \Gamma_{n,j}}{\Gamma_{n,j}}. \tag{26}$$

Since we restrict ourselves to $\theta_0 \in S^c_{m-1}$, functions I_j and $\Gamma_{n,j}$ are decreasing in the domain $1 \leqslant j \leqslant m-2$.

Let us first consider for some $\delta_n > 0$ the decision criterion

$$\Phi_n = \min\{j\colon \Gamma_{n,j} < \tau + \delta_n\} \quad \text{(cf. (23))}. \tag{27}$$

By (22), probability of error is defined by the formula

$$P(\Phi_n, \theta_0) = \mathsf{P}(\Phi_n \neq \max\{j\colon I_j \geqslant \tau\})$$

and Theorem 6 implies the following result.

Theorem 7. *If the sequence $\varepsilon_n = -\delta_n$ satisfies the assumptions of Theorem 5 then*

$$\lim_n P(\Phi_n, \theta_0) = 0 \quad \textit{for every } \theta_0 \in S^c_{m-1}.$$

The disadvantage of this information criterion is that the decisions Φ_n depend very strongly on the parameter $\tau > 0$, so that the models chosen for, say, $\tau = 10^{-1}$ and $\tau = 10^{-2}$ are often different. Therefore we make precise a variant of this criterion using local minima of the relative increments $\psi(j)$.

Let $\psi(j)$, $1 \leqslant j \leqslant m-1$ be the function defined by (26). If this function has no local minimum in the subdomain $1 < j < m-1$ then we put $\Phi_n = m-1$. Otherwise Φ_n equals the largest local minimum point $1 < j < m-1$. This is the *global information criterion* (GIC) to which we refer in the sequel.

If $1 < k < m-1$ is a local minimum of $\psi(j)$ then the values $\Gamma_{n,j}$ decrease rapidly (exponentially) in the domain $j_1 \leqslant j \leqslant k$ and slowly in the domain $k < j \leqslant j_2$ where $1 \leqslant j_1 < k$ and $k < j_2 \leqslant m-1$ are the closest local maxima of $\psi(j)$. This means that if k is the largest local minimum then $\Gamma_{n,k}$ can be expected to be small. By (21), for large n this means that with a high probability I_k will be small, i. e. that the hypothesis $\mathscr{H}_k^*$ with a small $\tau > 0$ will be very likely.

The GIC obviously differs from the information criterion proposed by Akaike (1973) (AIC, see also the references given in Section 1). It also differs from the other familiar criteria mentioned in Section 1. In the rest of section we consider an example serving for the demonstration of properties of the AIC throughout the book of Sahamoto et al. (1986). We will see that in this example the decision of GIC coincides with that of AIC.

E x a m p l e 3. Let us consider the sample space $\mathscr{X} = (0,1) \times \mathbb{R}$ and the exponential density

$$p_\theta(x_1, x_2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-[x_2 - \mu(x_1, \boldsymbol{a})]^2/2\sigma^2}, \quad \theta = (\sigma, \boldsymbol{a}) \in (0,\infty) \times \mathbb{R}^{21},$$

on this space, where $\boldsymbol{a} = (a_0, a_1, \ldots, a_{20})$ and

$$\mu(x, \boldsymbol{a}) = a_0 + \sum_{r=1}^{10} [a_{2r-1} \sin(2\pi r x) + a_{2r} \cos(2\pi r x)].$$

The vector θ from $\Theta = (0,\infty) \times \mathbb{R}^{21} \subset \mathbb{R}^{22}$ is assumed to be organized into 11 bivariate components $\vartheta_1, \ldots, \vartheta_{11}$, where

$$\vartheta_1 = (\sigma, a_0) \in (0,\infty) \times \mathbb{R}$$

and

$$\vartheta_{r+1} = (a_{2r-1}, a_{2r}) \in \mathbb{R}^2 \quad \text{for } 1 \leqslant r \leqslant 10.$$

By the result of the fifth example of Section 3, we have for $\theta_0 = (\sigma_0, \boldsymbol{a}_0)$ and $\theta = (\sigma, \boldsymbol{a})$ that

$$I(\theta_0;\theta) = \frac{1}{2}\int_0^1 \left[\frac{(\mu(x_1,\boldsymbol{a}_0)-\mu(x_1,\boldsymbol{a}))^2}{\sigma^2} + \frac{\sigma_0^2}{\sigma^2} - 1 - \ln\frac{\sigma_0^2}{\sigma^2}\right] \mathrm{d}x_1$$

$$= \frac{1}{2}\left[\frac{\sum\limits_{r=0}^{10}(a_{0r}-a_r)^2}{\sigma^2} + \frac{\sigma_0^2}{\sigma^2} - 1 - \ln\frac{\sigma_0^2}{\sigma^2}\right]. \tag{28}$$

