Akademie věd České republiky
Ústav teorie informace a automatizace

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

RESEARCH REPORT

Igor Vajda and Domingo Morales:

Generalized information criteria for optimal Bayes decisons

No. 2239 November 2008

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
Telex: 122018 atom c, Fax: (+42) (2) 688 4903
E-mail: utia@utia.cas.cz

Research Report No. 2239 December 2008

Generalized information criteria for optimal Bayes decisions

I. Vajda[1] and D. Morales[2]

[1] *Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Prague, Czech Republic. E-mail: vajda@utia.cas.cz*

[2] *Operations Research Center, University Miguel Hernández of Elche, 03202 Elche, Spain. E-mail: d.morales@umh.es*

This paper deals with Bayesian models given by statistical experiments and common types of loss functions. Probability of error of the Bayes identificator of state, and of more general types of Bayes risk are characterized by means of classical and generalized information criteria applicable to the experiment. In particular, the accuracy of the approximation is studied. A number of concrete numerical results and figures illustrate the obtained theoretical results.

Key words: Shannon entropy, Alternative Shannon entropy, Power entropies, Alternative power entropies, Bayes errors, Bayes risks, Sub-Bayes risks.

1 Introduction

In Morales, Pardo and Vajda (1996), we systematically studied *generalized measures of uncertainty* of stochastic systems with finite or countable state spaces Θ and probability distributions π on Θ, and *generalized measures of informativity* of random observations X with sample probability spaces $(\mathcal{X}, \mathcal{S}, P)$ and posterior distributions π_x on Θ when $X = x \in \mathcal{X}$. We investigated the general entropies $H(\pi)$ as appropriate concave or Schur concave functions of stochastic vectors π. As general *characteristics of informativity* of the whole stochastic observation experiment

$$\mathcal{E} = \langle (\Theta, \pi),\ (\mathcal{X}, \mathcal{S}, P) \rangle \tag{1.1}$$

we proposed the corresponding conditional entropies

$$H(\mathcal{E}) = \int_{\mathcal{X}} H(\pi_x)\mathrm{d}P(x) \tag{1.2}$$

closely related to the general information measures

$$I(\mathcal{E}) = H(\pi) - H(\mathcal{E}). \tag{1.3}$$

Particular attention was paid to the entropies of the form

$$H_\phi(\pi) = \sum_{\theta\in\Theta} \phi(\pi(\theta)) \tag{1.4}$$

for concave functions $\phi(t)$, $0 \le t \le 1$.

For $\phi(t) = -t\log t$ we obtain from (1.4) the classical Shannon entropy

$$H_1(\pi) = -\sum_{\theta\in\Theta} \pi(\theta)\ln \pi(\theta) \tag{1.5}$$

and from (1.2) and (1.3) the classical Shannon conditional entropy and Shannon information. For $\phi(t) = t(1-t)$ we obtain from (1.4) the alternative to the Shannon entropy

$$H_2(\pi) = 1 - \sum_{\theta\in\Theta} \pi^2(\theta) \tag{1.6}$$

called the *quadratic entropy* by Vajda (1968), and from (1.2) and (1.3) the corresponding quadratic conditional entropy $H_2(\mathcal{E})$ and quadratic information $I_2(\mathcal{E})$. In fact, Cover and Hart (1967) and Vajda (1968) introduced independently $H_2(\mathcal{E})$ as a measure of quality of decisions concerning the states $\theta \in \Theta$ achievable on the basis of observations X in the statistical experiments $\mathcal{E}$. For example, the probability of error $P_e(\mathcal{E})$ of the Bayes decisions $\delta_B : \mathcal{X} \longmapsto \Theta$ was estimated in Vajda (1968) as follows

$$H_2(\mathcal{E})\left(1 + \sqrt{1 - H_2(\mathcal{E})}\right) \le P_e(\mathcal{E}) \le H_2(\mathcal{E}) \tag{1.7}$$

so that the smooth information criterion $H_2(\mathcal{E})$ can replace the less easily evaluated $P_e(\mathcal{E})$ in the feature selection processes with low levels of error probability.

The quadratic entropy (1.6) requires the operation of multiplication and summation, and is thus computationaly simpler than the Shannon entropy (1.5) and also than the more general entropies of Rényi (1961)

$$\breve{H}_\alpha(\pi) = \frac{1}{\alpha - 1}\ln \sum_{\theta\in\Theta} \pi^\alpha(\theta), \quad \alpha > 0, \alpha \ne 1 \tag{1.8}$$

containing the Shannon entropy as the special limit case $H_1(\pi) = \breve{H}_1(\pi) \overset{\triangle}{=} \lim_{\alpha\to 1} \breve{H}_\alpha(\pi)$. Rényi introduced the entropies axiomatically by extending and parametrizing by α the additivity rule in the axioms used earlier by Faddeev (1957) to characterize the Shannon's $H_1(\pi)$. However, he emphasized also the alternative "pragmatic approach" to motivate

$H_1(\pi)$ and its extensions as characteristics of various statistical decision problems. In this sense for example Kovalevsky (1965) used $H_1(\mathcal{E})$ to obtain similar bounds as (1.7) to characterize the error probability $P_e(\mathcal{E})$ in pattern recognition problems which inspired among other the work of Vajda (1968). The bounds of Kovalevsky were later reinvented and applied in different areas of statistical decisions and information processing by several authors, e.g. Tebbe and Dwyer (1968) or Feder and Merhav (1994).

By appropriately modifying the extended additivity rule of Rényi (1961), Havrda and Charvát (1967) axiomatically introduced the one-one modification of the power entropies of Rényi,

$$H_\alpha(\pi) = \frac{1}{\alpha - 1}\left(1 - \sum_{\theta\in\Theta} \pi^\alpha(\theta)\right), \quad \alpha > 0, \alpha \neq 1 \tag{1.9}$$

with the limit $H_1(\pi) = \lim_{\alpha\to 1} H_\alpha(\pi)$. Vajda (1969) used the generalized informativity $H_\alpha(\mathcal{E})$ obtained by employing $H_\alpha(\pi)$ in (1.2) to evaluate bounds of the type (1.7) and proposed $H_\alpha(\mathcal{E})$ as a generalized feature extraction criterion. His criterion was used later by many authors, e.g. Kanal (1974), Devijver and Kittler (1982) or Devroye et al. (1996), and his bounds of the type (1.7) were later completed, modified or tightened by Salichov (1994), Toussaint (1977), Ben Bassat (1978), Ben Bassat and Raviv (1978) and Harremoes and Topsoe (2001).

Vajda and Vašek (1985) found a method for obtaining attainable bounds of the type (1.7) for arbitrary Schur concave entropy (1.2) applied later in Morales, Pardo and Vajda (1996). Here we use the results of these two papers to obtain some new attainable bounds for the probability of error $P_e(\mathcal{E})$ and apply these bounds to approximate Bayes risks $R_B(\mathcal{E})$ achieved in given experiments $\mathcal{E}$ for the most common types of loss functions. We address also the problem which information criteria provide the most accurate approximations of probabilities of errors and Bayes risks.

2 General loss modelL

Consider the classical model of Bayesian decision theory (cf. e.g. Berger (1986)) with state of nature θ from a finite set Θ, prior probability distributions of states $\pi = (\pi(\theta) > 0 : \theta \in \Theta)$ and observations (random samples) X conditionally distributed by probability measures P_θ on a measurable observation space $(\mathcal{X}, \mathcal{S})$ depending on the states $\theta \in \Theta$. We restrict ourselves to the important situation where the purpose of decision is identification of the unknown state θ. Thus our decisions (actions in the sense of Berger) are states θ from the action space Θ, and the loss functions are of the form

$$L : \Theta \times \Theta \mapsto [0, \infty), \quad L(\theta, \theta) = 0. \tag{2.10}$$

Thus we deal with the Bayesian model given by a statistical experiment

$$\mathcal{E} = \langle \pi, \mathcal{P} = \{P_\theta : \theta \in \Theta\}\rangle \tag{2.11}$$

and a loss function (2.10).

This is the standard decision-theoretic model of many real situations, in particular of the

(1) *pattern recognition* where the states of nature θ represent various possible patterns (images) and $L(\theta, \hat{\theta}) > 0$ is the loss incurred by the wrong identifications $\hat{\theta}$ of these patterns,
(2) *classification* where the states θ represent various classes of objects and $L(\theta, \hat{\theta}) > 0$ is the loss of misclassification
(3) *information transmission* where the states θ represent various possible messages transmitted via communication channel $(\Theta, \{P_\theta : \theta \in \Theta\}, \mathcal{X})$ with input alphabet Θ, output alphabet $\mathcal{X}$ and transition probability distributions P_θ describing distortion of messages by the channel noise.

