Estimation of entropies and divergences via nearest neighbors
Nikolai Leonenko, Luc Pronzato, Vippal Savani

To cite this version:

HAL Id: hal-00322783
https://hal.archives-ouvertes.fr/hal-00322783
Submitted on 18 Sep 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
ESTIMATION OF ENTROPIES AND DIVERGENCES VIA NEAREST NEIGHBORS

NIKOLAI LEONENKO — LUC PRONZATO — VIPPAL SAVANI

ABSTRACT. We extend the results in [L. F. Kozachenko, N. N. Leonenko: On statistical estimation of entropy of random vector, Problems Inform. Transmission 23 (1987), 95–101; Translated from Problemy Peredachi Informatsii 23 (1987), 9–16 (in Russian)] and [M. N. Goria, N. N. Leonenko, V. V. Mergel, P. L. Novi Inverardi: A new class of random vector entropy estimators and its applications in testing statistical hypotheses, J. Nonparametr. Statist. 17 (2005), 277–297] and show how kth nearest-neighbor distances in a sample of N i.i.d. vectors distributed with the probability density f can be used to estimate consistently Rényi and Tsallis entropies of the unknown f under minimal assumptions. The method is extended to the estimation of statistical distances between two distributions in the case when one i.i.d. sample from each is available.

1. Introduction

Let \( X \in \mathbb{R}^m \) be a random vector with probability measure \( \mu \) having the density \( f \) with respect to the Lebesgue measure \( \mu_L \). The Rényi entropy [24] of \( f \) is defined by

\[
H_q^* = \frac{1}{1-q} \log \int_{\mathbb{R}^m} f^q(x) \, dx, \quad q \neq 1, \tag{1}
\]


Keywords: entropy estimation, estimation of statistical distance, estimation of divergence, nearest-neighbor distances, Rényi entropy, Havrda-Charvát entropy, Tsallis entropy.

The first author gratefully acknowledges financial support from EPSRC grant RCMT 119. The work of the second author was partially supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors’s view.
and the H a v r d a—C h a r v á t entropy [8] (also called T s a l l i s entropy [25]) by

\[
H_q = \frac{1}{q-1} \left( 1 - \int_{\mathbb{R}^m} f^q(x) \, dx \right), \quad q \neq 1. \tag{2}
\]

When \( q \) tends to 1, both \( H_q \) and \( H_q^* \) tend to the Shannon entropy

\[
H_1 = -\int_{\mathbb{R}^m} f(x) \log f(x) \, dx. \tag{3}
\]

We consider the estimation of \( H_q^* \) and \( H_q \) from a sample of \( N \) independent and identically distributed (i.i.d.) random variables \( X_1, \ldots, X_N, N \geq 2 \), extending the approach proposed by K o z a c h e n k o and L e o n e n k o, see [12], [7], for the estimation of \( H_1 \). The method is based on nearest-neighbor distances in the sample (when \( m = 1 \), it is thus related to sample-spacing methods; see, e.g., [29] for Shannon entropy and [2] for a survey on entropy estimation). It is connected with the random-graph approach of R e d m o n d and Y u k i c h [23] who, supposing that the distribution is supported on \([0, 1]^m\) and with some smoothness assumptions on \( f \), construct a strongly consistent estimator of \( H_q^* \) for \( 0 < q < 1 \) (up to an unknown bias term independent of \( f \), related to the graph properties). For \( q \neq 1 \) our construction relies on the estimation of the integral

\[
I_q = I_q(f) = \mathbb{E}\{f^{q-1}(X)\} = \int_{\mathbb{R}^m} f^q(x) \, dx \tag{4}
\]

through the computation of conditional moments of nearest-neighbor distances (\( \mathbb{E} \) will always denote the expectation for \( f \)). It thus possesses some similarities with [6] where the asymptotic behavior of the moments of \( k \)th nearest-neighbor distances is considered: under the conditions that \( f \) is continuous, \( f > 0 \) on a compact convex subset \( C \) of \( \mathbb{R}^m \), with \( f \) having bounded partial derivatives on \( C \), the weak consistency of the estimator of \( I_q \) is established, \( N \to \infty \), for \( m \geq 2 \) and \( q < 1 \). Comparatively, our results cover a larger range of values for \( q \) and do not rely on regularity or bounded support assumptions for \( f \). The results for the estimation of the Shannon entropy (3) are derived from those obtained for \( q \neq 1 \).

