KYBERNETIKA — VOLUME 30 (1994), NUMBER 4, PAGES 445 - 460

DISCRETIZATION PROBLEMS ON GENERALIZED ENTROPIES AND R-DIVERGENCES[1]

L. PARDO, D. MORALES, K. FERENTINOS AND K. ZOGRAFOS

In many practical applications, data about an unknown continuous distribution arise in a grouped form. For these cases, estimation of population entropies and divergences must be done by means of their sample discretized estimates. In this paper, the problem of loss of information due to the discretization of the data is studied for (h,ϕ)-*entropies* and R_ϕ^h-*divergences*. Quadratic convergence theorems are given and asymptotic distributions are obtained.

1. INTRODUCTION

Let $(\mathcal{X}, \beta_\mathcal{X}, P_\theta)_{\theta\in\Theta}$ be a statistical space, where Θ is an open subset of $\mathbb{R}^M$. We shall assume that there exists a generalized probability density $f_\theta(x)$ for the distribution P_θ with respect to a σ-finite measure μ. In this context Csiszár [4], Burbea and Rao [2] considered the ϕ-entropy associated with $f_\theta(x)$ in the following way

$$H_\phi(f_\theta) = \int_\mathcal{X} \phi(f_\theta(x))\,\mathrm{d}\mu(x) \quad (\phi \text{ concave}). \tag{1}$$

Special choices of ϕ, such as $\phi_1(t) = -t\log t$, $\phi_2(t) = t - t^2$, $\phi_3(t) = t - t^3$, $\phi_4(t) = t-2t^2+2t^3-t^4$, $\phi_5(t) = -\log\int_0^\infty x^t e^{-x}\mathrm{d}x$, $\phi_6(t) = (1-\alpha)^{-1}(t^\alpha-1)$, $\alpha \neq 1$, $\alpha > 0$, $\phi_7(t) = (1+\lambda^{-1})\log(1+\lambda) - \lambda^{-1}(1+\lambda t)\log(1+\lambda t), \lambda > 0$, etc., give rise to Shannon's entropy, quadratic entropy, cubic entropy, genetic entropy, gamma entropy, entropy of degree α, hypoentropy, etc. But in the literature of Information Theory there exist other information measures, for instance, Rényi's entropy, Arimoto's entropy, Sharma and Mittal's entropies, etc., which cannot be obtained from (1) by specially choosing ϕ. For this reason Salicrú, Menéndez, Pardo and Morales [7] proposed the (h,ϕ)-entropy given by

$$H_\phi^h(f_\theta) = h\,(H_\phi(f_\theta)) \tag{2}$$

where either $\phi : [0,\infty) \to \mathbb{R}$ is concave and $h : (-\infty,\infty] \to (-\infty,\infty]$ increasing and concave or $\phi : [0,\infty) \to \mathbb{R}$ is convex and $h : (-\infty,\infty] \to (-\infty,\infty]$ decreasing

[1]This work is the result of a joint research effort under project No. STV-91-E-3025 of the European Economic Community ERASMUS program.

and concave. Note also that in the remaining cases, i.e., h increasing and convex and ϕ convex or h decreasing and convex and ϕ concave, $H_\phi^h(f_\theta)$ plays the role of a certainty function. In what follows, we assume that $H_\phi^h(f_\theta)$ is an entropy function. In the particular but very important case where the family $\{P_\theta : \theta \in \Theta\}$ is discrete, the entropies $H_\phi^h(f_\theta)$ defined in this manner have been considered by many authors; e. g., Vajda and Vašek [9], where arbitrary Schur-concave entropies $H(f_\theta)$ have been studied, and other references there in. In Salicrú et al. [7], the asymptotic distributions of estimates of (h,ϕ)-entropies under simple and stratified random sampling from multinomial populations were obtained.

Based on the following concavity property of the (h,ϕ)-entropy

$$H_\phi^h\left(\frac{f_{\theta_1}+f_{\theta_2}}{2}\right) \geq \frac{H_\phi^h(f_{\theta_1})+H_\phi^h(f_{\theta_2})}{2},$$

Morales, Pardo, Salicrú and Menéndez [6] defined the R_ϕ^h-divergence between the generalized probability densities f_{θ_1} and f_{θ_2} as follows

$$\begin{aligned} R_\phi^h(f_{\theta_1}, f_{\theta_2}) &= h\left(\int_{\mathcal{X}} \phi\left(\frac{f_{\theta_1}(x)+f_{\theta_2}(x)}{2}\right) \mathrm{d}\mu(x)\right) - \\ &\quad -\frac{1}{2}\left\{h\left(\int_{\mathcal{X}} \phi(f_{\theta_1}(x))\mathrm{d}\mu(x)\right) + h\left(\int_{\mathcal{X}} \phi(f_{\theta_2}(x))\,\mathrm{d}\mu(x)\right)\right\}, \end{aligned}$$

where $\theta_1 = (\theta_{11}, \ldots, \theta_{1M})$ and $\theta_2 = (\theta_{21}, \ldots, \theta_{2M})$. When $h(x) = x$, we have the J-divergence given by Burbea and Rao [2] and if $h(x) = (1-s)^{-1}x^{\frac{s-1}{r-1}}$ and $\phi(x) = x^r$, we have the R-divergence defined by Taneja [8].

In practice the values of a continuous random variable X cannot be measured exactly. Also, frequently continuous data are only available in a grouped form. This means that, in the case of univariate continuous data, the sample space is partitioned into disjoint intervals, each yielding a discrete value for X. More precisely, all values of X such that $k\varepsilon - \frac{\varepsilon}{2} < X \leq k\varepsilon + \frac{\varepsilon}{2}$ are coded $k\varepsilon$, $k = 0, \pm 1, \pm 2, \ldots$, where $\varepsilon > 0$ is the quantum of measurement. If $\mathcal{X}$ is an open subset of $\mathbb{R}$, μ is the Lebesgue measure on $\mathcal{X}$ and $\beta_{\mathcal{X}}$ is the corresponding Borel σ-field on $\mathcal{X}$, then the true distribution of the discretized random variable is

$$p_k(\varepsilon, \theta) = \int_{k\varepsilon-\frac{1}{2}\varepsilon}^{k\varepsilon+\frac{1}{2}\varepsilon} f_\theta(x)\mathrm{d}x, \qquad k = 0, \pm 1, \pm 2, \ldots$$

Ghurye and Johnson [5] have proved that under certain regularity conditions the discretized Kullback-Leibler divergence

$$I_\varepsilon^{\mathrm{KL}}(f_{\theta_1}, f_{\theta_2}) = \sum_k p_k(\varepsilon, \theta_1) \log \frac{p_k(\varepsilon, \theta_1)}{p_k(\varepsilon, \theta_2)}$$

differ by $O(\varepsilon^2)$ from (converges quadratically to) the theoretical Kullback–Leibler divergence

$$I^{\mathrm{KL}}(f_{\theta_1}, f_{\theta_2}) = \int_{\mathcal{X}} f_{\theta_1}(x) \log \frac{f_{\theta_1}(x)}{f_{\theta_2}(x)}\,\mathrm{d}x.$$

Through this result they were able to estimate the loss of information due to the discretization of the data. Zografos, Ferentinos and Papaioannou [10] have established that the discretized versions of the Csiszár and Rényi divergence measures as well as the discretized version of the Fisher information matrix converge under certain regularity conditions quadratically to their corresponding theoretical values. Furthermore the same result holds for the Vajda, Kagan and Matusita divergences as well as for the affinity between two distributions.

It is well known that (h,ϕ)-entropies of absolutely continuous distributions cannot be approximated by the (h,ϕ)-entropies of the corresponding discrete distributions; i.e.,

$$ {}_{\varepsilon}H_{\phi}^{h}(f_\theta) = h\left(\sum_{k}\phi(p_k(\varepsilon,\theta))\right)$$

does not converge to $H_{\phi}^{h}(f_\theta)$ as $\varepsilon \to 0$. Furthermore, there are many classical examples where $H_{\phi}^{h}(f_\theta) < \infty$ and ${}_{\varepsilon}H_{\phi}^{h}(f_\theta) \to \infty$ as $\varepsilon \to 0$; e.g., Shannon differential entropy $H_{\phi}(f_\theta)$ for $\phi(x) = -x\log x$, cf. Chap. 9 in Cover and Thomas [3]. Due to these problems, ${}_{\varepsilon}H_{\phi}^{h}(f_\theta)$ is not a good candidate to be used in measuring the discretized information. Thus for these mentioned cases, some questions arise when the observed values of the random variable are actually discrete. How do we measure the data information? How do we measure the amount of information lost by discretization of the data? How do we estimate theoretical (h,ϕ)-entropies? This paper is motived by the desire to give an answer to these questions on the basis of the (h,ϕ)-entropy measures.