These concrete interpretations and their various combinations appear also in the *detection theory* and *stochastic control theory.*

Let us briefly review basic concepts of Bayesian decision theory applicable in the present model. *Expected loss* of an individual identification action $\hat{\theta} \in \Theta$ is

$$\mathcal{L}(\pi, \hat{\theta}) = \sum_{\theta \in \Theta} L(\theta, \hat{\theta})\, \pi(\theta). \tag{2.12}$$

Each individual action $\theta_\pi \in \Theta$ with the property

$$\theta_\pi = \mathrm{argmin}_{\hat{\theta}}\, \mathcal{L}(\pi, \hat{\theta}) \tag{2.13}$$

is said to be *Bayes action* (Bayes decision without data) and the minimal a priori expected loss

$$L_B(\pi) = \mathcal{L}(\pi, \theta_\pi) \tag{2.14}$$

is a *prior Bayes loss.* Observation data $x \in \mathcal{X}$ are assumed to be used for identification by means of *identification rules*

$$\delta = \mathcal{X} \mapsto \Theta. \tag{2.15}$$

Technically, they are assumed to be $\mathcal{S}$-measurable and P_θ-integrable for all $\theta \in \Theta$. *Risk function* of the identification rule (2.15) is

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x))\, dP_\theta(x), \quad \theta \in \Theta$$

and its expected value

$$\mathcal{R}(\pi, \delta) = \sum_{\theta \in \Theta} R(\theta, \delta)\pi(\theta) = \sum_{\theta \in \Theta} \int_{\mathcal{X}} L(\theta, \delta(x))\pi(\theta)\, dP_\theta(x) \tag{2.16}$$

is simply a *risk.* The minimizer

$$\delta_B = \mathrm{argmin}_\delta\, \mathcal{R}(\pi, \delta) \tag{2.17}$$

is the *Bayes identification rule* and

$$R_B = \mathcal{R}(\pi, \delta_B) \tag{2.18}$$

the *Bayes risk* of identification in the model under consideration.

It is known that in this model the Bayes identification rule exists and is given by a relatively simple explicit formula. To demonstrate this and to find the Bayes identification rule formula, take first into account the marginal probability distribution

$$P = \sum_{\theta \in \Theta} \pi(\theta) P_\theta \tag{2.19}$$

on the observation space $(\mathcal{X}, \mathcal{S})$ which dominates each conditional distribution P_θ in the sense $P(S) = 0$ implies $P_\theta(S) = 0$ for $S \in \mathcal{S}$. Hence there exists the Radon-Nikodym density

$$p_\theta(x) = \frac{dP_\theta(x)}{dP(x)}$$

defined for all data $x \in \mathcal{X}$, with values uniquely given except possibly a set $S_\theta \in \mathcal{S}$ with $P(S_\theta) = 0$ (i.e. for P-almost all in symbols P-a.e. on $\mathcal{X}$). Then

$$\pi_x = (\pi_x(\theta) \overset{\triangle}{=} \pi(\theta) p_\theta(x) : \theta \in \Theta) \tag{2.20}$$

is the conditional (posterior) probability distribution on Θ given data x. Indeed, by the definition of Radon-Nikodym densities, $p_\theta(x)$

$$\min_\theta \pi_x(\theta) \geq 0 \quad \text{and} \quad \sum_\theta \pi_x(\theta) = \frac{dP(x)}{dP(x)} = 1 \qquad P\text{-a.e. on } \mathcal{X}.$$

Obviously, the statistical experiment (2.11) is equivalently described by the conditional distributions (2.20) for $x \in \mathcal{X}$ and the marginal distribution (2.19),

$$\mathcal{E} = \langle \pi, \mathcal{P} = \{P_\theta : \theta \in \Theta\} \rangle \equiv \langle P, \Pi = \{\pi_x : x \in \mathcal{X}\} \rangle. \tag{2.21}$$

Using the posterior distribution (2.20) and the concept of expected loss (2.12), we can rewrite the risk formula (2.16) into the simple form

$$\mathcal{R}(\pi_x, \delta) = \int_{\mathcal{X}} \mathcal{L}(\pi_x, \delta(x))\, dP(x). \tag{2.22}$$

From here and from (2.17) we see that an identification rule δ is Bayes (in symbols $\delta = \delta_B$) if and only if for P-almost all data $x \in \mathcal{X}$ the data based action $\delta_B(x)$ is Bayes for the posterior distribution, π_x, i.e. coincides with some θ_{π_x} defined in accordance with (2.13). Thus the Bayes identification rule can equivalently be defined by the formula

$$\delta_B(x) = \theta_{\pi_x} \qquad P\text{-a.e. on } \mathcal{X} \tag{2.23}$$

From here we deduce also that the Bayes risk R_B is the expected *posterior Bayes loss* given data x, denoted $L_B(\pi_x)$ and defined by (2.14) with the prior distribution π replaced by the posterior distribution π_x. In other words, we deduce that

$$\begin{aligned} R_B = \mathcal{R}(\pi, \delta_B) &= \int_{\mathcal{X}} \mathcal{L}(\pi_x, \theta_{\pi_x})\, dP(x) \qquad \text{(cf. (2.22), (2.23))} \\ &= \int_{\mathcal{X}} L_B(\pi_x)\, dP(x). \end{aligned} \tag{2.24}$$

3 Zero-one loss model

A prominent role in the applications of the model of previous section plays the error loss function

$$L_e : \Theta \times \Theta \mapsto \{0,1\}, \qquad L_e(\theta,\hat{\theta}) = \begin{cases} 1 & \text{if } \hat{\theta} \neq \theta, \\ 0 & \text{if } \hat{\theta} = \theta. \end{cases} \tag{3.25}$$

Here the general expected loss $\mathcal{L}(\pi,\hat{\theta})$ reduces to the *prior probability of error* of the identification action $\hat{\theta} \in \Theta$,

$$\mathcal{L}_e(\pi,\hat{\theta}) = \sum_{\theta\in\Theta} L_e(\theta,\hat{\theta})\pi(\theta) = 1 - \pi(\hat{\theta}) \tag{3.26}$$

The Bayes identification action θ_π thus minimizes this probability of error over $\hat{\theta} \in \Theta$. This means that the prior Bayes expected loss $L_B(\pi)$ given by (2.14) is the minimal prior probability of error given by the formula

$$e_B(\pi) = 1 - \pi(\theta_\pi), \tag{3.27}$$

and called simply *prior Bayes error*. Similarly the posterior. Bayes expected loss $L_B(\pi_x)$ for data $x \in \mathcal{X}$ is in this case the minimal posterior probability of error

$$e_B(\pi_x) = 1 - \pi_x(\theta_{\pi_x}) \tag{3.28}$$

called simply *posterior Bayes error*, as the Bayes identification action $\theta_{\pi_x} \in \Theta$ minimizes over $\hat{\theta} \in \Theta$ the posterior error probability $1 - \pi(\hat{\theta})$. Finally by (2.24) and the equality $L_B(\pi_x) = e_B(\pi_x)$, the Bayes risk R_B coincides with the *Bayes error* (average minimal posterior probability of error)

$$e_B = \int_{\mathcal{X}} e_B(\pi_x)\, dP(x). \tag{3.29}$$

As mentioned in the introduction, our intention is to evaluate or estimate performances of Bayes identification rules in the general loss function models by means of known performances of such rules in the simpler error loss function models. The rest of this section is devoted to the research of this eventuality. The achieved results serve in the next section to establish new bounds for the Bayes risk R_B based partly on the bounds for the Bayes error probability e_B established in previous literature and partly on new such bounds established in the next section.

Put in the general loss model (2.10)

$$\begin{aligned} L^+ &= \max\{L(\theta,\hat{\theta}) : \theta,\hat{\theta} \in \Theta\}, \\ L^- &= \min\{L(\theta,\hat{\theta}) : \theta,\hat{\theta} \in \Theta,\, L(\theta,\hat{\theta}) > 0\}, \end{aligned}$$

and impose the nontrivial condition $L^+ > 0$. Further, denote by

$$L^0 = \frac{L^+ + L^-}{2} > 0$$

the *median positive loss* and by

$$\Delta = (L^+ - L^0)100 = (L^0 - L^-)100$$

the *positive loss dispersion* in %. Obviously, $\Delta = 0$ if and only if $L(\theta, \hat{\theta})$ is proportional to the zero-one loss function $L_e(\theta, \hat{\theta})$.

Example 1. The error loss function L_e of (3.25) leads to $L^+ = L^- = 1$ so that the median loss is $L^0 = 1$ and the loss dispersion is $\Delta = 0\%$.