Consider now the reduced models with parameter subspaces

$$S_j = \{\theta \in \Theta\colon\ \vartheta_{j+1} = \ldots = \vartheta_{11} = 0\} \equiv \{\sigma, \boldsymbol{a}\colon\ a_{2j-1} = \ldots = a_{20} = 0\}$$

for $1 \leqslant j \leqslant 10$ and $S_0 = \emptyset, S_{11} = \Theta$. These models are considered throughout Chapter 4 in Sahamoto et al. (1986). For the same $\theta_0 = (\sigma_0, \boldsymbol{a}_0) \in \Theta - S_{10}$ and its ML-estimate $\hat\theta = (\hat\sigma, \hat{\boldsymbol{a}})$ based on $n = 500$ samples as considered on p. 60 ibid. (and given in the bold letters in Tables 1 and 2 below), we have computed the vectors $\tilde\theta_j = (\tilde\sigma_j, \tilde{\boldsymbol{a}}_j) \in S_j$ which minimize the I-divergence $I(\hat\theta;\theta)$ on $\partial S_j = S_j$.

If we define for every $1 \leqslant j \leqslant 10$

$$\tilde{\boldsymbol{a}}_j = (\hat a_0, \ldots, \hat a_{2j-2}, 0, \ldots, 0) \quad \text{and} \quad \tilde\sigma_j^2 = \sum_{i=2j-1}^{20} \hat a_i^2 + \hat\sigma^2 \stackrel{\triangle}{=} \hat\Delta_j^2 + \hat\sigma^2$$

then $\tilde\theta = (\tilde\sigma, \tilde{\boldsymbol{a}}) \in S_j$, and, by (28),

$$\inf_{\theta\in S_j} I(\hat\theta;\theta) = \inf_{(\sigma,\boldsymbol{a})\in S_j} I((\hat\sigma,\hat{\boldsymbol{a}});(\sigma,\boldsymbol{a}))$$

$$= \inf_{\sigma>0,\ (a_0,\ldots,a_{2j-2})\in\mathbb{R}^{2j-1}} \frac{1}{2}\left[\frac{\sum\limits_{i=0}^{2j-2}(\hat a_i - a_i)^2 + \sum\limits_{i=2j-1}^{20}\hat a_i^2 + \hat\sigma^2}{\sigma^2} - 1 - \ln\frac{\hat\sigma^2}{\sigma^2}\right]$$

$$= \inf_{\sigma>0}\frac{1}{2}\left[\frac{\sum\limits_{i=2j-1}^{20}\hat a_i^2 + \hat\sigma^2}{\sigma^2} - 1 - \ln\frac{\hat\sigma^2}{\sigma^2}\right] \qquad (\text{attained at } \boldsymbol{a} = \tilde{\boldsymbol{a}}_j)$$

$$= \frac{1}{2}\ln\left(\frac{\hat\sigma^2 + \sum\limits_{i=2j-1}^{20}\hat a_i^2}{\hat\sigma^2}\right) = \ln\sqrt{\frac{\hat\sigma^2 + \hat\Delta_j^2}{\hat\sigma^2}} \qquad (\text{attained at } \sigma^2 = \tilde\sigma_j^2)$$

$$= I((\hat a, \hat{\boldsymbol{a}});\ (\tilde\sigma_j\ \tilde{\boldsymbol{a}}_j)) = I(\hat\theta, \tilde\theta_j).$$

j	$\hat{a}_{2j-1}$	$\hat{a}_{2j}$	$\tilde{\sigma}_j^2 = \hat{\sigma}^2 + \hat{\Delta}_j^2$ ($\tilde{\sigma}_0^2 \equiv \hat{\sigma}^2$)	$\Gamma_{n,j} = \lvert I_{\hat{\theta}}(S_j)\rvert = \ln \sqrt{\frac{\tilde{\sigma}_j^2}{\hat{\sigma}^2}}$	$\Delta\,\Gamma_{n,j} = \Gamma_{n,j} - \Gamma_{n,j+1}$	$\psi(j) \times 100 = \frac{\Delta\,\Gamma_{n,j}}{\Gamma_{n,j}} \times 100$
0	—	8.011	0.901	∞	—	—
1	2.465	−3.847	28.20	1.722	0.675	39
2	2.249	−1.009	7.319	1.047	0.886	85
3	0.515	0.007	1.243	0.161	0.120	75
4	−0.139	−0.073	0.978	0.041	0.013	32
5	0.084	0.021	0.954	0.028	0.003	11
6	−0.060	0.044	0.946	0.025	0.003	12
7	−0.075	−0.011	0.940	0.022	0.004	18
8	0.142	−0.016	0.935	0.018	0.011	61
9	−0.035	0.039	0.914	0.007	0.001	14
10	0.040	−0.089	0.911	0.006	—	—

Table 1: Values of $\hat{\theta}_n$ and quantities depending on $\hat{\theta}_n$.