From now we assume that $\mathcal{X} \subset \mathbb{R}$ is open and all the positive functions $f_\theta(x)$, $\theta \in \Theta$, are continuous on $\mathcal{X}$. In this case it suffices to consider $\phi : (0,\infty) \to \mathbb{R}$ in our basic definitions (1) and (2). Let $\mathbb{Z}_\varepsilon = \{k \in \mathbb{Z}/k\varepsilon \in \mathcal{X}\}$, where $\mathbb{Z}$ is the set of integer numbers and $\mathcal{X} = S(f_\theta) = \{x \in \mathbb{R}/f_\theta(x) > 0\}$. We propose to estimate $H_{\phi}^{h}(f_\theta)$ by the following expression

$$H_{(h,\phi)}^{\varepsilon}(f_\theta) = h\left(\varepsilon\sum_{k\in\mathbb{Z}_\varepsilon}\phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right)\right), \tag{3}$$

which is motived by the fact that

$$H_{\phi}^{\varepsilon}(f_\theta) = \varepsilon\sum_{k\in\mathbb{Z}_\varepsilon}\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right)$$

is an approximate Riemann sum associated with the Riemann integral

$$\int_{\mathcal{X}}\phi(f(x))\,\mathrm{d}x.$$

This discretization is based on the one proposed by Csiszár [4] for ϕ-entropies. Csiszár proved that under some general conditions $H_\phi(f_\theta)$ can be approximated by its corresponding discretization. He proved that for every $\alpha > H_\phi(f_\theta)$ there exists $\varepsilon > 0$ such that $H_\phi(f_\theta) \le H_{\phi}^{\varepsilon}(f_\theta) < \alpha$.

Remark 1.1. Ghurye and Johnson [5] and Zografos, Ferentinos and Papaioannou [10] measured the discretized information by means of the divergence between the discretized distributions. For the measures which they studied, this is equal to approximate Riemann sums associated with the corresponding divergences between the absolutely continuous distributions; e. g., for the Kullback-Leibler information, $I^{\mathrm{KL}}(f_{\theta_1}, f_{\theta_2})$, the divergence between the discretized distributions is $I_\varepsilon^{\mathrm{KL}}(f_{\theta_1}, f_{\theta_2})$. However, as $f_{\theta_1}(k\varepsilon) \cong \frac{p_k(\varepsilon,\theta)}{\varepsilon}$ and $f_{\theta_2}(k\varepsilon) \cong \frac{q_k(\varepsilon,\theta)}{\varepsilon}$, the corresponding Riemann sum associated to $I^{\mathrm{KL}}(f_{\theta_1}, f_{\theta_2})$ is

$$I^{\mathrm{KL},\varepsilon}(f_{\theta_1}, f_{\theta_2}) = \varepsilon \sum_{k\in\mathbb{Z}_\varepsilon} f_{\theta_1}(k\varepsilon) \log \frac{f_{\theta_1}(k\varepsilon)}{f_{\theta_2}(k\varepsilon)},$$

which is approximately equal to $I_\varepsilon^{\mathrm{KL}}(f_{\theta_1}, f_{\theta_2})$. This is not the case when we deal with ϕ-entropies or (h,ϕ)-entropies.

Discrete approximations to the R_ϕ^h-divergence are defined as follows

$$R^\varepsilon_{(h,\phi)}(f_{\theta_1}, f_{\theta_2}) = H^\varepsilon_{(h,\phi)}\left(\frac{f_{\theta_1}+f_{\theta_2}}{2}\right) - \frac{1}{2}H^\varepsilon_{(h,\phi)}(f_{\theta_1}) - \frac{1}{2}H^\varepsilon_{(h,\phi)}(f_{\theta_2}). \tag{4}$$

In this paper we deal with a new problem; i. e., we examine the rate of convergence of $H^\varepsilon_{(h,\phi)}(f_\theta)$ and $R^\varepsilon_{(h,\phi)}(f_{\theta_1}, f_{\theta 2})$. We establish that these discretized versions converge under certain regularity conditions quadratically to their corresponding theoretical values; i. e., to $H^h_\phi(f_\theta)$ and $R^h_\phi(f_{\theta_1}, f_{\theta_2})$ respectively. Finally by using asymptotic distributions, the problem of estimating theoretical entropies and divergences through discretized data is also studied.

2. QUADRATIC CONVERGENCE OF H_ϕ^h-ENTROPIES AND R_ϕ^h-DIVERGENCES

In this section we are going to establish that under suitable conditions $H^\varepsilon_{(h,\phi)}(f_\theta)$ and $R^\varepsilon_{(h,\phi)}(f_{\theta_1}, f_{\theta_2})$ differ by $O(\varepsilon^2)$ from $H^h_\phi(f_\theta)$ and $R^h_\phi(f_{\theta_1}, f_{\theta_2})$ respectively as $\varepsilon \to 0$; i. e.,

$$\frac{H^h_\phi(f_\theta) - H^\varepsilon_{(h,\phi)}(f_\theta)}{\varepsilon^2} \quad \text{and} \quad \frac{R^h_\phi(f_{\theta_1}, f_{\theta_2}) - R^\varepsilon_{(h,\phi)}(f_{\theta_1}, f_{\theta_2})}{\varepsilon^2}$$

have a finite limite as $\varepsilon \to 0$. Theorem 2.1 establishes the quadratic convergence to the (h,ϕ)-entropies and its corollary the quadratic convergence to the R_ϕ^h-divergences. If we write $f \in C^{\mathrm{i}}(A)$ to denote that the real valued function f has a continuous ith derivative on the set A, then we obtain the following result.

Theorem 2.1. Under the assumptions listed below as regularity conditions, we have

$$\lim_{\varepsilon\to 0} \frac{H^h_\phi(f_\theta) - H^\varepsilon_{(h,\phi)}(f_\theta)}{\varepsilon^2} = \frac{1}{48}h'\left(\int_\chi \phi(f_\theta(x))\,\mathrm{d}x\right)\int_\chi \phi''(f_\theta(x))\, f'_\theta(x)^2\,\mathrm{d}x \tag{5}$$

Regularity conditions:

(i) $\int_{\mathcal{X}} |\phi(f_\theta(x))|\,\mathrm{d}x < \infty$,

(ii) $f_\theta \in C^2(\mathcal{X})$, $\phi \in C^2((0,\infty))$ and $h \in C^1(\mathbb{R})$.

(iii) $\phi''(f_\theta(x))\, f'_\theta(x)^2$ and $\phi'(f_\theta(x))\, f''_\theta(x)$ are Riemann integrable on $\mathcal{X}$.

(iv) If $\nu_{\varepsilon,k},\ z_{\varepsilon,k},\ w_{\varepsilon,k} \in \left[\left(k-\frac{1}{2}\right)\varepsilon, \left(k+\frac{1}{2}\right)\varepsilon\right]$, then

(a) $\lim_{\varepsilon\to 0} \varepsilon \sum_{k\in\mathbb{Z}_\varepsilon} \phi''(f_\theta(z_{\varepsilon,k}))\,(f'_\theta(z_{\varepsilon,k})^2 = \int_{\mathcal{X}} \phi''(f_\theta(x))\,(f'_\theta(x))^2\,\mathrm{d}x$

(b) $\lim_{\varepsilon\to 0} \varepsilon \sum_{k\in\mathbb{Z}_\varepsilon} \phi'(f_\theta(\nu_{\varepsilon,k})\, f''_\theta(w_{\varepsilon,k}) = \int_{\mathcal{X}} \phi'(f_\theta(x))\,(f''_\theta(x))^2\,\mathrm{d}x$.

