Literature Review

Karina Marks

February 25, 2017

1 Pre N.N-Estimator

1.1 1992 - Entropy Optimization Principles with Applications (J.N.Kapur and H.K.Kesavan) 8

This gives an overview of entropy, some principles on the optimisation of entropy, the interrelationships among these principles and applications of entropy and these principles.

This paper defines entropy as the probabilistic uncertainty - the uncertainty associated with the probability of outcomes.

Not useful for estimation of entropy, but shows the applications of exact entropies, the entropy optimisation principles considered are;

-

1.2 1998 - Limit Theorems for Non-Parametric Sample Entropy Estimators (K-S.Song) 6

Vasicek's sample entropy estimator conditions, theorems and proofs of this

2 N.N Entropy Estimator

2.1 2004 - Problems of Information Transmission - On statistical estimation of entropy of random vector (N.Leonenko) 1

This paper is a primary view at the estimator and at how conditions for asymptotic unbiasedness and consistency of the estimator can be established. Here we are looking at estimating the entropy of an absolutely continuous random sample of independent observations, with unknown probability density $f(x), x \in \mathbb{R}^d$, given by;

$$H = -\int_{\mathbb{R}^d} f(x) \log f(x) dx < \infty \tag{1}$$

As $f(x)$ is unknown this is not easily estimated for a random sample, and by just estimating the density $\hat{f}(x)$ to replace the actual density $f(x)$ in the formula for the entropy we get highly restrictive consistency conditions. Thus the following simple estimator for entropy was proposed;

$$H_N = d\log(\bar{\rho}) + \log(c(d)) + \log(\gamma) + \log(N-1) \tag{2}$$

where the sample $X_1, X_2, ..., X_N$, $N \geq 2$ is taken from the space $\mathbb{R}^d, d \geq 1$ and we have the following defined;

- Metric $\rho(x_1, x_2) = \left[\sum_{j=1}^{d}(x_1^{(j)} - x_2^{(j)})^2\right]^{\frac{1}{2}}$, where $x = (x^{(1)}, x^{(2)}, ..., x^{(d)}) \in \mathbb{R}^d$
- Volume of the d-dimensional unit ball, $v(y,r) = \{x \in \mathbb{R}^d : \rho(x,y) < r\}$, is given by $c(d) = |v(y,r)| = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$
- $\bar{\rho} = \left[\prod_{i=1}^{N}\rho_i\right]^{\frac{1}{N}}$, where $\rho_i = min(\rho(X_i, X_j) : j \in \{1,2,...,N\}\backslash\{i\})$
- Euler constant $\gamma = \exp\left[-\int_0^\infty e^{-t}\log(t)dt\right]$

Henceforth, here Leonenko establishes some conditions under which the estimator is asymptotically unbiased and consistent. The two main results about this estimator are, as follows;

Theorem 1 *For exact entropy H, Kozachenko-Leonenko estimator $\hat{H}_{N,k}$, and density function $f(x)$, for some $\epsilon > 0$ if both*

$$\int_{\mathbb{R}^d} |log(f(x))|^{1+\epsilon} f(x)dx < \infty \tag{3}$$

and

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |log(\|x-y\|)|^{1+\epsilon} f(x)f(y)dxdy < \infty \tag{4}$$

Then we have that

$$\lim_{N\to\infty} \mathbb{E}(\hat{H}_{N,k}) = H$$

Thus $\hat{H}_{N,k}$ is an asymptotically unbiased estimator of H.

Theorem 2 *For exact entropy H, Kozachenko-Leonenko estimator $\hat{H}_{N,k}$, and density function $f(x)$, for some $\epsilon > 0$ if both*

$$\int_{\mathbb{R}^d} |log(f(x))|^{2+\epsilon} f(x)dx < \infty$$

and

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |log(\|x-y\|)|^{2+\epsilon} f(x)f(y)dxdy < \infty$$

Then $\hat{H}_{N,k}$ for $N \to \infty$ is a consistent estimator of H.

(An estimator is consistent if the probability that it is in error by more than a given amount tends to zero as the sample become large $\Leftrightarrow$ for error $\delta > 0$, we have $\lim_{N\to\infty} \mathbb{P}(|\hat{H}_{N,k} - H| < \delta) = 1$)

In this paper, the estimator is in its simplest form, which is later developed into something more sophisticated, using the nearest neighbour method, where the consistency and asymptotic unbias of the estimator holds under less constrained conditions.

2.1.1 Summary - paper 1

This paper looks at estimating the Shannon entropy of an absolutely continuous random sample of independent observations, with unknown probability density $f(x), x \in \mathbb{R}^d$. As $f(x)$ is unknown this is not easily estimated accurately for a random sample, and by just estimating the density $\hat{f}(x)$ to replace the actual density $f(x)$ in the formula for the entropy we get highly restrictive consistency conditions.