Example 2. Consider the state space $\Theta = \{1, \dots, n\}$, the loss function (2.10) given in the matrix form

$$(L(\theta, \hat{\theta}))_{\theta, \hat{\theta}=1}^{n} = \begin{pmatrix} 0 & 4/5 & 4/5 & \dots & 4/5 & 6/5 \\ 4/5 & 0 & 4/5 & \dots & 4/5 & 6/5 \\ 4/5 & 4/5 & 0 & \dots & 4/5 & 6/5 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 4/5 & 4/5 & 4/5 & \dots & 0 & 6/5 \\ 6/5 & 6/5 & 6/5 & \dots & 6/5 & 0 \end{pmatrix}$$

where the unknown state $1 \le i \le n-1$ (say a battlefield decision of an enemy) means an artillery attacks of a type i and the state n means an aircraft attack. Here

$$L^+ = 6/5 > L^0 = 1 > L^- = 4/5 \quad \text{and} \quad \Delta = 20\%.$$

Theorem 1. Let the general loss model of Section 3 satisfy the condition $L(\theta, \hat{\theta}) > 0$ for $\theta \neq \hat{\theta}$. If the median loss is L^0 and the loss dispersion is $\Delta \geq 0$, then

(i) the prior Bayes loss L_B and the prior Bayes error e_B satisfy the relation

$$|L_B(\pi) - L^0 \, e_B(\pi)| \le \frac{\Delta}{100} \, e_B(\pi),$$

(ii) for P-almost all $x \in \mathcal{X}$, the posterior Bayes loss $L_B(\pi_x)$ and the posterior Bayes error $e_B(\pi_x)$ satisfy the relation

$$|L_B(\pi_x) - L^0 \, e_B(\pi_x)| \le \frac{\Delta}{100} \, e_B(\pi_x), \tag{3.30}$$

(iii) the Bayes risk R_B and the Bayes error satisfy the relation

$$|R_B - L^0 \, e_B| \le \frac{\Delta}{100} \, e_B.$$

Proof. (I) If $\theta \neq \hat{\theta}$ then by assumption $L(\theta, \hat{\theta}) > 0$ so that $L(\theta, \hat{\theta}) \in [L^-, L^+]$. If $L(\theta, \hat{\theta}) \in [L^0, L^+]$ then, by definition of L^+ and dispersion Δ,

$$L(\theta, \hat{\theta}) - L^0 \le L^+ - L^0 = \Delta/100.$$

If $L(\theta,\hat{\theta}) \in [L^-, L^0]$ then, similarly,

$$L^0 - L(\theta,\hat{\theta}) \le L^0 - L^- = \Delta/100.$$

Hence

$$|L(\theta,\hat{\theta}) - L^0| \le \Delta/100 \quad \text{for all} \quad \theta \neq \hat{\theta}. \tag{3.31}$$

Now by (2.18) and by the assumptions, for every π and $\hat{\theta} \in \Theta$

$$\mathcal{L}(\pi,\hat{\theta}) = \sum_{\theta \neq \hat{\theta}} L(\theta,\hat{\theta})\pi(\theta). \tag{3.32}$$

Hence (2.20) implies for every π

$$L_B(\pi) = \sum_{\theta \neq \theta_\pi} L(\theta,\theta_\pi)$$

Further, (3.27) implies for every π

$$\sum_{\theta \neq \theta_\pi} \pi(\theta) = e_B(\pi),$$

which is positive by the assumed positivity of all $\pi(\theta)$. Therefore multiplying the left side of (3.31) by $\pi(\theta)/e_B(\pi)$, summing over all $\theta \neq \theta_\pi$ and using the Jensen inequality, we get

$$\left| \frac{1}{e_B(\pi)} \sum_{\theta \neq \theta_\pi} L(\theta,\hat{\theta}) - L^0 \right| \le \frac{\Delta}{100}.$$

Thus it remains to apply (3.32) to complete the proof of (i).

(II) Since π_x, given in Section 2, are shown to be probability distribution on Θ for P-almost all $x \in \mathcal{X}$, (ii) follows from(i).

(III) Integrating both sides of (3.30) over $\mathcal{X}$ with respect to the measure P and using the Jensen inequality, we get

$$\left| \int_{\mathcal{X}} L_B(\pi_x) dP(x) - L^0 \int_{\mathcal{X}} e_B(\pi_x) dP(x) \right| \le \frac{\Delta}{100} \int_{\mathcal{X}} e_B(\pi_x) dP(x).$$

The desired result of (iii) follows from here and from the formulas (2.24) and (3.29). ■

Denote for a while by δ_e the Bayes identifier in the simpler error loss model, to distinguish it from the Bayes identifier δ_B in the general loss model of the previous section. By definition, $\delta_e(x)$ maximizes the posterior probability $\pi_x(\theta)$ on Θ under observation $x \in \mathcal{X}$. Therefore $L(\delta_e(x),\hat{\theta})$ is the lowest loss among all losses $L(\theta,\hat{\theta})$ resulting from the decision $\hat{\theta}$. If we replace in the definition of the Bayes identification $\hat{\theta} = \delta_B(x)$ the posteriori expected loss

$$\mathcal{L}(\pi_x,\hat{\theta}) = \sum_{\theta \in \Theta} L(\theta,\hat{\theta})\pi_x(\theta) \quad \text{(c.f. (2.23) and (2.13))}$$

by the posteriori most probable loss $L(\delta_e(x),\hat{\theta})$ then the corresponding identifier

$$\delta_{SB}(x) = \text{argmin}_{\hat{\theta}} L(\delta_e(x),\hat{\theta}) \tag{3.33}$$

is an interesting alternative to the Bayes $\delta_B(x)$. We call it a *sub-Bayes identifier.* It is simpler than the Bayes identifier since (3.33) minimizes one particular loss function while (2.13) is minimizes the mixture (2.12) of such functions. It may be useful when fast Bayes actions δ_B are required in a model with fixed δ_e and frequently varying loss functions $L(\theta,\hat{\theta})$.

The following Theorem 2 deals with the *sub-Bayes risk*

$$R_{SB} = \mathcal{R}(\pi,\delta_{SB}). \tag{3.34}$$

It assumes less than Theorem 1 and at the same time provides tighter bounds when $e_B(\pi_x) < 1/2$ or $e_B > 1/2$.

Theorem 2. Consider the general loss model of Section 3 with median loss $L^0 > 0$ and loss dispersion $\Delta \geq 0$.

(i) For P-almost all $x \in \mathcal{X}$ the posterior Bayes loss $L_B(\pi_x) = \mathcal{L}(\pi_x,\delta_B(x))$ and the posterior sub-Bayes loss $\mathcal{L}(\pi_x,\delta_{SB}(x))$ satisfy the relation

$$0 \leq \mathcal{L}(\pi_x,\delta_{SB}) - \mathcal{L}(\pi_x,\delta_B) \leq 2e_B(\pi_x)\Delta/100,$$

where $e_B(\pi_x)$ is the posterior Bayes error (3.28).

(ii) The Bayes risk R_B and the sub-Bayes risk R_{SB} satisfy the relation

$$0 \leq R_{SB} - R_B \leq 2e_B\Delta/100,$$

where e_B is the Bayes error (3.29).

Proof. (I) Since for P-almost all $x \in \mathcal{X}$

$$\delta_B(x) = \text{argmin}_{\hat{\theta}} \mathcal{L}(\pi_x,\delta),$$

the left inequality in (i) is clear. By (2.12)

$$\mathcal{L}(\pi_x,\delta_{SB}) = L(\delta_e(x),\delta_{SB}(x))\pi_x(\delta_e(x)) + \xi(x)$$

for

$$\xi(x) = \sum_{\theta\neq\delta_e(x)} L(\theta,\delta_{SB}(x)) \leq (L^0 + \Delta/100)[1-\pi_x(\delta_e(x))].$$

Similarly,

$$\mathcal{L}(\pi_x,\delta_B) = L(\delta_e(x),\delta_B(x))\pi_x(\delta_e(x)) + \eta(x) \leq L(\delta_e(x),\delta_{SB}(x))\pi_x(\delta_e(x)) + \eta(x)$$

for

$$\eta(x) = \sum_{\theta\neq\delta_e(x)} L(\theta,\delta_B(x)) \geq (L^0 + \Delta/100)[1-\pi_x(\delta_e(x))].$$

Therefore

$$\mathcal{L}(\pi_x, \delta_{SB}) - \mathcal{L}(\pi_x, \delta_B) \leq [1 - \pi_x(\delta_e(x))]2\Delta/100$$

and (i) follows from (3.28) and from the fact that $\delta_e(x)$ is the Bayes identifier action $\theta_{\pi_x} \in \Theta$ considered in (3.28).

(II) By (3.27)

$$R_B = \int_{\mathcal{X}} \mathcal{L}(\pi_x, \delta_B(x))\, dP(x)$$

and by (3.34) and (2.16)

$$R_{SB} = \int_{\mathcal{X}} \mathcal{L}(\pi_x, \delta_{SB}(x))\, dP(x).$$

Thus (ii) obviously follows from the already proved inequality in (i) and from the formula (3.29) for the Bayes error e_B. ■

4 Generalized information criteria

In this section 4 we denote by $n = |\Theta|$ the number of parameters in Θ. We study estimates of Bayes errors $e_B(\pi)$, $e_B(\pi_x)$ and e_B (or more generally, the Bayes risks $R_B(\pi)$, $R_B(\pi_x)$, R_B) by means of information criteria represented by measures of uncertainties (entropies) $H(\pi)$, $H(\pi_x)$ and

$$H = \int_{\mathcal{X}} H(\pi_x)\, dP(x)$$

of realizations of states of nature θ from individual stochastic sources (Θ, π), (Θ, π_x), or from systems of such sources $\{(\Theta, \pi_x) : x \in \mathcal{X}\}$ depending on data (samples) x which are realizations of random observations X with the sample space $(\mathcal{X}, \mathcal{S}, P)$. For details about these concepts and notations see sections 2 and 3.