Proof. A Taylor's expansion of $h\left(\varepsilon \sum_{k\in\mathbb{Z}_\varepsilon} \phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right)\right)$ around the point $\int_{\mathcal{X}} \phi(f_\theta(x))\,\mathrm{d}x$ yields

$$H^h_\phi(f_\theta) - H^\varepsilon_{(h,\phi)}(f_\theta) = h'\left(\int_{\mathcal{X}} \phi(f_\theta(x))\,\mathrm{d}x\right)\left(\int_{\mathcal{X}} \phi(f_\theta(x))\,\mathrm{d}x - \sum_{k\in\mathbb{Z}_\varepsilon} \varepsilon\phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right)\right) + \\ + o\left(\sum_{k\in\mathbb{Z}_\varepsilon} \varepsilon\phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right) - \int_{\mathcal{X}} \phi(f_\theta(x))\,\mathrm{d}x\right).$$

Therefore

$$\lim_{\varepsilon\to 0} \frac{H^h_\phi(f_\theta) - H^\varepsilon_{(h,\phi)}(f_\theta)}{\varepsilon^2} = h'\left(\int_{\mathcal{X}} \phi(f_\theta(x))\,\mathrm{d}x\right) \lim_{\varepsilon\to\infty} \frac{\int_{\mathcal{X}} \phi(f_\theta(x))\,\mathrm{d}x - \sum_{k\in\mathbb{Z}_\varepsilon} \varepsilon\phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right)}{\varepsilon^2}.$$

First, we calculate the difference

$$\int_{\mathcal{X}} \phi(f_\theta(x))\,\mathrm{d}x - \sum_{k\in\mathbb{Z}_\varepsilon} \varepsilon\phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right) = \sum_{k\in\mathbb{Z}_\varepsilon} \left(\int_{k\varepsilon-\frac{1}{2}\varepsilon}^{k\varepsilon+\frac{1}{2}\varepsilon} \phi(f_\theta(x))\,\mathrm{d}x - \varepsilon\phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right)\right).$$

Observe that if $\int_{k\varepsilon-a}^{k\varepsilon+a} \phi(f_\theta(x))\,\mathrm{d}x = F(k\varepsilon+a) - F(k\varepsilon-a)$, then

$$F(k\varepsilon+a) = F(k\varepsilon) + \phi(f_\theta(k\varepsilon))\,a + \phi'(f_\theta(k\varepsilon))\, f'_\theta(k\varepsilon)\,\frac{a^2}{2} + \\ + \left\{\phi''(f_\theta(z_{\varepsilon,k}))\, f'_\theta(z_{\varepsilon,k})^2 + \phi'(f_\theta(z_{\varepsilon,k}))\right\}\frac{a^3}{6}.$$

So, by taking $a = \frac{\varepsilon}{2}$, we obtain

$$\int_{k\varepsilon-\frac{1}{2}\varepsilon}^{k\varepsilon+\frac{1}{2}\varepsilon} \phi(f_\theta(x))\,\mathrm{d}x = \phi(f_\theta(k\varepsilon))\,\varepsilon + \frac{1}{24}\left\{\phi''(f_\theta(z_{\varepsilon,k}))\, f'(z_{\varepsilon,k})^2 + \right. \\ \left. + \phi'(f_\theta(z_{\varepsilon,k}))\, f''_\theta(z_{\varepsilon,k})\right\}\varepsilon^3. \qquad (6)$$

A Taylor's expansion of $\phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right)$ around the point $f_\theta(k\varepsilon)$ yields

$$\phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right) = \phi(f_\theta(k\varepsilon)) + \left(\frac{p_k(\varepsilon,\theta)}{\varepsilon} - f_\theta(k\varepsilon)\right)\phi'(r),$$

with r a point belonging to the interval determined by the points $f_\theta(k\varepsilon)$ and $p_k(\varepsilon,\theta)/\varepsilon$. In fact, since $f_\theta(x)$ is continuous in all $\left[(k-\frac{1}{2})\varepsilon,(k+\frac{1}{2})\varepsilon\right]$, $\varepsilon>0$, $k\in\mathbb{Z}_\varepsilon$ we can easily see that there exists a point $w_{\varepsilon,k}\in\left[(k-\frac{1}{2})\varepsilon,(k+\frac{1}{2})\varepsilon\right]$, such that $r=f_\theta(w_{\varepsilon,k})$.

Taking $\phi(x)=x$ in (6), for $\nu_{\epsilon,k}\in\left[(k-\frac{1}{2})\,\varepsilon,\,(k+\frac{1}{2})\,\varepsilon\right]$, we have

$$p_k(\varepsilon,\theta) = \varepsilon\, f_\theta(k\varepsilon) + \frac{\varepsilon^3}{24} f''_\theta(\nu_{\epsilon,k}).$$

Then

$$\phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right) = \phi(f_\theta(k\varepsilon)) + \frac{\varepsilon^2}{24} f''_\theta(\nu_{\epsilon,k})\,\phi'(f_\theta(w_{\epsilon,k})).$$

Therefore, we obtain

$$\int_{k\varepsilon-\frac{1}{2}\varepsilon}^{k\varepsilon+\frac{1}{2}\varepsilon}\phi(f_\theta(x))\,\mathrm{d}x - \phi\left(\frac{p_k(\varepsilon,\theta)}{\varepsilon}\right)\varepsilon = \frac{\varepsilon^2}{24}\left\{\phi''(f_\theta(z_{k,\varepsilon})\,f'_\theta(z_{k,\varepsilon})^2 + \right.$$
$$\left. +\phi'(f_\theta(z_{k,\varepsilon})\,f''(z_{k,\varepsilon}) - f''_\theta(\nu_{\epsilon,k})\,\phi'(f_\theta(w_{\epsilon,k}))\right\}$$

and finally

$$\lim_{\epsilon\to 0}\frac{H^h_\phi(f_\theta) - H^\varepsilon_{(h,\phi)}(f_\theta)}{\varepsilon^2} = \frac{1}{48}h'\left(\int_{\mathcal{X}}\phi(f_\theta(x))\,\mathrm{d}x\right)\int_{\mathcal{X}}\phi''(f_\theta(x))\,f'_\theta(x)^2\,\mathrm{d}x.$$

Remark 2.1. The regularity conditions for this theorem are essentially the same as those of Ghurye and Johnson [5], except that the convexity of f_θ on the tails of $S(f_\theta)$ is not required. Conditions (i)–(iii) are fairly easy to check. For condition (iv), Lemmas 1, 2 and 3 of [5] are applicable. The requirement that the functions f_1 and f_2 used in these lemmas should be positive valued at the tails of their domains of definition may be relaxed to the requirement that it should not change sign at these tails as has been shown by Zografos, Ferentinos and Papaioannou [1].

Remark 2.2. Let Δ be the right hand side expression of (5), which is negative because of the conditions assumed on h and ϕ, i.e., (increasing, concave) or (decreasing, convex). It is immediate to obtain Δ for several discretized versions of entropy measures. For instance, if we consider $h(x)=x$ and $\phi(x)=-x\log x$, we have the discretized version of Shannon's entropy and in this case $-\Delta$ is given by

$$\frac{1}{48}\int_{\mathcal{X}}\left(\frac{\partial\log f_\theta(x)}{\partial x}\right)^2 f_\theta(x)\,\mathrm{d}x = \frac{1}{48}E_\theta\left(\left(\frac{\partial\log f_\theta(X)}{\partial x}\right)^2\right),$$

which is Fisher's information number for $f_\theta(x)$.

In the following table we present some examples:

Distribution Family	$-\Delta$
Gamma Family ($\alpha > 2,\ \lambda > 0$)	$\lambda^2/(48(\alpha-2))$
Normal Family ($\mu \in \mathbb{R},\ \sigma > 0$)	$(48\sigma^2)^{-1}$
Beta Family ($\alpha > 2,\ \beta > 2$)	$((\alpha+\beta-1)(\alpha+\beta-2)(\alpha+\beta-4))/(48(\alpha-2)(\beta-2))$

Note that for the Normal family, $-\Delta$ is proportional to the inverse of the variance. In the remaining cases $-\Delta$ is approximately proportional to the inverse of the variance.