The method can also be applied to the estimation of statistical distances. Here we only consider the Kullback-Leibler relative entropy, defined by

\[
K(f, g) = \int_{\mathbb{R}^m} f(x) \log \frac{f(x)}{g(x)} \, dx = \tilde{H}_1 - H_1, \tag{5}
\]
where $H_1$ is given by (3) and
\[ \tilde{H}_1 = -\int_{\mathbb{R}^m} f(x) \log g(x) \, dx. \] (6)

The estimation of $H_1$ and $\tilde{H}_1$ is then based on $N$ independent observations $X_1, \ldots, X_N$ distributed with the density $f$ and $M$ observations $Y_1, \ldots, Y_M$ distributed with $g$. Estimation of other statistical distances is considered in [14].

Section 2 gives some properties of $I_q$, $H^*_q$ and $H_q$ and lists some applications of entropy estimation. The main results of the paper are presented in Section 3, where the nearest-neighbor estimators of the quantities (1–5) are defined and their asymptotic properties are summarized.

2. Some properties of $I_q$, $H_q$, $H^*_q$ and applications

One may notice that $H^*_q$ can be expressed as a function of $H_q$, $H^*_q = \log[1 - (q - 1)H_q] / (1 - q)$, with $d(H^*_q) / d(H_q) = 1/I_q$ and $d^2(H^*_q) / d(H_q)^2 = (q - 1)/I^2_q$ for any $q$. $H^*_q$ is thus a strictly increasing concave (resp. convex) function of $H_q$ for $q < 1$ (resp. $q > 1$). A distribution that maximizes $H^*_q$ therefore also maximizes $H_q$ and will be called $q$-entropy maximizing. The entropy $H_q$ is a concave (resp. convex) function of the density for $q > 0$ (resp. $q < 0$). Hence, $q$-entropy maximizing distributions, under some specific constraints, are uniquely defined for $q > 0$. For instance, when the constraint is that the distribution is finitely supported, then the $q$-entropy maximizing distribution is uniform. Also, for any dimension $m \geq 1$ the $q$-entropy maximizing distribution with a given covariance matrix is of the multidimensional Student-$t$ type if $m/(m + 2) < q < 1$ and has a finite support if $q > 1$, see [30]. This generalizes the well-known property that Shannon entropy $H_1$ is maximized for the normal distribution. Such entropy-maximization properties can be used to derive nonparametric statistical tests, following the same approach as in [29] where normality is tested with $H_1$; see also [7]. The $q$-entropy maximizing property of the Student distribution can be used to test that a given sample is Student distributed, which finds applications in financial mathematics, see [11]. The entropy (2) is of interest in the study of nonlinear Fokker-Planck equations, see [26]. Values of $q \in [1, 3]$ are used in [1] to study the behavior of fractal random walks. Applications for quantizer design, characterization of time-frequency distributions, image registration and indexing, texture classification, image matching etc., are considered in [10], [9], [20]. Entropy minimization is used in [22], [32] for parameter estimation in semi-parametric models. Entropy estimation is a basic tool for independent component analysis in signal processing, see, e.g., [18]. The Kullback-Leibler relative
entropy (5) can be used to construct a measure of mutual information (MI) between statistical distributions, with applications in image [31], [20] and signal processing [18].

3. The estimators and their properties

3.1. The estimators

Suppose that $X_1, \ldots, X_N$, $N \geq 2$, are i.i.d. with a probability measure $\mu$ having a density $f$ with respect to the Lebesgue measure. Let $\rho(x, y)$ denote the Euclidean distance between two points $x, y$ of $\mathbb{R}^m$. For a given sample $X_1, \ldots, X_N$, and a given $X_i$ in the sample, from the $N - 1$ distances $\rho(X_i, X_j)$, $j = 1, \ldots, N$, $j \neq i$, we form the order statistics $\rho_{1,N-1}^{(i)} \leq \rho_{2,N-1}^{(i)} \leq \cdots \leq \rho_{N-1,N-1}^{(i)}$, so that $\rho_{k,N-1}^{(i)}$ is the $k$th nearest-neighbor distance from $X_i$ to some other $X_j$ in the sample, $j \neq i$. We estimate $I_q$ (4) for $q \neq 1$, by

$$\hat{I}_{N,k,q} = \frac{1}{N} \sum_{i=1}^{N} (\zeta_{N,i,k})^{1-q},$$

with $\zeta_{N,i,k} = (N - 1) C_k V_m \left( \rho_{k,N-1}^{(i)} \right)^m$, where $V_m = \pi^{m/2}/\Gamma(m/2 + 1)$ is the volume of the unit ball $B(0, 1)$ in $\mathbb{R}^m$ and $C_k = \left[ \Gamma(k)/\Gamma(k + 1 - q) \right]^{1/(1-q)}$. Then we estimate $H^*_q$ (1) and $H_q$ (2) respectively by

$$\hat{H}^*_{N,k,q} = \log(\hat{I}_{N,k,q})/(1 - q),$$
$$\hat{H}_{N,k,q} = (1 - \hat{I}_{N,k,q})/(q - 1).$$

For the estimation of $H_1$ (3) we take the limit of $\hat{H}_{N,k,q}$ as $q \to 1$, which gives

$$\hat{H}_{N,k,1} = \frac{1}{N} \sum_{i=1}^{N} \log \xi_{N,i,k},$$

with $\xi_{N,i,k} = (N - 1) \exp\left[ -\Psi(k) \right] V_m \left( \rho_{k,N-1}^{(i)} \right)^m$, where $\Psi(z) = \Gamma'(z)/\Gamma(z)$ is the digamma function.