Therefore, the following estimator was proposed for the Shannon entropy of a random sample $X_1, X_2, ..., X_N$ of d-dimensional observations;

$$H_N = d\log(\bar{\rho}) + \log(c(d)) + \log(\gamma) + \log(N-1) \tag{5}$$

where $c(d) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$ is the volume of the d-dimensional unit ball, the Euler constant is $\gamma = \exp\left[-\int_0^\infty e^{-t}\log(t)dt\right]$ and $\bar{\rho} = \left[\prod_{i=1}^N \rho_i\right]^{\frac{1}{N}}$, with ρ_i the nearest neighbour distance from X_i to another member of the sample X_j, $i \neq j$.

Under some strong conditions on the density function, this estimator is asymptotically unbiased and a consistent estimator for the Shannon entropy.

The estimator here is in a simple form, which is later developed into something more sophisticated, using the nearest neighbour method, but considering larger values of k (here $k = 1$). This estimator is developed so that the consistency and asymptotic unbias of the estimator holds under less constrained conditions.

2.2 2006 - Causality Detection Based on Information-Theoretic Approaches in Time Series Analysis (K.H-Schindler, M.Palus, M.Vejmelka, J.Bhattacharya) 7

considers different types of estimators including NN

2.3 2007 - a Class of Renyi Information Estimators for Mulitdimensional densities (N.Leonenko, L.Pronzato, V.Savani) 2

(Shows that entropy can be estimated consistently with minimal assumptions on the density of the distribution, f. Moreover, this can be extended to estimate

the statistical difference between two distributions using one i.i.d sample from each.)

This paper looks at estimating both the Rényi and Tsallis entropies for a random vector $X \in \mathbb{R}^d$ with density function f, defined as;

Rényi entropy

$$H_q^* = \frac{1}{1-q} log(I_q) \qquad (q \neq 1) \tag{6}$$

Tsallis entropy

$$H_q = \frac{1}{q-1}(1 - I_q) \qquad (q \neq 1) \tag{7}$$

where in both the above $I_q = \int_{\mathbb{R}^d} f^q(x)dx$.

When the order of the entropy $q \to 1$, both the Rényi, (12), and Tsallis, (13), entropies tend to the Shannon entropy. The Shannon entropy is a special case for when $q = 1$, defined by;

$$H = -\int_{x:f(x)>0} f(x)log(f(x))dx \tag{8}$$

For $q \neq 1$, the construction of the estimator of entropy relies upon the estimation of the integral I_q, which is given by;

$$\hat{I}_{N,k,q} = \frac{1}{N}\sum_{i=1}^{N}(\zeta_{N,k,q})^{1-q} \tag{9}$$

where we have taken a sample $X_1, X_2, ..., X_N$ with all $X_i \in \mathbb{R}^d$, and k is the size of the nearest neighbour method to be used. We have also defined;

- $\zeta_{N,k,q} = (N-1)C_k V_d(\rho_{k,N-1}^{(i)})^d$
- Volume of d-dimensional unit ball $V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$
- $C_k = \left[\frac{\Gamma(k)}{\Gamma(k+1-q)}\right]^{\frac{1}{1-q}}$
- $\rho_{k,N-1}^{(i)}$ is the kth nearest neighbour distance from the observation X_i to some other X_j

Thus, here it is also shown that the estimator $\hat{I}_{N,k,q}$ for I_q is asymptotically unbiased and consistent, given certain conditions;

Theorem 3 *The estimator $\hat{I}_{N,k,q}$, given above satisfies;*

$$\mathbb{E}(\hat{I}_{N,k,q}) \to I_q, \qquad (N \to \infty) \tag{10}$$

for $q > 1$ provided that I_q exists, and for any $q \in (1, k+1)$ if f is bounded. Thus, $\hat{I}_{N,k,q}$ is an asymptotically unbiased estimator for I_q.

Theorem 4 *The estimator $\hat{I}_{N,k,q}$, given above satisfies;*

$$\hat{I}_{N,k,q} \to^{L_2} I_q, \qquad (N \to \infty) \tag{11}$$

(and thus $\hat{I}_{N,k,q} \to^{p} I_q, (N \to \infty)$) for $q > 1$, provided that I_{2q-1} exists and for any $q \in (1, \frac{k+1}{2})$ when $k \geq 2$ if f is bounded. Thus, $\hat{I}_{N,k,q}$ is a consistent estimator for I_q.

Therefore, under the conditions of Theorem 4 we have the estimators for the following entropies;

Rényi entropy

$$\hat{H}_q^* = \frac{1}{1-q} log(\hat{I}_{N,k,q}) \qquad (q \neq 1) \tag{12}$$

Tsallis entropy

$$\hat{H}_q = \frac{1}{q-1}(1 - \hat{I}_{N,k,q}) \qquad (q \neq 1) \tag{13}$$

which are both consistent estimators of the Rényi and Tsallis entropies of the sample.