In a similar way to the previous theorem, if we suppose that the supports of the generalized density functions f_{θ_1} and f_{θ_2} are such that $S(f_{\theta_1}) = S(f_{\theta_2}) = \mathcal{X} \subset \mathbb{R}$ is an open set, then we obtain that the rate of convergence of the discrete approximation to the R^h_ϕ-divergence is quadratic.

Corollary 2.1. Under a straightforward extension of the regularity conditions given in Theorem 2.1, to f_{θ_1}, f_{θ_2} and $f_{\theta_1} + f_{\theta_2}$, we have

$$\begin{aligned}
\Delta^* &= \lim_{\varepsilon \to 0} \frac{R^h_\phi(f_{\theta_1}, f_{\theta_2}) - R^\varepsilon_{(h\phi)}(f_{\theta_1}, f_{\theta_2})}{\varepsilon^2} = \\
&= \frac{1}{24}\left\{ h'\left(\int_{\mathcal{X}} \phi\left(\frac{f_{\theta_1}(x) + f_{\theta_2}(x)}{2}\right) \mathrm{d}x\right) \cdot \right. \\
&\quad \cdot \int_{\mathcal{X}} \phi''\left(\frac{f_{\theta_1}(x) + f_{\theta_2}(x)}{2}\right)\left(\frac{f'_{\theta_1}(x) + f'_{\theta_2}(x)}{2}\right)^2 \mathrm{d}x - \\
&\quad -\frac{1}{2}h'\left(\int_{\mathcal{X}} \phi(f_{\theta_1}(x))\,\mathrm{d}x\right)\int_{\mathcal{X}} \phi''(f_{\theta_1}(x))\, f'_{\theta_1}(x)^2\,\mathrm{d}x - \\
&\quad \left. -\frac{1}{2}h'\left(\int_{\mathcal{X}} \phi(f_{\theta_2}(x))\,\mathrm{d}x\right)\int_{\mathcal{X}} \phi''(f_{\theta_2}(x))\, f'_{\theta_2}(x)^2\,\mathrm{d}x \right\}.
\end{aligned}$$

Remark 2.3. If we consider $h(x) = x$ and $\phi(x) = -x\log x$, i.e., for the Information Radius, then we obtain that Δ^* is given by

$$\frac{1}{48}\left\{ E_{\theta_1}\left(\left(\frac{\partial \log f_{\theta_1}(X)}{\partial x}\right)^2 - \left(\frac{\partial \log f_{\theta_1}(X) + f_{\theta_2}(X))}{\partial x}\right)\right) + E_{\theta_2}\left(\left(\frac{\partial \log f_{\theta_2}(X)}{\partial x}\right)^2 - \left(\frac{\partial \log f_{\theta_1}(X) + f_{\theta_2}(X))}{\partial x}\right)\right)\right\}.$$

3. ON ESTIMATING THEORETICAL ENTROPIES AND DIVERGENCES THROUGH DISCRETIZED DATA. NONPARAMETRIC APPROACH.

In the previous discretization scheme, the amount of information lost due to discretization or grouping of the data is given by

$$D(\varepsilon) = H^h_\phi(f) - H^\varepsilon_{(h,\phi)}(f) \quad \text{or} \quad B(\varepsilon) = R^h_\phi(f_1, f_2) - R^\varepsilon_{(h,\phi)}(f_1, f_2)$$

for the case of entropies or R-divergences respectively. If f and f_1 or f_2 are unknown but grouped data are available from them, which is the case of many practical applications, then we can use statistical methods to estimate $H^h_\phi(f)$ and/or $R^h_\phi(f_1, f_2)$.

We consider two possibilities depending whether $\mathcal{X}$ is bounded or not. Let us first suppose that $\mathcal{X} = (a, b)$, where $k\varepsilon - \frac{\varepsilon}{2} < a < k\varepsilon + \frac{\varepsilon}{2}$ and $(k + K - 1)\varepsilon - \frac{\varepsilon}{2} < b < (k + K - 1)\varepsilon + \frac{\varepsilon}{2}$ for some $k \in \mathbb{Z}_\varepsilon$ and some $K \in \mathbb{N}$; i.e., we have K classes with probabilities

$$p_1(\varepsilon) = \int_a^{k\varepsilon+\frac{1}{2}\varepsilon} f_\theta(x)\,\mathrm{d}x, \qquad p_K(\varepsilon) = \int_{(k+K-1)\varepsilon-\frac{1}{2}\varepsilon}^{b} f_\theta(x)\,\mathrm{d}x$$

and

$$p_j(\varepsilon) = \int_{j\varepsilon-\frac{1}{2}\varepsilon}^{j\varepsilon+\frac{1}{2}\varepsilon} f_\theta(x)\,\mathrm{d}x, \quad j = 2, \ldots, K-1. \tag{7}$$

To estimate $H_\phi^h(f)$ on the basis of a discretized random sample of size n from f, we define the discretized sample estimate as follows:

$$H_{(h,\phi)}^\varepsilon(\hat{P}) = h\left(\varepsilon \sum_{j=1}^{K} \phi\left(\frac{\hat{p}_j(\varepsilon)}{\varepsilon}\right)\right)$$

where $\hat{P} = (\hat{p}_1(\varepsilon), \ldots, \hat{p}_K(\varepsilon))^t$ and $\hat{p}_j(\varepsilon)$ is the relative frequence associated to the probability $p_j(\varepsilon)$. Let us also define

$$H_{(h,\phi)}^\varepsilon(P) = H_{(h,\phi)}^\varepsilon(f) = h\left(\varepsilon \sum_{j=1}^{K} \phi\left(\frac{p_j(\varepsilon)}{\varepsilon}\right)\right),$$

where $P = (p_1(\varepsilon), \ldots, p_K(\varepsilon))^t$.

The following theorem gives the asymptotic distribution of $H_{(h,\phi)}^\varepsilon(\hat{P})$.

Theorem 3.1. If $h \in C^1(\mathbb{R})$ and $\Phi \in C^1((0, \infty))$, then

$$n^{1/2}\left(H_{(n,\phi)}^\varepsilon(\hat{P}) - H_{(n,\phi)}^\varepsilon(P)\right) \xrightarrow[n\to\infty]{L} N(0, \sigma^2),$$

where

$$\sigma^2 = \sum_{i=1}^{K} w_i^2\, p_i(\varepsilon) - \left(\sum_{i=1}^{K} w_i\, p_i(\varepsilon)\right)^2$$

and

$$w_i = \frac{\partial H_{(h,\phi)}^\varepsilon(P)}{\partial p_i(\varepsilon)} = h'\left(\varepsilon \sum_{j=1}^{K} \phi\left(\frac{p_j(\varepsilon)}{\varepsilon}\right)\right) \phi'\left(\frac{p_i(\varepsilon)}{\varepsilon}\right).$$

Proof. We consider the Taylor expansion of $H_{(h,\phi)}^\varepsilon(\hat{P})$ around the point P

$$H_{(h,\phi)}^\varepsilon(\hat{P}) = H_{(h,\phi)}^\varepsilon(P) + \sum_{i=1}^{K} \frac{\partial H_{(h,\phi)}^\varepsilon(P)}{\partial p_i(\varepsilon)}(\hat{p}_i(\varepsilon) - p_i(\varepsilon)) + R_n.$$

As $n^{1/2} R_n \xrightarrow[n\to\infty]{L} 0$, applying the Central Limit Theorem, we conclude that

$$n^{1/2}\left[H_{(h,\phi)}^\varepsilon(\hat{P}) - H_{(h,\phi)}^\varepsilon(P)\right] \xrightarrow[n\to\infty]{L} N(0, W^t \Sigma_P W),$$

where $\Sigma_P = \mathrm{diag}(P) - P\,P^{\mathrm{t}}$ and $W = (w_1, \ldots, w_K)^{\mathrm{t}}$. □

Now we consider the loss due to estimating the theoretical (h, ϕ)- entropy through discretized data; i.e., $\tilde{D}(\varepsilon) = H^h_\phi(f) - H^\varepsilon_{(h,\phi)}(\hat{P}) \cdot \tilde{D}(\varepsilon)$ is still an unknown quantity, but we know that

$$\begin{aligned} D(\varepsilon) - \tilde{D}(\varepsilon) &= \left(H^h_\phi(f) - H^\varepsilon_{(h,\phi)}(P)\right) - \left(H^h_\phi(f) - H^\varepsilon_{(h,\phi)}(\hat{P})\right) = \\ &= H^\varepsilon_{(h,\phi)}(\hat{P}) - H^\varepsilon_{(h,\phi)}(P), \end{aligned}$$

whose asymptotic distribution is given in Theorem 3.1. So, a $(1-\alpha)\,100\,\%$ large sample confidence interval for $D(\varepsilon) - \tilde{D}(\varepsilon)$, is

$$\left(-z_{\alpha/2}\frac{\hat{\sigma}}{n^{1/2}},\, z_{\alpha/2}\frac{\hat{\sigma}}{n^{1/2}}\right),$$

where $\hat{\sigma}$ is obtained by replacing $p_k(\varepsilon)$ by $\hat{p}_k(\varepsilon)$ in Theorem 3.1. Finally, if $\hat{\sigma}$ is "small", then $D(\varepsilon) \approx \tilde{D}(\varepsilon)$ and as $D(\varepsilon) \approx \varepsilon^2\Delta$ is also "small", so is $\tilde{D}(\varepsilon)$.