Classical Shannon information criteria are based on the *Shannon entropy* (here measured in *nats* instead of *bits*)

$$H(\pi) = \sum_{\theta \in \Theta} \phi(\pi(\theta)), \quad \phi(t) = -t \ln t.$$

In Section 1 we mentioned their generalizations based on the *power entropies*

$$H_\alpha(\pi) = \sum_{\theta \in \Theta} \phi_\alpha(\pi(\theta)), \qquad \alpha > 0. \tag{4.35}$$

where for $\alpha \neq 1$

$$\phi_\alpha(t) = \frac{t(1 - t^{\alpha-1})}{\alpha - 1} \quad \text{and} \quad \phi_1(t) = \lim_{\alpha \to 1} \phi_\alpha(t) = -t \ln t. \tag{4.36}$$

Hence

$$H_\alpha(\pi) = \frac{1}{\alpha - 1}\left[1 - \sum_{\theta \in \Theta} \pi(\theta)^\alpha\right] \quad \text{if} \quad \alpha \neq 1$$

and

$$H_1(\pi) = \lim_{\alpha\to 1} H_\alpha(\pi) = -\sum_{\theta\in\Theta} \pi(\theta)\ln \pi(\theta).$$

As argued in Morales, Pardo and Vajda (1996), the desired information-theoretic properties of the power entropies follow from the concavity of functions $\phi_\alpha(t)$ on $[0,1]$ and from their extremal values $\phi_\alpha(0) = \phi_\alpha(1) = 0$. As an example we can take the *information processing property*

$$0 = H_\alpha(\pi_D) \le H_\alpha(\pi T^{-1}) \le H_\alpha(\pi) \le H_\alpha(\pi_U) = (n - n^{1-\alpha})/(\alpha - 1),$$

where $T : \Theta \mapsto \mathcal{T}$ is a mapping which leads to the new distribution

$$\pi T^{-1}(\tau) = \sum_{\theta : T(\theta) = \tau} \pi(\theta)$$

on the new states $\tau \in \mathcal{T}$ and as such represents an information processing on the state space. The remaining symbols π_D, π_U are Dirac and uniform probability distributions on Θ. The concavity argument applies also to the *alternative power functions* $\tilde{\phi}_\alpha(t) = \phi_\alpha(1-t)$ so that the same information-theoretic properties are shared by the corresponding *alternative power entropies*

$$\tilde{H}_\alpha(\pi) = \sum_{\theta\in\Theta} \tilde{\phi}_\alpha(\pi(\theta)), \quad \alpha > 0, \tag{4.37}$$

i.e.

$$\tilde{H}_\alpha(\pi) = \frac{1}{\alpha - 1}\left[n^* - 1 - \sum_{\theta\in\Theta}(1 - \pi(\theta))^\alpha\right] \quad \text{if} \quad \alpha \neq 1,$$

where $n^* = \#\{\theta \in \Theta : \pi(\theta) > 0\}$, and

$$\tilde{H}_1(\pi) = \lim_{\alpha\to 1} \tilde{H}_\alpha(\pi) = -\sum_{\theta\in\Theta}(1 - \pi(\theta))\ln(1 - \pi(\theta)).$$

Note that the alternative Shannon entropy was introduced as a measure of diversity by Zvárová (2008).

Similarly as the classical Shannon entropy, the generalized entropies $H_\alpha(\pi)$ and $\tilde{H}_\alpha(\pi)$ are measures of the information obtained by observing the state from Θ a priori distributed by π. One can thus expect that the minimal error probability $e_B(\pi)$ of identification of this state on the basis of π is intimately related to these entropies. Since the Bayes error $e_B = e_B(\mathcal{E})$ in the general experiment $\mathcal{E}$ (c.f. (2.21)) is the average minimal error probability

$$e_B(\mathcal{E}) = \int_{\mathcal{X}} e_B(\pi_x)\mathrm{d}P(x) \qquad \text{(c.f. (3.29))}, \tag{4.38}$$

it must be similarly related to the average generalized entropies $H_\alpha(\mathcal{E})$ and $\tilde{H}_\alpha(\mathcal{E})$ defined as analogous stochastic mixtures

$$H_\alpha(\mathcal{E}) = \int_{\mathcal{X}} H_\alpha(\pi_x)\mathrm{d}P(x) \quad \text{and} \quad \tilde{H}_\alpha(\mathcal{E}) = \int_{\mathcal{X}} \tilde{H}_\alpha(\pi_x)\mathrm{d}P(x). \tag{4.39}$$

In what follows we investigate this relation.

In the next theorem we evaluate the upper and lower bounds

$$\mathcal{H}_\alpha^+(e_B) = \max_{e_B(\mathcal{E})=e_B} H_\alpha(\mathcal{E}) \quad \text{and} \quad \mathcal{H}_\alpha^-(e_B) = \min_{e_B(\mathcal{E})=e_B} H_\alpha(\mathcal{E}), \tag{4.40}$$

using for $\alpha > 0$ and $n =\mid \Theta \mid$ the auxiliary constants

$$a_{\alpha,k} = \begin{cases} \frac{1-k^{1-\alpha}}{\alpha-1} & \text{if } \alpha \neq 1 \\ \lim_{\alpha\to 1} a_{\alpha,k} = \ln k & \text{if } \alpha = 1 \end{cases}, \; c_k = \frac{k-1}{k}, \quad 1 \le k \le n, \tag{4.41}$$

$$b_{\alpha,k} = \frac{a_{\alpha,k+1} - a_{\alpha,k}}{c_{k+1} - c_k}, \quad 1 \le k \le n-1 \tag{4.42}$$

and the auxiliary function

$$h(t) = -t\ln t - (1-t)\ln(1-t), \quad 0 \le t \le 1 \;\; \text{where} \;\; 0\ln 0 = 0. \tag{4.43}$$

Theorem 3. For every $\alpha > 0$ and $0 \le e_B \le c_n \equiv (|\Theta|-1)/|\Theta|$ the *power entropy upper bounds* (4.40) are given by the formulas

$$\mathcal{H}_\alpha^+(e_B) = \frac{1-(n-1)^{1-\alpha}e_B^\alpha - (1-e_B)^\alpha}{\alpha - 1} \tag{4.44}$$

if $\alpha \neq 1$ and

$$\mathcal{H}_1^+(e_B) = \lim_{\alpha\to 1} \mathcal{H}_\alpha^+(e_B) = h(e_B) + e_B\ln(n-1) \tag{4.45}$$

if $\alpha = 1$ while the *power entropy lower bounds* (4.40) are given by the formulas

$$\mathcal{H}_\alpha^-(e_B) = a_{\alpha,k} + b_{\alpha,k}(e_B - c_k) \quad \text{when} \quad c_k \le e_B \le c_{k+1}, \quad 1 \le k \le n-1 \tag{4.46}$$

if $0 < \alpha < 2$ and

$$\mathcal{H}_\alpha^-(e_B) = \frac{a_{\alpha,n}}{c_n} e_B \tag{4.47}$$

if $\alpha \ge 2$. The bounds $\mathcal{H}_\alpha^+(e_B)$ and $\mathcal{H}_\alpha^-(e_B)$ coincide only at the endpoints $c_1 = 0$ and c_n of the domain of e_B where

$$\mathcal{H}_\alpha^+(0) = \mathcal{H}_\alpha^-(0) = 0 \quad \text{and}\; \mathcal{H}_\alpha^+(c_n) = \mathcal{H}_\alpha^-(c_n) = a_{\alpha,n} > 0. \tag{4.48}$$

Proof. (I) The Bayesian errors $e(\pi)$ and e_B take on values in the interval

$$0 \le e(\pi), \; e_B \le c_n. \tag{4.49}$$

By Theorem 2 in Morales et al. (1996), for every $0 \le e \le c_n$

$$e(\pi) = e \quad \text{implies} \quad H_\alpha^-(e) \le H_\alpha(\pi) \le H_\alpha^+(e) \tag{4.50}$$

where the lower and upper bounds $H_\alpha^\pm(e)$ are attained by the entropies $H_\alpha(\pi^\pm)$ for special distributions $\pi^\pm = (\pi^\pm(\theta) : \theta \in \Theta)$. If $\alpha \neq 1$ then, by using the method proposed by Vajda and Vašek (1985), these bounds were evaluated in the mentioned Theorem 2 as follows:

$$H_\alpha^+(e) = \frac{1-(n-1)^{1-\alpha}e^\alpha - (1-e)^\alpha}{\alpha - 1} \tag{4.51}$$

and

$$H_\alpha^-(e) = \frac{1-[1-k(1-e)]^\alpha - k(1-e)^\alpha}{\alpha-1} \tag{4.52}$$

when $c_k \le e \le c_{k+1}$ and $1 \le k \le n-1$. If $\alpha = 1$ then the upper bound was evaluated as the corresponding limit

$$H_1^+(e) = \lim_{\alpha\to 1} H_\alpha^+(e) = h(e) + e\ln(n-1) \tag{4.53}$$

and the lower bound was evaluated as the limit

$$H_1^-(e) = \lim_{\alpha\to 1} H_\alpha^-(e) = -[1-k(1-e)]\ln[1-k(1-e)] - k(1-e)\ln(1-e) \tag{4.54}$$

on the intervals $c_k \le e \le c_{k+1}$ for $1 \le k \le n-1$.