Suppose now that $X_1, \ldots, X_N$ are i.i.d. with the density $f$ and that $Y_1, \ldots, Y_M$ are i.i.d. with the density $g$. For any $X_i$ in the sample, $i \in \{1, \ldots, N\}$, consider $\hat{\rho}(X_i, Y_j)$, $j = 1, \ldots, M$, and form the order statistics $\hat{\rho}_{1,M}^{(i)} \leq \hat{\rho}_{2,M}^{(i)} \leq \cdots \leq \hat{\rho}_{M,M}^{(i)}$, so that $\hat{\rho}_{k,M}^{(i)}$ is the $k$th nearest-neighbor distance from $X_i$ to some $Y_j$, ...
ESTIMATION OF ENTROPIES AND DIVERGENCES VIA NEAREST NEIGHBORS

\( j \in \{1, \ldots, M\} \). Then we estimate \( \hat{H}_1 \) (6) and \( K(f, g) \) (5) respectively by

\[
\hat{H}_{N,M,k} = \frac{1}{N} \sum_{i=1}^{N} \log \left\{ M \exp \left[ -\Psi(k) \right] V_m \left( \hat{\rho}_{k,M}^{(i)} \right)^m \right\}, 
\]

\[
\hat{K}_{N,M,k} = \hat{H}_{N,M,k} - \hat{H}_{N,k,1} = m \log \left[ \prod_{i=1}^{N} \frac{\hat{\rho}_{k,M}^{(i)}}{\hat{\rho}_{k,N}^{(i)}} \right]^{1/N} + \log \frac{M}{N - 1}. 
\]

### 3.2. Asymptotic properties

The properties of the estimators \( \hat{I}_{N,k,q} \) (7) are summarized in Table 1, from which one can deduce those of \( \hat{H}_{N,k,q} \) (8) and \( \hat{H}_{N,k,q} \) (9). Table 2 gives the properties of the estimators \( \hat{H}_{N,k,1} \) (10) and \( \hat{H}_{N,M,k} \) (11), from which those of the estimator \( \hat{K}_{N,M,k} \) (12) can be read directly. As indicated, \( L_2 \) (and thus weak) consistency is obtained without any smoothness assumption on the underlying density \( f \) (or densities \( f \) and \( g \)) or any bounded-support assumption, which improves the results of existing methods, see [2]. The proofs are rather technical and are omitted due to space limitation, see [14]. They rely on an application of Lebesgue’s bounded convergence theorem, on Theorem 2.5.1 of [4], p. 34, on the generalized Helly-Bray Lemma, see [16], p. 187 and on the following.

**Lemma 1** (Lebesgue, [13]). If \( g \in L_1(\mathbb{R}^m) \), then for any sequence of open balls \( B(x, R_k) \) of radius tending to zero as \( k \to \infty \) and for \( \mu_L \)-almost any \( x \in \mathbb{R}^m \),

\[
\lim_{k \to \infty} \left[ \frac{1}{(V_m R_k^m)} \right] \int_{B(x, R_k)} g(t) \, dt = g(x).
\]

**Table 1.** Asymptotic properties of the estimator (7) as \( N \to \infty \).

<table>
<thead>
<tr>
<th>( q )</th>
<th>assumption on ( f )</th>
<th>( \hat{I}_{N,k,q} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q &lt; 1 )</td>
<td>( \mathbb{E}{f^{q-1}(X)} &lt; \infty )</td>
<td>asympt. unbiased</td>
</tr>
<tr>
<td>( q &lt; 1 )</td>
<td>( \mathbb{E}{f^{2(q-1)}(X)} &lt; \infty )</td>
<td>( L_2 )-consistent</td>
</tr>
<tr>
<td>( 1 &lt; q &lt; k+1 )</td>
<td>( f ) bounded</td>
<td>asympt. unbiased</td>
</tr>
<tr>
<td>( 1 &lt; q &lt; (k+1)/2 )</td>
<td>( f ) bounded, ( k \geq 2 )</td>
<td>( L_2 )-consistent</td>
</tr>
<tr>
<td>( 1 &lt; q &lt; 3/2 )</td>
<td>( f ) bounded, ( k = 1 )</td>
<td>( L_2 )-consistent</td>
</tr>
</tbody>
</table>