The values $\tilde{\sigma}_j^2 = \hat{\sigma}^2 + \hat{\Delta}_j^2$ and $\hat{\sigma}^2$ are tabulated for $1 \leqslant j \leqslant 10$ in Table 1 below, together with the minimal I-divergences

$$\Gamma_{n,1} = I(\hat{\theta}; \tilde{\theta}_1) > \ldots > \Gamma_{n,10} = I(\hat{\theta}; \tilde{\theta}_{10})$$

and with quantities $\Delta\,\Gamma_{n,j}$ and $\psi(j)$ defined by (26).

We see from Table 1 and from the full line graph in Figure 1 that the GIC reduces the dimension to $j = 5$, i. e. that the best in the stated sense is the exponential submodel with the parameter space S_5. This submodel has also been calculated on p. 75 in Sahamoto et al. (1986) as the best in the sense of Akaike's AIC.

The values I_k, $\Delta\,I_k$ and $\varphi(k)$ defined by (25) and obtained by replacing in the formulas above the maximum likely parameters $\hat{\theta} = (\hat{\sigma}, \hat{\boldsymbol{a}})$ by the true values $\theta_0 = (\sigma_0, \boldsymbol{a}_0)$, are presented in Table 2 and Figure 1 (the interrupted line). We see two local minima of $\varphi(j)$ in the domain $1 < j < 9$. One at $j = 5$ as in the case of $\psi(k)$, and the other at $j = 7$. The relative loss of information caused by dropping out the coordinate pair $\theta_7 = (a_{0,11}, a_{0,12}) = (-0.011, 0.042)$ is considerable, but the negative information $I_{\theta_0}(S_6) = -I_6$ is still negligible, of order 10^{-3}. Thus the submodel with the parameter subspace S_7 does not seem to have much advantage over that with S_6. Since the information distance I_5 is of order 10^{-3} too, but I_4 is already of order 10^{-2}, the above considered GIC decision for S_5 seems to be justified also by the direct analysis of global information distance between θ_0 and the subspaces S_j, $1 \leqslant j \leqslant 10$.

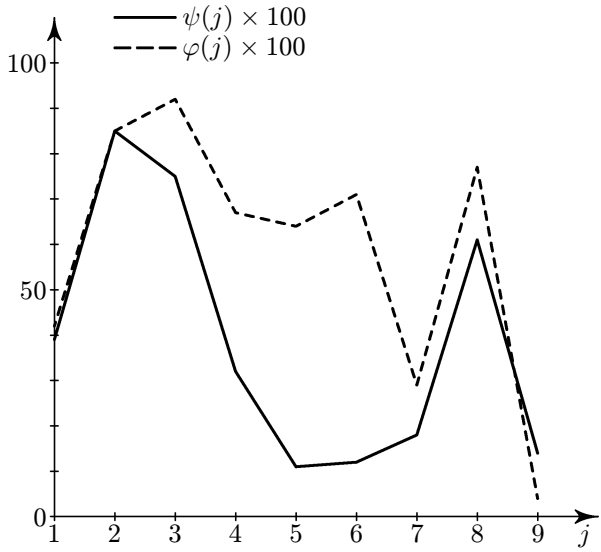

Figure 1. Graphs of the relative information distance increments $\psi(j)$ and $\varphi(j)$.

j	$a_{0,2j-1}$	$a_{0,2j}$	$\tilde{\sigma}_j^2 = \sigma_0^2 + \Delta_j^2$ $(\tilde{\sigma}_0^2 \equiv \sigma_0^2)$	$I_j = \lvert I_{\theta_0}(S_j)\rvert$ $= \ln \sqrt{\frac{\sigma_j^2}{\sigma_0^2}}$	$\Delta I_j = I_j - I_{j+1}$	$\varphi(j) \times 100$ $= \frac{\Delta I_j}{I_j} \times 100$
0	—	8.000	1.000	∞	—	—
1	2.415	−3.806	27.13	1.650	0.691	42
2	2.119	−0.997	0.959	0.819	0.679	85
3	0.545	0.069	1.324	0.140	0.129	92
4	−0.094	−0.078	1.022	0.011	0.007	67
5	−0.021	−0.065	1.007	0.004	2.7×10^{-3}	64
6	−0.011	0.042	1.003	1.3×10^{-3}	9.2×10^{-4}	71
7	−0.011	0.010	1.000	3.8×10^{-4}	1.1×10^{-4}	29
8	0.020	−0.004	1.000	2.7×10^{-4}	2.1×10^{-4}	77
9	0.002	0.001	1.000	6.1×10^{-5}	3×10^{-6}	4
10	−0.006	−0.009	1.000	5.8×10^{-5}	—	—

Table 2: Values of θ_0 and quantities depending on θ_0.