(II) Consider now arbitrary parameter $\alpha > 0$, arbitrary constants $0 \le \tilde{c} < c \le c_n$ and arbitrary distributions $\pi, \tilde{\pi}$ such that $e(\pi) = c$ and $\tilde{e}(\tilde{\pi}) = \tilde{c}$. Then the linear function

$$tH_\alpha(\pi) + (1-t)H_\alpha(\tilde{\pi}) \text{ of variable } 0 \le t \le 1$$

must be bounded above by the function $\mathcal{H}_\alpha^+(tc + (1-t)\tilde{c})$ and bounded below by the function $\mathcal{H}_\alpha^-(tc+(1-t)\tilde{c})$. This implies that $\mathcal{H}_\alpha^+$ must be concave and $\mathcal{H}_\alpha^-$ convex on the interval $[\tilde{c}, c] \subseteq [0,1]$. At the same time $\mathcal{H}_\alpha^+$ must be minimal but above H_α^+ and $\mathcal{H}_\alpha^-$ must be maximal but below H_α^-. Since H_α^+ is concave itself, this implies $\mathcal{H}_\alpha^+ = H_\alpha^+$ so that (4.44) and (4.45) follow from (4.51) and (4.53). On the other hand, H_α^- given by (4.52) and (4.54) is piecewise concave in the intervals between the cutpoints c_k, $1 \le k \le n-1$. The piecewise linear function $\Phi_\alpha(t)$ of variable $t \in [0, c_n]$ connecting the points $[c_k, H_\alpha^-(c_k)] \equiv [c_k, a_k]$ for $1 \le k \le n$ is

$$\Phi_\alpha(t) = a_{\alpha,k} + b_{\alpha,k}(t-c_k) \quad \text{for} \quad c_k \le t \le c_{k+1}, \quad 1 \le k \le n-1. \tag{4.55}$$

This function is convex (concave) if the sequence

$$\frac{\Phi_\alpha(c_k)}{c_k} = \frac{a_{\alpha,k}}{c_k} = \begin{cases} \frac{k\left(1-k^{1-\alpha}\right)}{(\alpha-1)(k-1)} & \text{if } \alpha \ne 1 \\ \lim_{\alpha\to 1} a_{\alpha,k} = \frac{k}{k-1}\ln k & \text{if } \alpha = 1 \end{cases}$$

is increasing (decreasing) for $k = 2, 3, ...$Obviously, it is constant equal 1 if $\alpha = 2$, increasing if $0 < \alpha < 2$ and decreasing if $\alpha > 2$. Therefore $\mathcal{H}_\alpha^-(e_B) = \Phi_\alpha(e_B)$ if $0 < \alpha \le 2$ and $\mathcal{H}_\alpha^-(e_B)$ is linear in e_B, equal $[\Phi_\alpha(c_n) - \Phi_\alpha(0)]\, e_B/c_n \equiv a_n e_B/c_n$, if $\alpha > 2$. This proves (4.46) and (4.47). The last assertion including relations (4.48) is a consequence of what has already been proved. In Figures 1 and 2 are drawn the curves $\mathcal{H}_\alpha^\pm(e_B)$ as functions of variable e_B for $\alpha = 1/2, 1$ and $\alpha = 2, 3$. The lower bounds $\mathcal{H}_\alpha^-(e_B)$ for $\alpha \ge 2$ are linear in e_B.

Remark. Relation (4.53) is the well known Fano bound of information theory and (4.51) is its extension obtained previously in Vajda (1968) for $\alpha = 2$ and in Morales, Pardo and Vajda (1996) and other references mentioned there for remaining $\alpha > 0$.

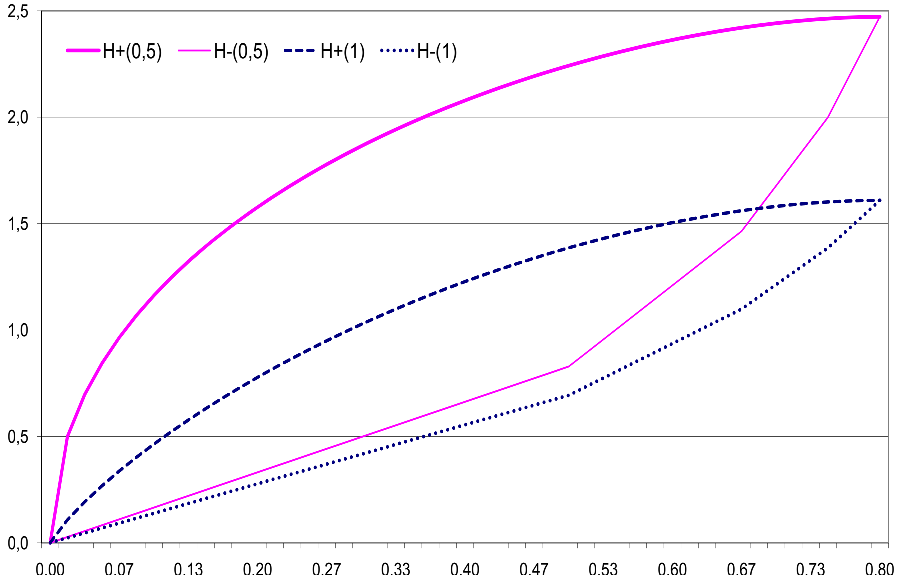

Figure 1: $\mathcal{H}_\alpha^\pm(e_B)$ as functions of variable e_B for $\alpha = 1/2, 1$.

The next theorem evaluates the upper and lower bounds

$$\tilde{\mathcal{H}}_\alpha^+(e_B) = \max_{e_B(\mathcal{E})=e_B} \tilde{H}_\alpha(\mathcal{E}) \quad \text{and} \quad \tilde{\mathcal{H}}_\alpha^-(e_B) = \min_{e_B(\mathcal{E})=e_B} \tilde{H}_\alpha(\mathcal{E}). \tag{4.56}$$

It uses the same c_k as Theorem 3 and for every $\alpha > 0$ the constants

$$\tilde{a}_{\alpha,k} = \begin{cases} \frac{k-1}{\alpha-1}\left[1-\left(\frac{k-1}{k}\right)^{\alpha-1}\right] & \text{if } \alpha \neq 1 \\ \lim_{\alpha\to1}\tilde{a}_{\alpha,k} = (k-1)\ln\frac{k}{k-1} & \text{if } \alpha = 1 \end{cases} \tag{4.57}$$

for $0\ln 0 = 0$, $1 \le k \le n$, and

$$\tilde{b}_{\alpha,k} = \frac{\tilde{a}_{\alpha,k+1} - \tilde{a}_{\alpha,k}}{c_{k+1} - c_k} \quad \text{for} \quad 1 \le k \le n-1.$$

Theorem 4. Let $\alpha > 0$ be arbitrary fixed. The *alternative power entropy upper bounds* (4.56) is for every $0 \le e_B \le c_n$ explicitly given by the formula

$$\tilde{\mathcal{H}}_\alpha^+(e_B) = \frac{1}{\alpha-1}\left[n-1-e_B^\alpha-(n-1)\left(1-\frac{e_B}{n-1}\right)^\alpha\right] \tag{4.58}$$

if $\alpha \neq 1$ and

$$\tilde{\mathcal{H}}_1^+(e_B) = \lim_{\alpha\to1}\tilde{\mathcal{H}}_\alpha^+(e_B) = -e\ln e - (n-1-e)\ln\left(\frac{n-1-e}{n-1}\right) \tag{4.59}$$

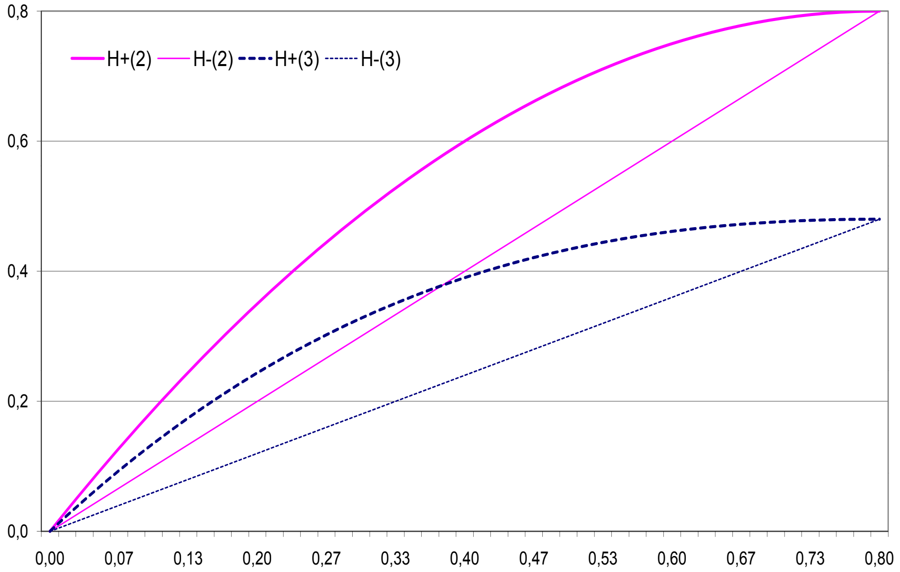

Figure 2: $\mathcal{H}^{\pm}_{\alpha}(e_B)$ as functions of variable e_B for $\alpha = 2, 3$.