Now we treat the case where $\mathcal{X}$ is an open and not bounded interval of $\mathbb{R}$. For any $c > 0$, let us define the open interval $(a, b) = \mathcal{X} \cap (-c, c)$, where $k\varepsilon - \frac{\varepsilon}{2} < a < k\varepsilon + \frac{\varepsilon}{2}$ and $(k + K - 1)\varepsilon - \frac{\varepsilon}{2} < b < (k + K - 1)\varepsilon + \frac{\varepsilon}{2}$ for some $k \in \mathbb{Z}_\varepsilon$ and some $K \in \mathbb{N}$; i.e., we have K classes whose probabilities $p_i(\varepsilon)$, $i = 2, \ldots, K-1$, are given in (7) and the remaining probabilities are

$$p_1(\varepsilon) = \int_{-\infty}^{k\varepsilon+\frac{1}{2}\varepsilon} f_\theta(x)\,\mathrm{d}x \quad \text{and} \quad p_K(\varepsilon) = \int_{(k+K-1)\varepsilon-\frac{1}{2}\varepsilon}^{+\infty} f_\theta(x)\,\mathrm{d}x.$$

Let us also define

$$H^{\varepsilon,c}_{(h,\phi)}(\hat{P}) = h\left(\varepsilon \sum_{j=1}^{K} \phi\left(\frac{\hat{p}_j(\varepsilon)}{\varepsilon}\right)\right).$$

where $\hat{P} = (\hat{p}_1(\varepsilon), \ldots, \hat{p}_K(\varepsilon))^{\mathrm{t}}$ is the relative frequence vector associated to the probability vector $P = (p_1(\varepsilon), \ldots, p_K(\varepsilon))^{\mathrm{t}}$. So Theorem 3.1 can be applied to obtain the asymptotic distribution of $n^{1/2}\left(H^{\varepsilon,c}_{(h,\phi)}(\hat{P}) - H^{\varepsilon,c}_{(h,\phi)}(P)\right)$.

Now we consider the loss due to estimating the theoretical (h, ϕ)-entropy through truncated discretized data; i.e., $\tilde{D}^c(\varepsilon) = H^h_\phi(f) - H^{\varepsilon,c}_{(h,\phi)}(\hat{P}).\tilde{D}^c(\varepsilon)$ is still an unknown quantity, but we know that

$$D(\varepsilon) - \tilde{D}(\varepsilon) = H^{\varepsilon,c}_{(h,\phi)}(\hat{P}) - H^\varepsilon_{(h,\phi)}(P).$$

If we now suppose that for a sufficiently small $\eta > 0$, there exist a $c > 0$ such that $\left|H^{\varepsilon,c}_{(h,\phi)}(P) - H^\varepsilon_{(h,\phi)}(P)\right| < \eta$, then a $(1-\alpha)\,100\,\%$ large sample confidence interval for $D(\varepsilon) - \tilde{D}^c(\varepsilon)$, is

$$\left(-\eta - z_{\alpha/2}\frac{\hat{\sigma}}{n^{1/2}},\, \eta + z_{\alpha/2}\frac{\hat{\sigma}}{n^{1/2}}\right),$$

where $\hat{\sigma}$ is obtained by replacing $p_k(\varepsilon)$ by $\hat{p}_k(\varepsilon)$ in Theorem 3.1. Finally, if $\hat{\sigma}$ and η are "small", then $D(\varepsilon) \approx \tilde{D}^c(\varepsilon)$ and as $D(\varepsilon) \approx \varepsilon^2\Delta$ is also "small", so is $\tilde{D}^c(\varepsilon)$.

To estimate $R_\phi^h(f_1, f_2)$, we consider two possibilities when $\mathcal{X}_1 = \mathcal{X}_2 = (a, b)$: (1) f_1 unknown, (2) f_1 and f_2 unknown. If (a, b) is partitioned in K disjoint intervals, let us define according to (7) $p_i(\varepsilon)$ and $q_i(\varepsilon)$, $i = 1, \ldots, K$, to be the corresponding probabilities under f_1 and f_2 respectively. Let us consider the probability vectors $P = (p_1(\varepsilon), \ldots, p_K(\varepsilon))^t$ and $Q = (q_1(\varepsilon), \ldots, q_K(\varepsilon))^t$ and the relative frequency vectors $\hat{P} = (\hat{p}_1(\varepsilon), \ldots, \hat{p}_K(\varepsilon))^t$ and $\hat{Q} = (\hat{q}_1(\varepsilon), \ldots, \hat{q}_K(\varepsilon))^t$ when independent discretized random samples of sizes n and m are observed from f_1 and f_2 respectively. Let us also define $R^\varepsilon_{(h,\phi)}(P, Q) = R^\varepsilon_{(h,\phi)}(f_1, f_2)$ and

$$R^\varepsilon_{h,\phi}(\hat{P}, Q) = H^\varepsilon_{(h,\phi)}\left(\frac{\hat{P}+Q}{2}\right) - \frac{1}{2}H^\varepsilon_{(h,\phi)}(\hat{P}) - \frac{1}{2}H^\varepsilon_{(h,\phi)}(Q)$$

when f_1 is unknown, and

$$R^\varepsilon_{h,\phi}(\hat{P}, \hat{Q}) = H^\varepsilon_{(h,\phi)}\left(\frac{\hat{P}+\hat{Q}}{2}\right) - \frac{1}{2}H^\varepsilon_{(h,\phi)}(\hat{P}) - \frac{1}{2}H^\varepsilon_{(h,\phi)}(\hat{Q})$$

when both f_1 and f_2 are unknown. In an analogous way to the case of $H^h_\phi(f)$, one obtains the following theorem.

Theorem 3.2. (a) If $h \in C^1(\mathbb{R})$ and $\Phi \in C^1((0, \infty))$, then

$$n^{1/2}\left(R^\varepsilon_{(h,\phi)}(\hat{P}, Q) - R^\varepsilon_{(h,\phi)}(P, Q)\right) \xrightarrow[n\to\infty]{L} N(0, \sigma_1^2),$$

where

$$\sigma_1^2 = \sum_{i=1}^K t_i^2\, p_i(\varepsilon) - \left(\sum_{i=1}^K t_i\, p_i(\varepsilon)\right)^2$$

and

$$\begin{aligned} t_i &= -\frac{1}{2}h'\left(\varepsilon\sum_{j=1}^K \phi\left(\frac{p_j(\varepsilon)}{\varepsilon}\right)\right)\phi'\left(\frac{p_i(\varepsilon)}{\varepsilon}\right) + \\ &\quad + \frac{1}{2}h'\left(\varepsilon\sum_{j=1}^K \phi\left(\frac{p_j(\varepsilon)+q_j(\varepsilon)}{2\varepsilon}\right)\right)\phi'\left(\frac{p_i(\varepsilon)+q_i(\varepsilon)}{2\varepsilon}\right). \end{aligned}$$

(b) If $\frac{m}{n+m} \xrightarrow[n\to\infty]{L} \lambda \in (0, 1)$, then

$$\left(\frac{mn}{n+m}\right)^{1/2}\left(R^\varepsilon_{(h,\phi)}(\hat{P}, \hat{Q}) - R^\varepsilon_{(h,\phi)}(P, Q)\right) \xrightarrow[n\to\infty]{L} N(0, \sigma^2),$$

where $\sigma^2 = \lambda\,\sigma_1^2 + (1-\lambda)\,\sigma_2^2$, $\sigma_2^2 = \sum_{i=1}^K s_i^2\, q_i(\varepsilon) - \left(\sum_{i=1}^K s_i\, q_i(\varepsilon)\right)^2$,

$$s_i = -\frac{1}{2}h'\left(\varepsilon\sum_{j=1}^K \phi\left(\frac{q_j(\varepsilon)}{\varepsilon}\right)\right)\phi'\left(\frac{q_i(\varepsilon)}{\varepsilon}\right) +$$

$$+\frac{1}{2}h'\left(\varepsilon\sum_{j=1}^{K}\phi\left(\frac{p_j(\varepsilon)+q_j(\varepsilon)}{2\varepsilon}\right)\right)\phi'\left(\frac{p_i(\varepsilon)+q_i(\varepsilon)}{2\varepsilon}\right)$$

and n and m are the sizes of the samples of f_1 and f_2 respectively.