The question is to what extent the above achieved coincidence of decisions by the criteria GIC and AIC depends on the set of data randomly generated for the example in Sahamoto et al. (1986). To answer this question, Nikolov (1996) simulated new data of the same sample size $n = 500$ with the aim to compare the decisions by GIC and AIC. The performances of both criteria were found to be similar, slightly

in favour of GIC. Namely, the average difference between decisions by AIC and GIC observed in 2000 simulated samples was +0.485. Since it is known that AIC overestimates the "quasitrue model" (see p. 402 in Nishii (1988)), this result favorizes the GIC. Moreover, the observed dispersion of decisions by GIC was significantly smaller.

References

[1] *Akaike, H.*: Information theory and an extension of the maximum likelihood principle. Proceedings of the Second International Symposium on Information Theory (B. N. Petrov et al., eds.). Akademiai Kiado, Budapest, 1973, pp. 267–281.

[2] *Bauer, P., Pötscher B. M. and Hackl P.*: Model selection by multiple test procedures. Statistics *19* (1988), 39–44.

[3] *Berlinet, A. and Francq, Ch.*: Identification of a univariate ARMA model. Comp. Statist. *9* (1994), 117–133.

[4] *Brown, L. D.*: Fundamentals of Statistical Exponential Families. Inst. of Mathem. Statist, Hayword, California, 1986.

[5] *Chernoff, H.*: On the distribution of likelihood ratio. Ann. Math. Statist. *25* (1954), 573–578.

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

[7] *Lehmann, E. L.*: Testing Statistical Hypotheses. 2nd edition, Wiley, New York, 1986.

[8] *Liese, F. and Vajda, J.*: Convex Statistical Distances. Teubner, Leipzing, 1987.

[9] *Nikolov, V.*: Regression and Autoregression Models of Signals and their Recognition by Neural Nets. Diploma Theses. Faculty of Physical and Nuclear Engineering, Czech Tech. University, Prague, 1996. (In Czech.)

[10] *Nishii, R.*: Maximum likelihood principle and model selection when the true model is unspecified. Journal of Multivariate Analysis *27* (1988), 392–403.

[11] *Pötscher, B. M.*: Order estimation in Arma–models by Lagrangian multiplier test. Ann. Statist. *11* (1983), 872–885.

[12] *Rissanen, J.*: Modelling by shortest data description. Automatica *14* (1978), 465–471.

[13] *Rockefellar, R. T.*: Convex Analysis. Princeton University Press, Princeton, 1970.

[14] *Ronchetti, E.*: Robust model selection. In Transactions of the Twelfth Prague Conference on Information Theory, ... (J. Á. Víšek, and P. Lachout, eds.). Academy of Sciences of the Czech Republic, Prague, 1994, pp. 200–202.

[15] *Rydén, T.*: Estimating the order of hidden Markov models. Statistics *26* (1995), 345–354.

[16] *Sahamoto, Y., Ishiguro, M. and Kitagawa, G.*: Akaike Information Criterion Statistics. Reidel, Dordrecht, 1986.

[17] *Schwartz, G.*: Estimating the dimension of a model. Annals of Statistics *6* (1978), 461–464.

[18] *Spanier, J. and Oldham, K. B.*: An Atlas of Functions. Springer, Berlin, 1987.

[19] *Speed, T. P. and Yu, B.*: Model collection and prediction: Normal regression. Ann. Inst. Statist. Math. *45* (1993), 35–54.

[20] *Vajda, I.*: Global statistical information, likelihood ratio tests and maximum likelihood estimators. Kybernetika (submitted) (1997).

[21] *Vieu, P.*: Order choice of nonlinear autoregressive models. Statistics *26* (1995), 307–328.

Authors' addresses: *I. Vajda*, Institute of Information Theory and Automation, CZ-182 08 Prague, Czech Republic; *E. van der Meulen*, Katolieke Universiteit Leuven, B-3001 Leuven, Belgium.