if $\alpha = 1$ while the *alternative power entropy lower bounds* (4.56) are given by the formulas

$$\tilde{\mathcal{H}}^{-}_{\alpha}(e_B) = \tilde{a}_{\alpha,k} + \tilde{b}_{\alpha,k}(e_B - c_k) \quad \text{when} \quad c_k < e_B < c_{k+1}, \quad 1 \leq k \leq n-1 \tag{4.60}$$

if $\alpha > 2$ and

$$\tilde{\mathcal{H}}^{-}_{\alpha}(e_B) = \frac{\tilde{a}_{\alpha,n}}{c_n} e_B \tag{4.61}$$

if $0 < \alpha \leq 2$. The bounds $\tilde{\mathcal{H}}^{+}_{\alpha}(e_B)$ and $\tilde{\mathcal{H}}^{-}_{\alpha}(e_B)$ coincide only at the endpoints $c_1 = 0$ and c_n of the domain of e_B where

$$\mathcal{H}^{+}_{\alpha}(0) = \mathcal{H}^{-}_{\alpha}(0) = 0 \quad \text{and } \mathcal{H}^{+}_{\alpha}(c_n) = \mathcal{H}^{-}_{\alpha}(c_n) = \tilde{a}_{\alpha,n} > 0. \tag{4.62}$$

Proof. (I) As before, the Bayes errors $e(\pi)$ and e_B take on values in the interval

$$0 \leq e(\pi),\ e_B \leq c_n.$$

By Theorem 1 in Vajda and Vašek (1985), for every $0 \leq e \leq c_n$

$$e(\pi) = e \quad \text{implies} \quad \tilde{H}^{-}_{\alpha}(e) \leq \tilde{H}_{\alpha}(\pi) \leq \tilde{H}^{+}_{\alpha}(e) \tag{4.63}$$

where the lower and upper bounds $H^{\pm}_{\alpha}(e)$ are attained by the entropies $H_{\alpha}(\pi^{\pm})$ for the special distributions

$$\pi^{+} = \left(1 - e, \frac{e}{n-1}, \frac{e}{n-1} ..., \frac{e}{n-1}\right)$$

and

$$\pi^{-} = (1 - e, 1 - e, , ..., 1 - e, 1 - k(1 - e), 0, 0, ..., 0)$$

provided $c_k \le e \le c_{k+1}$ for $1 \le k \le n-1$. Hence for $\alpha \ne 1$

$$\tilde{H}^+_\alpha(e) = \tilde{H}_\alpha(\pi^+) = \frac{1}{\alpha-1}\left[n-1-e^\alpha-(n-1)\left(1-\frac{e}{n-1}\right)^\alpha\right] \tag{4.64}$$

and

$$\tilde{H}^-_\alpha(e) = \tilde{H}_\alpha(\pi^-) = \frac{k-ke^\alpha-k^\alpha(1-e)^\alpha}{\alpha-1} \tag{4.65}$$

when $c_k \le e \le c_{k+1}$ and $1 \le k \le n-1$. For $\alpha = 1$ we get

$$\tilde{H}^+_1(e) = \tilde{H}_1(\pi^+) = \lim_{\alpha\to1}\tilde{H}^+_\alpha(e) = -e\ln e-(n-1-e)\ln\left(\frac{n-1-e}{n-1}\right) \tag{4.66}$$

and

$$\tilde{H}^-_1(e) = \tilde{H}^-_1(\pi^-) = \lim_{\alpha\to1}\tilde{H}^-_\alpha(e) = -ke-k(1-e)\ln\left[k(1-e)\right] \tag{4.67}$$

on the intervals $c_k \le e \le c_{k+1}$ for $1 \le k \le n-1$.

(II) Consider now arbitrary parameter $\alpha > 0$, arbitrary constants $0 \le \tilde{c} < c \le c_n$ and arbitrary distributions $\pi, \tilde{\pi}$ such that $e(\pi) = c$ and $\tilde{e}(\tilde{\pi}) = \tilde{c}$. Then the linear function

$$t\tilde{H}_\alpha(\pi)+(1-t)\tilde{H}_\alpha(\tilde{\pi}) \text{ of variable } 0 \le t \le 1$$

must be bounded above by the function $\tilde{\mathcal{H}}^+_\alpha(tc+(1-t)\tilde{c})$ and bounded below by the function $\tilde{\mathcal{H}}^-_\alpha((tc+(1-t)\tilde{c})$. Similarly as in the previous proof, this implies that $\tilde{\mathcal{H}}^+_\alpha$ must be concave and $\tilde{\mathcal{H}}^-_\alpha$ convex on the interval $[\tilde{c}, c] \subseteq [0,1]$. At the same time $\tilde{\mathcal{H}}^+_\alpha$ must be minimal but above $\tilde{H}^+_\alpha$ and $\tilde{\mathcal{H}}^-_\alpha$ must be maximal but below $\tilde{H}^-_\alpha$. Since $\tilde{H}^+_\alpha$ is concave itself, this implies $\tilde{\mathcal{H}}^+_\alpha = \tilde{H}^+_\alpha$ so that (4.58) and (4.59) follow from (4.64) and (4.66). On the other hand, $\tilde{H}^-_\alpha$ given by (4.65) and (4.67) is piecewise concave in the intervals between the cutpoints c_k, $1 \le k \le n-1$. The piecewise linear function $\tilde{\Phi}_\alpha(t)$ of variable $t \in [0, c_n]$ connecting the points $[c_k, \tilde{H}^-_\alpha(c_k)] \equiv [c_k, \tilde{a}_k]$ for $1 \le k \le n$ is

$$\tilde{\Phi}_\alpha(t) = \tilde{a}_{\alpha,k}+\tilde{b}_{\alpha,k}(t-c_k) \quad \text{for} \quad c_k \le t \le c_{k+1}, \quad 1 \le k \le n-1$$

This function is convex for (concave) if the sequence

$$\frac{\tilde{\Phi}_\alpha(c_k)}{c_k} = \frac{\tilde{a}_{\alpha,k}}{c_k} = \begin{cases} \frac{k}{\alpha-1}\left[1-\left(\frac{k-1}{k}\right)^{\alpha-1}\right] & \text{if } \alpha \ne 1 \\ \lim_{\alpha\to1}\tilde{a}_{\alpha,k} = -k\ln\frac{k-1}{k} & \text{if } \alpha = 1 \end{cases}$$

is increasing (decreasing) for $k = 2, 3, ...$Obviously, it is constant equal 1 if $\alpha = 2$, decreasing if $0 < \alpha < 2$ and increasing if $\alpha > 2$. Therefore $\mathcal{H}^-_\alpha(e_B) = \Phi_\alpha(e_B)$ if $\alpha > 2$ and $\mathcal{H}^-_\alpha(e_B)$ is linear in e_B equal $[\Phi_\alpha(c_n)-\Phi_\alpha(0)]\, e_B/c_n \equiv a_n e_B/c_n$ if $0 < \alpha \le 2$. This proves (4.60)and (4.61). The last assertion including relations (4.62) is a consequence of what was already proved above.

In Figures 3 and 4 are drawn the curves $\tilde{\mathcal{H}}^\pm_\alpha(e_B)$ as functions of variable e_B for $\alpha = 1/2, 1$ and $\alpha = 2, 3$.