Remark 3.1. For the case $h(x) = x$ and $\phi(x) = x\log x$, i.e., for the Information Radius, we get

$$t_i = \frac{1}{2}\log\frac{p_i(\varepsilon)}{p_i(\varepsilon)+q_i(\varepsilon)} \quad \text{and} \quad s_i = \frac{1}{2}\log\frac{q_i(\varepsilon)}{p_i(\varepsilon)+q_i(\varepsilon)}.$$

Now, the loss due to estimating the theoretical R_φ^h-divergence through discretized data is

$$\tilde{B}_1(\varepsilon) = R_\phi^h(f_1, f_2) - R_{(h,\phi)}^\varepsilon(\hat{P}, Q)$$

if f_1 is unknown, and

$$\tilde{B}_2(\varepsilon) = R_\phi^h(f_1, f_2) - R_{(h,\phi)}^\varepsilon(\hat{P}, \hat{Q})$$

if f_1 and f_2 are unknown.

Asymptotic distributions of $B(\varepsilon) - \tilde{B}_1(\varepsilon)$ and $B(\varepsilon) - \tilde{B}_2(\varepsilon)$ are given in Theorem 3.2. So, $(1-\alpha)\,100\,\%$ large sample confidence intervals for $B(\varepsilon) - \tilde{B}_1(\varepsilon)$ and $B(\varepsilon) - \tilde{B}_2(\varepsilon)$, are

$$\left(-z_{\alpha/2}\frac{\hat{\sigma}_1}{n^{1/2}}, z_{\alpha/2}\frac{\hat{\sigma}_1}{n^{1/2}}\right) \quad \text{and} \quad \left(-z_{\alpha/2}\frac{\hat{\sigma}}{\left(\frac{mn}{n+m}\right)^{1/2}}, z_{\alpha/2}\frac{\hat{\sigma}}{\left(\frac{mn}{n+m}\right)^{1/2}}\right),$$

where $\hat{\sigma}_1$ and $\hat{\sigma}$ are obtained by replacing $p_k(\varepsilon)$ and/or $q_k(\varepsilon)$ by $\hat{p}_k(\varepsilon)$ and $\hat{q}_k(\varepsilon)$ respectively in Theorem 3.2. Finally, the case of unbounded support can be treated as in the entropy case and the same considerations can be given.

4. ON ESTIMATING THEORETICAL ENTROPIES AND DIVERGENCES THROUGH DISCRETIZED DATA. PARAMETRIC APPROACH

In Section 3, the problem of estimating theoretical entropies and divergences through discretized data was treated when f and f_1 or f_2 could not be included in any parametric family of distributions. In this section, we again consider a statistical space $(\mathcal{X}, \beta_\mathcal{X}, P_\theta)_{\theta\in\Theta}$, where $\mathcal{X}$ is an open subset of $\mathbb{R}$, Θ is an open subset of $\mathbb{R}^M$ and f_θ and f_{θ_1} or f_{θ_2} are the Radon–Nikodym derivatives of P_θ and P_{θ_1} or P_{θ_2} with respect to the Lebesgue measure in $(\mathcal{X}, \beta_\mathcal{X})$.

In the previous discretization scheme, where $\{P_\theta\}_{\theta\in\Theta}$ is a well known family of probability functions (Gammma, Normal, Beta,...) and only discretized data is available, it seems more reasonable to estimate $H_\phi^h(\theta) = H_\phi^h(f_\theta)$ and $R_\phi^h(\theta_1, \theta_2) = R_\phi^h(f_{\theta_1}, f_{\theta_2})$ better by means of $H_\phi^h(\tilde{\theta})$ and $R_\phi^h(\tilde{\theta}_1, \tilde{\theta}_2)$, where $\tilde{\theta}$, $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are the maximum likelihood estimators (M.L.E.) of θ, θ_1 and θ_2 respectively based on the available discretized data, than by means of $H_{(h,\phi)}^{\varepsilon,c}(\hat{P})$ and $R_{(h,\phi)}^{\varepsilon,c}(\hat{P}, \hat{Q})$.

In this section we first develop a general notation useful for discussing a variety of issues that arise in testing and estimation for the multinomial distribution. Let $X = (X_1, \ldots, X_K)^t$ be a K-dimentional random vector with the multinomial distribution $X \stackrel{d}{=} M_K(n, P)$, where $P = (p_1, \ldots, p_K)^t$ is a vector of cell probabilities and $n = \sum_{i=1}^{K} X_i$. We let Δ_K be the set of all possible K-dimensional probability vectors; i.e.,

$$\Delta_K = \left\{ P \in \mathbb{R}^K : p_i \geq 0,\ i = 1, \ldots, K, \sum_{i=1}^{K} p_i = 1 \right\}.$$

The vector of observed proportions $\hat{P} = n^{-1}X$ is also a point of Δ_K. There exist a function $g(\theta)$ that maps each value of a vector $\theta = (\theta_1, \ldots, \theta_M)^t$ into a point in Δ_K. When we assume that a given multinomial parametric model is correct, we are really just assuming that there exist a parameter value θ^0 in Θ such that the true cell probability vector verifies $P = g(\theta^0)$. In this section we assume the following six regularity conditions given by [1]:

1. The point θ^0 is an interior point of Θ.
2. $p_i = g_i(\theta^0) > 0$ for $i = 1, \ldots, K$.
3. The mapping $g : \Theta \to \Delta_K$ is totally differentiable at θ^0, so that the partial derivatives of g_i with respect to each θ_j exists at θ^0 and $g(\theta)$ has a linear approximation at θ^0 given by

$$g_i(\theta) = g_i(\theta^0) + \sum_{j=1}^{M} (\theta_j - \theta_j^0) \frac{\partial g_i(\theta^0)}{\partial \theta_j} + o\left(\|\theta - \theta^0\|\right).$$

   as $\theta \to \theta^0$.
4. The Jacobian matrix $\left(\frac{\partial g}{\partial \theta}\right)$, whose (i, j) element is $\frac{\partial g_i(\theta^0)}{\partial \theta_j}$, is of full rank; i.e., rank M.
5. The inverse mapping $g^{-1} : \Theta \to \Delta_K$ is continuous at $g(\theta^0) = p$.
6. The mapping $g : \Theta \to \Delta_K$ is continuous at every point θ in Θ.

Birch in [1] gives the asymptotic distribution of the M.L.E. $\tilde{\theta}$ based on the discretized data. This is given in the next theorem.

Theorem 4.1. Under conditions 1–6 and assuming that $P = g(\theta^0)$, the asymptotic distribution of $\tilde{\theta}$ is given by

$$n^{1/2}(\tilde{\theta} - \theta) \xrightarrow[n \to \infty]{L} N(0, (A^t A)^{-1}),$$

where $A_{K \times M} = \operatorname{diag}\left(g(\theta^0)^{-1/2}\right) \left(\frac{\partial g(\theta^0)}{\partial \theta}\right)$.

From Theorem 4.1, we obtain the asymptotic distributions of $H_\phi^h(\tilde{\theta})$ and $R_\phi^h(\tilde{\theta}_1, \tilde{\theta}_2)$.