**Table 2.** Asymptotic properties of the estimators (10) and (11).

<table>
<thead>
<tr>
<th>assumptions on ( f ) and ( g )</th>
<th>property</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f ) bounded, ( \exists \varepsilon &gt; 0: \mathbb{E}{f^{-\varepsilon}(X)} &lt; \infty )</td>
<td>( \hat{H}_{N,k,1} ) is ( L_2 )-consistent, ( N \to \infty )</td>
</tr>
<tr>
<td>( g ) bounded, ( \exists \varepsilon &gt; 0: \mathbb{E}{g^{-\varepsilon}(X)} &lt; \infty )</td>
<td>( \hat{H}_{N,M,k} ) is ( L_2 )-consistent, ( N,M \to \infty )</td>
</tr>
</tbody>
</table>
EXAMPLE. Figure 1 presents $H_q^*$ as a function of $q$ (solid line) for the three-dimensional ($m = 3$) Student distribution $T(\nu, I_3)$ with zero mean, scaling matrix the identity $I_3$ (and covariance matrix $\nu/(\nu - 2)$ times the identity) and $\nu = 5$ degrees of freedom, the p.d.f. of which is

$$f_{\nu}(x) = \frac{1}{(\nu \pi)^{m/2}} \frac{\Gamma\left(\frac{m+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left[1 + x^\top x/\nu\right]^{(m+\nu)/2}}, \quad x \in \mathbb{R}^m.$$  

The associated Rényi entropy $H_q^*$ is given by

$$H_q^* = \frac{1}{1 - q} \log B\left(\frac{q(m+\nu)}{2}, \frac{m}{2}\right) - \frac{1}{2} \log[(\pi\nu)^{m}] - \log \Gamma\left(\frac{m}{2}\right)$$

with $B(u, v) = \Gamma(u)\Gamma(v)/\Gamma(u + v)$ the Beta function, and is defined for $q > m/(m + \nu) = 3/8 = 0.375$. The estimates $\hat{H}_{N,k,q}^*$ for $k = 1, \ldots, 5$ obtained from a sample of size $N = 1000$ are plotted on the same figure. Note that $\hat{H}_{N,k,q}^*$ is defined only for $q < k + 1$.

Further developments. Here nearest neighbors are defined for the Euclidean distance, but the metric could be adapted to the observed sample. Indeed, for $X_1, \ldots, X_N$ a sample having a non-spherical distribution, its empirical covariance matrix $\hat{\Sigma}_N$ could be used to define a new metric through $\|x\|_{\hat{\Sigma}_N}^2 = x^\top \hat{\Sigma}_N^{-1} x$, the volume $V_m$ of the unit ball in this metric becoming $|\hat{\Sigma}_N|^{1/2} \pi^{m/2} / \Gamma(m/2 + 1)$.  

270
Few results exist concerning $\sqrt{N}$-consistency of entropy estimators. For instance, $\sqrt{N}$-consistency of an estimator of $H_1$ based on nearest-neighbor distances ($k = 1$) is proved in [27] for $m = 1$ and sufficiently regular densities $f$ with unbounded support. Concerning the method proposed here, $\sqrt{N}$-consistency of the estimator $\hat{I}_{N,k,q}$ is still an open issue. As for the case of spacing methods, where the spacing $m$ can be taken as an increasing function of the sample size $N$, see, e.g., [29], [28], it seems reasonable to let $k = k_N$ increase with $N$. Properties of nearest-neighbor distances with $k_N \to \infty$ are considered for instance in [17], [19], [5], [15].

A central limit theorem for functions $h(\rho)$ of nearest-neighbor distances is obtained in [3] for $k = 1$ and in [21] for $k \geq 1$. However, these results are restricted to the case of bounded functions, which does not cover the situation $h(\rho) = \rho^{m(1-q)}$, see (7), or $h(\rho) = \log(\rho)$, see (10). Conditions for the asymptotic normality of $\hat{I}_{N,k,q}$ are under current investigation.

Acknowledgements. The authors wish to thank Anatoly A. Zhigljavsky from Cardiff School of Mathematics for helpful discussions.

REFERENCES


ESTIMATION OF ENTROPIES AND DIVERGENCES VIA NEAREST NEIGHBORS


Received September 20, 2006

Nikolai Leonenko
Vippal Savani
Cardiff University
Cardiff School of Mathematics
Senghennydd Road
Cardiff CF24 4AG
UNITED KINGDOM
E-mail: leonenkon@cardiff.ac.uk
savaniv@cardiff.ac.uk

Luc Pronzato
Laboratoire I3S, UNSA-CNRS
2000 Route des Lucioles, BP.121
06903 Sophia Antipolis - Cedex
FRANCE
E-mail: pronzato@i3s.unice.fr