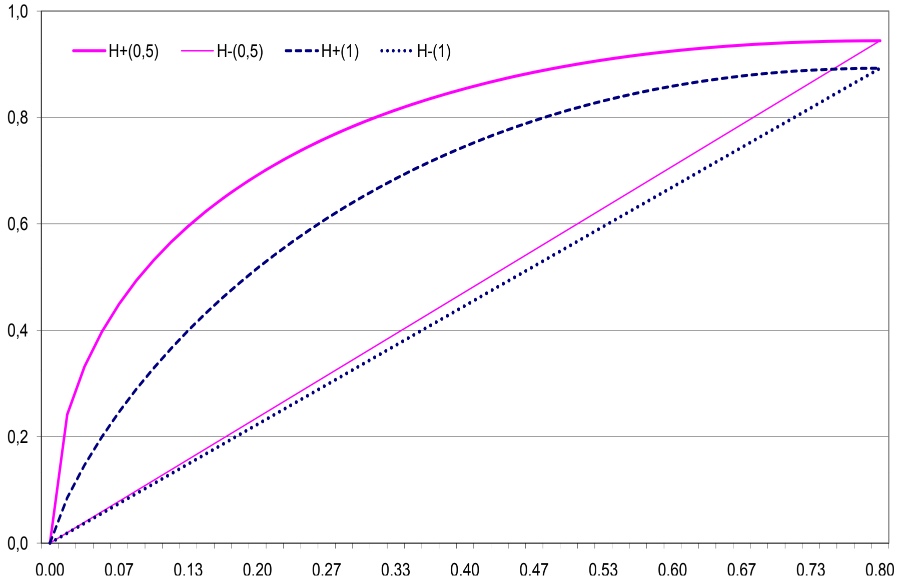

Figure 3: $\tilde{\mathcal{H}}^{\pm}_{\alpha}(e_B)$ as functions of variable e_B for $\alpha = 1/2, 1$.

The last step of our research was evaluation of the integrals

$$\int_0^{c_n} \mathcal{H}^+_\alpha(e_B)\,\mathrm{d}e_B = \begin{cases} \frac{1}{\alpha-1}\left[\frac{n-1}{n} - \frac{n^\alpha+n-2}{(\alpha+1)n^\alpha}\right] & \text{if } \ \alpha \neq 1 \\ \frac{1}{2n}\left[n-1+(n-2)\ln n\right] & \text{if } \ \alpha = 1 \end{cases} \tag{4.68}$$

and

$$\int_0^{c_n} \mathcal{H}^-_\alpha(e_B)\,\mathrm{d}e_B = \begin{cases} \frac{1}{2(\alpha-1)}\sum_{k=1}^{n-1}\frac{2-k^{1-\alpha}-(k+1)^{1-\alpha}}{k(k+1)} & \text{if } \ 0<\alpha<2, \alpha\neq 1 \\ \frac{1}{2}\sum_{k=1}^{n-1}\frac{\ln[k(k+1)]}{k(k+1)} & \text{if } \ \alpha = 1 \\ \frac{(n-1)(1-n^{1-\alpha})}{2(\alpha-1)n} & \text{if } \ \alpha \geq 2. \end{cases} \tag{4.69}$$

for the Bayes error bounds $\mathcal{H}^{\pm}_\alpha(e_B)$. Average differences

$$\tau_{\alpha,n} = \frac{1}{c_n}\int_0^{c_n}\left(\mathcal{H}^+_\alpha(e_B) - \mathcal{H}^-_\alpha(e_B)\right)\,\mathrm{d}e_B. \tag{4.70}$$

between these bounds represent *average inaccuracies* of the power (in particular Shannon) conditional entropies as information criteria of Bayes errors. Concrete numerical values of these inaccuracy measures as functions of $\alpha > 0$ are given in Table 1 at the end of the paper.

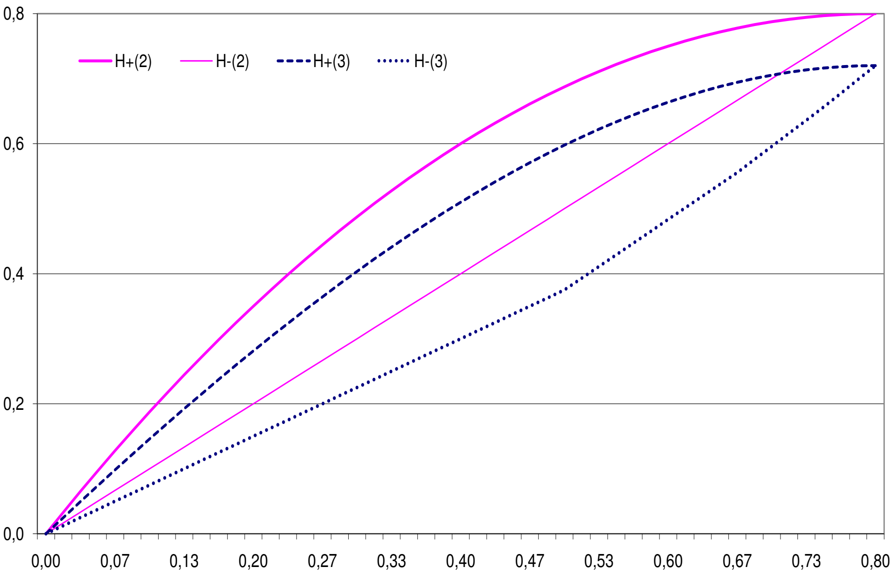

Figure 4: $\tilde{\mathcal{H}}_\alpha^\pm(e_B)$ as functions of variable e_B for $\alpha = 2, 3$.

Similar steps and ideas were applied also to the alternative bounds $\tilde{\mathcal{H}}_\alpha^\pm(e_B)$, namely we obtained the integrals

$$\int_0^{c_n} \tilde{\mathcal{H}}_\alpha^+(e_B)\,\mathrm{d}e_B = \begin{cases} \frac{1}{\alpha-1}\left[\frac{(n-1)^2}{n} - \frac{(n-1)^2}{\alpha+1} + \frac{n(n-2)}{\alpha+1}\left(\frac{n-1}{n}\right)^{\alpha+1}\right] & \text{if } \alpha \neq 1 \\ \frac{(n-1)^2}{2n}\left[1 + (n-2)\ln\frac{n-1}{n}\right] & \text{if } \alpha = 1 \end{cases} \tag{4.71}$$

and

$$\int_0^{c_n} \tilde{\mathcal{H}}_\alpha^-(e_B)\,\mathrm{d}e_B = \begin{cases} \frac{(n-1)^2}{2n(\alpha-1)}\left[1 - \left(\frac{n-1}{n}\right)^{\alpha-1}\right] & \begin{matrix}\text{if } 0 < \alpha \leq 2, \\ \alpha \neq 1\end{matrix} \\ \frac{(n-1)^2}{2n}\ln\frac{n}{n-1} & \text{if } \alpha = 1 \\ \frac{1}{2(\alpha-1)}\sum_{k=1}^{n-1}\frac{2k-1-(k-1)\left(\frac{k-1}{k}\right)^{\alpha-1}-k\left(\frac{k}{k+1}\right)^{\alpha-1}}{k(k+1)} & \text{if } \alpha > 2. \end{cases} \tag{4.72}$$

leading to the inaccuracy measures

$$\tilde{\tau}_{\alpha,n} = \frac{1}{c_n}\int_0^{c_n}\left(\tilde{\mathcal{H}}_\alpha^+(e_B) - \tilde{\mathcal{H}}_\alpha^-(e_B)\right)\,\mathrm{d}e_B. \tag{4.73}$$

for conditional alternative power entropies (in particular, alternative Shannon entropies). Concrete numerical values of $\tilde{\tau}_{\alpha,n}$ as functions of $\alpha > 0$ are given in Table 2 at the end of the paper.

Criteria for Bayes and sub-Bayes risk are now easily obtained by plugging into the bounds and inaccuracy measures of the present section the results of sections 2 and 3. This and application of the accuaracy measures for selection of the most accurate information criteria will be the last step of this research.

5 Acknowledgements

The research for this paper was supported by grants GAČR 102/07/1131 and MTM2006-05693.

6 References

M. Ben Bassat (1978). f-entropies, probability of error, and feature selection. *Information and Control.* **39**, 227-242.

M. Ben Bassat and J. Raviv (1978). Rényi's entropy and probability of error. *IEEE Transactions on Information Theory*, **24,** 324-331.

Berger J. O. (1986). *Statistical Decision Theory and Bayesian Analysis.* 2-nd Ed., Springer, Berlin.

T. M.Cover and P. E. Hart (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13,** 21-27.

P. Devijver and J. Kittler (1982). *Pattern Recognition. A Statistical Approach.* Prentice Hall, Englewood Cliffs, New Jersey.

L. Devroye, L. Györfi and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, Berlin.

D. K. Faddeev (1957). Zum Begriff der Entropie einer endlichen Wahrscheinlichkeitsschemas. *Arbeits zur informationstheorie,* vol. I, Deutscher Verlag der Wissenschaften, Berlin.

M. Feder and N. Merhav (1994). Relations between entropy and error probability. *IEEE Transactions on Information Theory*, **40,** 259-266.

J. Havrda and F. Charvát (1967). Concept of structural a-entropy. *Kybernetika* **3**, 30-35.

L. Kanal (1974). Patterns in pattern recognittion. *IEEE Transactions on Information Theory*, **20,** 697-707.

V. A. Kovalevskij (1965). The problem of character recognition from the point of view of mathematical statistics. In *Reading Automata and Pattern Recognition* (in Russian), Naukova Dumka, Kyjev. English translation in: *Character Readers and Pattern Recognition* , 3-30. Spartan Books, New York (1968).

D. Morales, L. Pardo and I. Vajda (1996). Uncertainty of discrete stochastic systems: general theory and statistical inference. *IEEE Transactions on System, Man and Cybernetics, Part A*, **26**, 1-17.