Theorem 4.2. Under conditions 1–6, if θ is the true value of the parameter, $f_\theta \in C^1(\mathcal{X})$, $h \in C^1(\mathbb{R})$, $\phi \in C^1((0,\infty))$ and $\left|\frac{\partial}{\partial\theta_1}\phi(f_\theta(x))\right| < \Phi(x)$, where Φ is finitely integrable in $\mathcal{X}$, then

$$n^{1/2}\left[H_\phi^h(\tilde{\theta}) - H_\phi^h(\theta)\right] \xrightarrow[n\to\infty]{L} N(0,\sigma^2),$$

where $\sigma^2 = T^{\mathrm{t}}(A^{\mathrm{t}}\,A)^{-1}\,T$, $T = (t_1,\ldots,t_m)^{\mathrm{t}}$ and

$$t_i = \frac{\partial H_\phi^h(\theta)}{\partial\theta_i} = h'\left(\int_{\mathcal{X}}\phi(f_\theta(x))\,\mathrm{d}x\right)\int_{\mathcal{X}}\phi'(f_\theta(x))\,\frac{\partial f_\theta(x)}{\partial\theta_i}\,\mathrm{d}x, \quad i = 1,\ldots,M.$$

Proof. We consider the Taylor expansion of $H_\phi^h(\tilde{\theta})$ around the point θ

$$H_\phi^h(\tilde{\theta}) = H_\phi^h(\theta) + T^{\mathrm{t}}(\tilde{\theta}-\theta) + R_n.$$

Due to the fact that $n^{1/2}R_n$ converges in probability to 0 when $n \to \infty$, one gets the result. □

Now we consider the loss due to estimating the theoretical (h,ϕ)-entropies through discretized data under a parametric model assumption. First we suppose that $\mathcal{X} = (a,b)$, where $k\varepsilon - \frac{\varepsilon}{2} < a < k\varepsilon + \frac{\varepsilon}{2}$ and $(k+K-1)\varepsilon - \frac{\varepsilon}{2} < b < (k+K-1)\varepsilon + \frac{\varepsilon}{2}$ for some $k \in \mathbb{Z}_\varepsilon$ and some $K \in \mathbb{N}$; i.e., we have K classes whose probabilities $p_i(\varepsilon), i = 1,\ldots,K$, are given in (7). In this case $D^*(\varepsilon) = H_\phi^h(\theta) - H_\phi^h(\tilde{\theta})$, where $\tilde{\theta}$ is the M.L.E. of θ based on the multinomial model with $P = (p_1(\varepsilon),\ldots,p_K(\varepsilon))^{\mathrm{t}}$. As the asymptotic distribution of $D^*(\varepsilon)$ is given in Theorem 4.2, a $(1-\alpha)\,100\,\%$ large sample confidence interval for $D^*(\varepsilon)$ is

$$\left(-z_{\alpha/2}\frac{\hat{\sigma}}{n^{1/2}},\ z_{\alpha/2}\frac{\hat{\sigma}}{n^{1/2}}\right),$$

where $\hat{\sigma}$ is obtained by replacing θ by $\tilde{\theta}$ in Theorem 4.2.

For the case that $\mathcal{X}$ is an open and not bounded interval of $\mathbb{R}$, we suppose that for any $\eta > 0$, there exist a $c > 0$ such that $\mathcal{X}\cap(-c,c) = (a,b)$ and $|{}^cH_h^\phi(\theta) - H_h^\phi(\theta)| < \eta$, where

$${}^cH_\phi^h(\theta) = h\left(\int_a^b \phi(f_\theta(x))\,\mathrm{d}x\right).$$

So, as in Section 3, a $(1-\alpha)\,100\,\%$ large sample confidence interval for $D_c^*(\varepsilon) = H_h^\phi(\theta) - {}^cH_h^\phi(\tilde{\theta})$, is

$$\left(-\eta - z_{\alpha/2}\frac{\hat{\sigma}}{n^{1/2}},\ \eta + z_{\alpha/2}\frac{\hat{\sigma}}{n^{1/2}}\right),$$

where $\hat{\sigma}$ is obtained by replacing θ by $\tilde{\theta}$ in Theorem 4.2.

To estimate $R_\phi^h(\theta_1,\theta_2)$, we consider two possibilities when $\mathcal{X}_1 = \mathcal{X}_2 = (a,b)$: (1) θ_1 unknown, (2) θ_1 and θ_2 unknown. If (a,b) is partitioned in K disjoint

intervals, let us define according to (7) $p_i(\varepsilon)$ and $q_i(\varepsilon), i = 1, \ldots, K$, to be the corresponding probabilities under f_{θ_1} and f_{θ_2} respectively. Let us consider the probability vectors $P = (p_1(\varepsilon), \ldots, p_K(\varepsilon))^t$ and $Q = (q_1(\varepsilon), \ldots, q_K(\varepsilon))^t$ and also the relative frequency vectors $\hat{P} = (\hat{p}_1(\varepsilon), \ldots, \hat{p}_K(\varepsilon))^t$ and $\hat{Q} = (\hat{q}_1(\varepsilon), \ldots, \hat{q}_K(\varepsilon))^t$ when independent discretized random samples of sizes n and m are observed from f_{θ_1} and f_{θ_2} respectively. Let us define $\theta_1 = (\theta_{11}, \ldots, \theta_{1M})$, $\theta_2 = (\theta_{21}, \ldots, \theta_{2M})$,

$$A_{K\times M} = \operatorname{diag}\left(g(\theta_1)^{-1/2}\right)\left(\frac{\partial g(\theta_1)}{\partial \theta_1}\right), \quad B_{K\times M} = \operatorname{diag}\left(g(\theta_2)^{-1/2}\right)\left(\frac{\partial g(\theta_2)}{\partial \theta_2}\right),$$

$$R^h_\phi(\tilde{\theta}_1, \theta_2) = H^h_\phi\left(\frac{\tilde{\theta}_1 + \theta_2}{2}\right) - \frac{1}{2}H^h_\phi(\tilde{\theta}_1) - \frac{1}{2}H^h_\phi(\theta_2),$$

when θ_1 is unknown, and

$$R^h_\phi(\tilde{\theta}_1, \tilde{\theta}_2) = H^h_\phi\left(\frac{\tilde{\theta}_1 + \tilde{\theta}_2}{2}\right) - \frac{1}{2}H^h_\phi(\tilde{\theta}_1) - \frac{1}{2}H^h_\phi(\tilde{\theta}_2),$$

when both θ_1 and θ_2 are unknown. Let us also define $T = (t_1, \ldots, t_M)^t$ and $S = (s_1, \ldots, s_M)^t$ with

$$\begin{aligned} t_i &= \frac{1}{2}\left\{h'\left(\int_{\mathcal{X}} \phi(f_{\theta_1}(x))\,\mathrm{d}x\right)\left(\int_{\mathcal{X}} \phi'(f_{\theta_1}(x))\frac{\partial f_{\theta_1}(x)}{\partial \theta_{1i}}\,\mathrm{d}x\right) - \right. \\ &- \left. h'\left(\int_{\mathcal{X}} \phi\left(\frac{f_{\theta_1}(x) + f_{\theta_2}(x)}{2}\right)\mathrm{d}x\right)\left(\int_{\mathcal{X}} \phi'\left(\frac{f_{\theta_1}(x) + f_{\theta_2}(x)}{2}\right)\frac{\partial f_{\theta_1}(x)}{\partial \theta_{1i}}\,\mathrm{d}x\right)\right\}, \end{aligned}$$

and

$$\begin{aligned} s_i &= \frac{1}{2}\left\{h'\left(\int_{\mathcal{X}} \phi(f_{\theta_2}(x))\,\mathrm{d}x\right)\left(\int_{\mathcal{X}} \phi'(f_{\theta_2}(x))\frac{\partial f_{\theta_2}(x)}{\partial \theta_{2i}}\,\mathrm{d}x\right) - \right. \\ &- \left. h'\left(\int_{\mathcal{X}} \phi\left(\frac{f_{\theta_1}(x) + f_{\theta_2}(x)}{2}\right)\mathrm{d}x\right)\left(\int_{\mathcal{X}} \phi'\left(\frac{f_{\theta_1}(x) + f_{\theta_2}(x)}{2}\right)\frac{\partial f_{\theta_2}(x)}{\partial \theta_{2i}}\,\mathrm{d}x\right)\right\}. \end{aligned}$$

In an analogous way to the case of $H^h_\phi(\theta)$, one obtains the following theorem.

Theorem 4.3. Let $\tilde{\theta}_1$ and $\tilde{\theta}_2$ be the M.L.E. of θ_1 and θ_2 based on independent discretized samples of size n and m respectively. Let us suppose that conditions 1-6 hold, θ_1 and θ_2 are the true values of the parameter for f_1 and f_2 respectively, $f_{\theta_i} \in C^1(\mathcal{X})$, $i = 1, 2$, $h \in C^1(\mathbb{R})$, $\phi \in C^1((0, \infty))$ and $\left|\frac{\partial}{\partial \theta_{ij}}\phi(f_{\theta_i}(x))\right| < \Phi(x)$ and $\left|\frac{\partial}{\partial \theta_{ij}}\phi\left(\frac{f_{\theta_1}(x)+f_{\theta_2}(x)}{2}\right)\right| < \Phi(x)$, $i = 1, 2$, $j = 1, \ldots, M$, where $\Phi(x)$ is finitely integrable in $\mathcal{X}$.

(a) If θ_2 is known, then

$$n^{1/2}\left(R^h_\phi(\tilde{\theta}_1, \theta_2) - R^h_\phi(\theta_1, \theta_2)\right) \xrightarrow[n\to\infty]{L} N\left(0, T^t(A^t A)^{-1} T\right).$$

(b) If $\frac{m}{n+m} \xrightarrow[n,m\to\infty]{L} \lambda \in (0,1)$, then

$$\left(\frac{m\,n}{n+m}\right)^{1/2}\left(R_\varphi^h(\tilde\theta_1,\tilde\theta_2)-{}_\varepsilon R_\varphi^h(\theta_1,\theta_2)\right) \xrightarrow[n\to\infty]{L}$$
$$\xrightarrow[n\to\infty]{L} N\left(0,\ \lambda T^{\mathrm{t}}(A^{\mathrm{t}}A)^{-1}T+(1-\lambda)S^{\mathrm{t}}(B^{\mathrm{t}}B)^{-1}S\right).$$

Remark 4.1. For the case $h(x)=x$ and $\phi(x)=x\log x$, i.e., for the Information Radius, we get

$$t_i=\int_{\mathcal{X}}\frac{1}{2}\frac{\partial f_{\theta_1}(x)}{\partial\theta_{1i}}\log\frac{2\,f_{\theta_1}(x)}{f_{\theta_1}(x)+f_{\theta_2}(x)}\,\mathrm{d}x$$

and

$$s_i=\int_{\mathcal{X}}\frac{1}{2}\frac{\partial f_{\theta_2}(x)}{\partial\theta_{2i}}\log\frac{2\,f_{\theta_2}(x)}{f_{\theta_1}(x)+f_{\theta_2}(x)}\,\mathrm{d}x.$$

Now, the loss due to estimating the theoretical R_ϕ^h-divergence through discretized data is

$$B_1^*(\varepsilon)=R_\phi^h(\theta_1,\theta_2)-R_\phi^h(\tilde\theta_1,\theta_2)$$

if θ_1 is unknown, and

$$B_2^*(\varepsilon)=R_\phi^h(\theta_1,\theta_2)-R_\phi^h(\tilde\theta_1,\tilde\theta_2)$$

if θ_1 and θ_2 are unknown.

Asymptotic distributions of $B_1^*(\varepsilon)$ and $B_2^*(\varepsilon)$ are given in Theorem 4.3. So, $(1-\alpha)\,100\,\%$ large sample confidence intervals for $B_1^*(\varepsilon)$ and $B_2^*(\varepsilon)$ are

$$\left(-z_{\alpha/2}\frac{\sigma_1}{n^{1/2}},\ z_{\alpha/2}\frac{\sigma_1}{n^{1/2}}\right)\quad\text{and}\quad\left(-z_{\alpha/2}\frac{\hat\sigma}{\left(\frac{m\,n}{n+m}\right)^{1/2}},\ z_{\alpha/2}\frac{\hat\sigma}{\left(\frac{m\,n}{n+m}\right)^{1/2}}\right),$$

where $\hat\sigma_1$ and $\hat\sigma$ are obtained by replacing θ_1 and/or θ_2 by $\tilde\theta_1$ and $\tilde\theta_2$ respectively in Theorem 4.3. Finally, the case of unbounded support can be treated as in the entropy case and the same considerations can be given. Now, we give some applications to testing statistical hypotheses.

Remark 4.2. The previous results giving the asymptotic distribution of the R_ϕ^h-divergence statistics in random sampling can be used in various settings to construct confidence intervals and to test statistical hypotheses based on one or more samples. We give some examples.

(1) To test the hypothesis that the divergence between θ and θ_0, a predicted value of θ available beforehand to the experimenter, is of a certain magnitude R_0, i.e., $H_0: R_\varphi^h(\theta,\theta_0)=R_0$, we can use the statistic

$$Z=\frac{n^{1/2}\left(R_\phi^h(\tilde\theta,\theta_0)-R_0\right)}{\hat\sigma},$$

which has approximately a standard normal distribution under H_0 for sufficiently large n, and $\hat{\sigma}$ is obtained from Theorem 4.3 by replacing θ by its maximum likelihood estimator $\tilde{\theta}$ in $(T^t(A^tA)^{-1}T)^{1/2}$.

(2) To test the hypothesis that the divergence between θ_1 and θ_2 is of a certain magnitude R_0, i.e., $H_0 : R_\phi^h(\theta_1, \theta_2) = R_0$, we can use the statistic

$$Z' = \left(\frac{mn}{n+m}\right)^{1/2} \left(\frac{R_\phi^h(\tilde{\theta_1}, \tilde{\theta_2}) - R_0}{\tilde{\sigma}}\right),$$

which has approximately a standard normal distribution under H_0 for sufficiently large n and m , and $\hat{\sigma}$ is obtained from Theorem 4.3 by replacing λ by $\frac{m}{n+m}$ and θ_1 and θ_2 by their maximum likelihood estimators $\tilde{\theta}_1$ and $\tilde{\theta}_2$ in $\lambda T^t(A^tA)^{-1}T+$ $(1-\lambda)S^t(B^tB)^{-1}$S.

(Received September 15, 1992.)

REFERENCES

[1] M. W. Birch: A new proof of the Pearson–Fisher Theorem. Ann. Math. Statist. *35* (1964), 817–824.

[2] J. Burbea and C. R. Rao: Entropy differential metric, distance and divergence measures in probability spaces: An unified approach. J. Multivariate Anal. *12* (1982), 575–596.

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

[4] I. Csiszár: Generalized entropy and quantization problem. In: Trans. of the Sixth Prague Conference, Academia, Prague 1973, pp. 159–174.

[5] S. G. Ghurye and B. Johnson: Discrete approximations to the information integral. Canad. J. Statist. *9* (1981), 27–37.

[6] D. Morales, L. Pardo, M. Salicrú and M. L. Menéndez: Information measures associated to R-divergences. In: Multivariate analysis: Future directions 2. (C. M. Cuadras and C. R. Rao, eds.) Elsevier Science Publishers, B. V. 1982.

[7] M. Salicrú, M. L. Menéndez, L. Pardo and D. Morales: Asymptotic distribution of (h, ϕ)-entropies. Comm. Statist. A – Theory Methods (to appear).

[8] I. J. Taneja: On generalized information measures and their applications. Adv. Elect. and Elect. Phys. *76* (1989), 327–413.

[9] I. Vajda and K. Vašek: Majorization, concave entropies and comparison of experiments. Problems Control Inform. Theory *14* (1985), 105–115.

[10] K. Zografos, K. Ferentinos and T. Papaioannou: Discrete approximations to the Csiszár, Rényi, and Fisher measures of information. Canadian J. Statist. *14* (1986), 4, 355–366.

Professor L. Pardo and Professor D. Morales, Departamento de Estadística e I. O., Facultad de Matemáticas, Universidad Complutense de Madrid, 28040-Madrid. Spain.

Professor K. Ferentinos and Professor K. Zografos, Department of Mathematics, Probability-Statistics and O. R. Unit, University of Ioannina, 45110-Ioannina. Greece.