A. Rényi (1961). On measures of entropy and information. In *Proceedings of 4-th Berkel;ey Symp. on Probab. Statist.* Univ. of California Press, Berkeley, California.

N. P. Salichov (1974). Confirmation of a hypothesis of I. Vajda. *Information Transmission Problems,* **10**, 114-115.

D. L. Tebbe and S. J. Dwyer III (1968). Uncertainty and probability of error. *IEEE Transactions on Information Theory*, **14,** 516-518.

G. T. Toussaint (1977). A generalization of Shannon's equivocation and the Fano bound. *IEEE Transactions on System, Man and Cybernetics*, **7**, 300-302.

I. Vajda (1968). Bounds on the minimal error probability and checking a finite or countable number of hypotheses. *Information Transmisssion Problems,* **4**, 9-17.

I. Vajda (1969). A contribution to informational analysis of patterns. In *Methodologies of Pattern Recognition* (Ed. M. S. Watanabe). Academic Press, New York.

I. Vajda and K. Vašek (1985). Majorization, concave entropies and comparison of experiments. *Problems of Control and Information Theory* **14**, 105-115.

Annex of Tables

n	0.125	0.25	0.5	1	1.5	2	3	4
2	0.4123	0.3455	0.2525	0.1534	0.1071	0.3333	0.2500	0.2000
3	1.0019	0.8146	0.5575	0.2908	0.1705	0.2569	0.1649	0.1227
4	1.6083	1.2764	0.8322	0.3939	0.2099	0.2037	0.1306	0.1011
5	2.2213	1.7288	1.0849	0.4773	0.2376	0.1818	0.1203	0.0958
6	2.8364	2.1717	1.3202	0.5474	0.2587	0.1716	0.1168	0.0944
7	3.4514	2.6055	1.5415	0.6082	0.2754	0.1663	0.1157	0.0941
8	4.0649	3.0309	1.7511	0.6619	0.2891	0.1635	0.1156	0.0942
9	4.6763	3.4484	1.9507	0.7100	0.3006	0.1619	0.1158	0.0945
10	5.2853	3.8588	2.1416	0.7535	0.3105	0.1610	0.1161	0.0948
20	11.2287	7.6589	3.7372	1.0523	0.3664	0.1610	0.1195	0.0970
30	16.9402	11.0848	5.0025	1.2353	0.3926	0.1623	0.1211	0.0979
40	22.4755	14.2697	6.0850	1.3680	0.4088	0.1631	0.1220	0.0984
50	27.8726	17.2803	7.0469	1.4723	0.4200	0.1637	0.1226	0.0987
100	53.4660	30.7249	10.8646	1.8025	0.4487	0.1651	0.1238	0.0993
200	100.7473	53.5218	16.3214	2.1392	0.4699	0.1659	0.1244	0.0997
300	145.1650	73.5789	20.5287	2.3382	0.4794	0.1661	0.1246	0.0998
400	187.7885	92.0259	24.0828	2.4800	0.4852	0.1663	0.1247	0.0998
500	229.1154	109.3564	27.2176	2.5903	0.4892	0.1663	0.1248	0.0999
600	269.4330	125.8440	30.0537	2.6806	0.4922	0.1664	0.1248	0.0999
700	308.9277	141.6610	32.6630	2.7571	0.4945	0.1664	0.1248	0.0999
800	347.7298	156.9250	35.0927	2.8233	0.4963	0.1665	0.1248	0.0999
900	385.9354	171.7210	37.3753	2.8818	0.4979	0.1665	0.1249	0.0999
1000	423.6182	186.1130	39.5347	2.9342	0.4992	0.1665	0.1249	0.0999
1500	605.8886	253.4550	48.9962	3.1359	0.5037	0.1666	0.1249	0.1000
2000	780.6168	315.3187	56.9761	3.2792	0.5064	0.1666	0.1249	0.1000
2500	949.9337	373.4003	64.0082	3.3904	0.5082	0.1666	0.1250	0.1000
3000	1115.0557	428.6341	70.3668	3.4814	0.5095	0.1666	0.1250	0.1000
3500	1276.7668	481.6060	76.2147	3.5583	0.5106	0.1666	0.1250	0.1000
4000	1435.6130	532.7152	81.6582	3.6249	0.5115	0.1666	0.1250	0.1000
4500	1591.9953	582.2495	86.7712	3.6837	0.5122	0.1666	0.1250	0.1000
5000	1746.2203	630.4252	91.6074	3.7363	0.5128	0.1666	0.1250	0.1000
5500	1898.5295	677.4104	96.2074	3.7839	0.5133	0.1666	0.1250	0.1000
6000	2049.1181	723.3393	100.6028	3.8273	0.5137	0.1666	0.1250	0.1000
6500	2198.1467	768.3208	104.8187	3.8673	0.5141	0.1666	0.1250	0.1000
7000	2345.7497	812.4453	108.8755	3.9043	0.5145	0.1666	0.1250	0.1000
7500	2492.0411	855.7887	112.7898	3.9388	0.5148	0.1666	0.1250	0.1000
8000	2637.1191	898.4155	116.5757	3.9710	0.5151	0.1666	0.1250	0.1000
8500	2781.0685	940.3815	120.2452	4.0013	0.5154	0.1666	0.1250	0.1000
9000	2923.9639	981.7349	123.8082	4.0299	0.5156	0.1666	0.1250	0.1000
9500	3065.8708	1022.5178	127.2736	4.0569	0.5158	0.1666	0.1250	0.1000
10000	3206.8477	1062.7677	130.6490	4.0825	0.5160	0.1667	0.1250	0.1000

Table 1. Average inaccuracies $\tau_{\alpha,n}$ for selected α and n.

n	0.125	0.25	0.5	1	1.5	2	3	4
2	0.8889	0.8000	0.6667	0.5000	0.4000	0.3333	0.0625	0.0542
3	0.5551	0.5009	0.4234	0.3346	0.2867	0.2569	0.1042	0.1017
4	0.4893	0.4376	0.3628	0.2765	0.2307	0.2037	0.1262	0.1288
5	0.4671	0.4166	0.3430	0.2566	0.2097	0.1818	0.1398	0.1463
6	0.4572	0.4076	0.3349	0.2483	0.2005	0.1716	0.1490	0.1584
7	0.4521	0.4032	0.3311	0.2446	0.1961	0.1663	0.1557	0.1673
8	0.4491	0.4007	0.3292	0.2428	0.1939	0.1635	0.1608	0.1742
9	0.4473	0.3994	0.3283	0.2420	0.1928	0.1619	0.1647	0.1796
10	0.4462	0.3985	0.3278	0.2417	0.1923	0.1610	0.1679	0.1840
20	0.4436	0.3975	0.3287	0.2436	0.1935	0.1610	0.1825	0.2044
30	0.4435	0.3980	0.3298	0.2453	0.1952	0.1623	0.1875	0.2114
40	0.4436	0.3984	0.3306	0.2463	0.1962	0.1631	0.1899	0.2150
50	0.4438	0.3986	0.3311	0.2470	0.1969	0.1637	0.1914	0.2172
100	0.4440	0.3993	0.3321	0.2484	0.1983	0.1651	0.1944	0.2215
200	0.4442	0.3996	0.3327	0.2492	0.1991	0.1659	0.1960	0.2237
300	0.4443	0.3997	0.3329	0.2495	0.1994	0.1661	0.1965	0.2244
400	0.4443	0.3998	0.3330	0.2496	0.1996	0.1663	0.1967	0.2248
500	0.4444	0.3998	0.3331	0.2497	0.1997	0.1663	0.1969	0.2250
600	0.4444	0.3999	0.3331	0.2497	0.1997	0.1664	0.1970	0.2252
700	0.4444	0.3999	0.3332	0.2498	0.1998	0.1664	0.1970	0.2253
800	0.4444	0.3999	0.3332	0.2498	0.1998	0.1665	0.1971	0.2253
900	0.4444	0.3999	0.3332	0.2498	0.1998	0.1665	0.1971	0.2254
1000	0.4444	0.3999	0.3332	0.2498	0.1998	0.1665	0.1972	0.2255
1500	0.4444	0.3999	0.3333	0.2499	0.1999	0.1666	0.1973	0.2256
2000	0.4444	0.4000	0.3333	0.2499	0.1999	0.1666	0.1973	0.2257
2500	0.4444	0.4000	0.3333	0.2499	0.1999	0.1666	0.1973	0.2257
3000	0.4444	0.4000	0.3333	0.2499	0.1999	0.1666	0.1974	0.2257
3500	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
4000	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
4500	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
5000	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
5500	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
6000	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
6500	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
7000	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
7500	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
8000	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
8500	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
9000	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
9500	0.4444	0.4000	0.3333	0.2500	0.2000	0.1666	0.1974	0.2258
10000	0.4444	0.4000	0.3333	0.2500	0.2000	0.1667	0.1974	0.2258

Table 2. Average inaccuracies $\tilde{\tau}_{\alpha,n}$ for selected α and n.