Moreover, this paper also goes on to discuss the estimator for the Shannon entropy, given by;

$$\hat{H}_{N,k,1} = \frac{1}{N} \sum_{i=1}^{N} \log(\xi_{N,i,k}) \tag{14}$$

where we have taken a sample $X_1, X_2, ..., X_N$ with all $X_i \in \mathbb{R}^d$, and k is the size of the nearest neighbour method to be used. We have also defined;

- $\xi_{N,i,k} = (N-1)\exp[-\Psi(k)]V_d(\rho_{k,N-1}^{(i)})^d$
- V_d and $\rho_{k,N-1}^{(i)}$ are as defined above
- Digamma function $\Psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$

Then, we also have the following Theorem, for the Shannon entropy;

Theorem 5 *Suppose that f is bounded and that I_{q_1} exists for some $q_1 < 1$. Then H_1 exists and the estimator 14 satisfies $\hat{H}_{N,k,1} \to^{L_2} H_1$ as $N \to \infty$.*

In this paper, we consistently consider a fixed value of k, and assume that q depends upon this k for Theorems 3 and 4 to hold. I wish to examine whether or not a fixed value of k is appropriate for this estimator.

2.3.1 Summary - paper 2

This paper looks at estimating both the Rényi (H_q^*) and Tsallis (H_q) entropies for a random vector $X \in \mathbb{R}^d$ with density function $f(x)$, when $q \neq 1$, by using the kth nearest neighbour method, with a fixed values of k. This is achieved by considering the integral $I_q = \int_{\mathbb{R}^d} f^q(x)dx$, and generating its estimator, which is defined as $\hat{I}_{N,k,q} = \frac{1}{N}\sum_{i=1}^{N}(\zeta_{N,k,q})^{1-q}$. Where, $\zeta_{N,k,q} = (N-1)C_k V_d(\rho_{k,N-1}^{(i)})^d$, $V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$ is the volume of d-dimensional unit ball, $C_k = \left[\frac{\Gamma(k)}{\Gamma(k+1-q)}\right]^{\frac{1}{1-q}}$ and $\rho_{k,N-1}^{(i)}$ is the kth nearest neighbour distance from the observation X_i to some other X_j.

The estimator $\hat{I}_{N,k,q}$, provided $q > 1$ and I_q exists - and for any $q \in (1, k+1)$ if f is bounded - is thus found to be an asymptotically unbiased estimator for I_q. Also, provided $q > 1$ and I_{2q-1} exists - and for any $q \in (1, \frac{k+1}{2})$, when $k \geq 2$ if f is bounded - $\hat{I}_{N,k,q}$ is thus a consistent estimator for I_q. Moreover, by simple formulas both the Rényi and Tsallis entropies can be written in terms of this estimated value;

$$\hat{H}_q^* = \frac{1}{1-q} log(\hat{I}_{N,k,q}) \tag{15}$$

$$\hat{H}_q = \frac{1}{q-1}(1 - \hat{I}_{N,k,q}) \tag{16}$$

thus, under the latter conditions, provide consistent estimates of these entropies as $N \to \infty$.

Furthermore, this paper goes on to discuss an estimator for the Shannon entropy, H_1 by taking the limit of the estimator for the Tsallis entropy, $\hat{H}_{N,k,q}$ as $q \to 1$, again with a fixed value of k. This estimator is given by $\hat{H}_{N,k,1} = \frac{1}{N}\sum_{i=1}^{N}\log(\xi_{N,i,k})$, where $\xi_{N,i,k} = (N-1)\exp[-\Psi(k)]V_d(\rho_{k,N-1}^{(i)})^d$, with V_d and $\rho_{k,N-1}^{(i)}$ defined as in the estimation of I_q and the Digamma function $\Psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$. Under the following conditions; f is bounded, I_{q_1} exists for some $q_1 > 1$; the H_1 exists and the estimator $\hat{H}_{N,k,1}$ is a consistent estimator for the Shannon entropy.

2.4 2009 - Statistical Inference of ϵ -Entropy and the Quadratic Renyi Entropy (N.Leonenko, O.Seleznjev) 5

Show asymptotic properties of the nn-estimator an the quadratic one?

2.5 Feb 2016 - On Kozachenko-Leonenko Entropy Estimator (S.Delattre, N.Fournier) 3

Studies the bias (which is found to be $O(N^{\frac{-2}{d}})$, for dimension d) and variance of the estimator, prove CLT for $d = 1, 2$ and find explicit asymptotic confidence intervals.

2.6 Jul 2016 - Efficient Multivariate Entropy Estimation via k-Nearest Neighbour Distances (T.B.Berrett, R.J.Samworth, M.Yuan) 4

THIS ONE

Under certain conditions (considering k dependent of n) the k-nearest neighbour estimator only works for dimension $d \leq 3$, for a higher dimension gives problems. They introduced a new estimator which is formed as a weighted average of k-nearest neighbour estimators for different values of k.

They also show that the bias of the K-L estimator is dependent on α and β from certain conditions which imply for $d \leq 3$, that the bias is $O\left(max\left\{\frac{k^{\frac{\alpha}{\alpha+d}+\epsilon}}{n^{\frac{\alpha}{\alpha+d}+\epsilon}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}}\right\}\right)$.

From paper 4 (TODO reference), we have the following, conditions which will be used in Theorems 7 and 8;

Condition 1 *(β) For density f bounded, denoting $m := \lfloor \beta \rfloor$ and $\eta := \beta - m$, we have that f is m times continuously differentiable and there exists $r_* > 0$ and a Borel measurable function g_* such that for each $t = 1, 2, ..., m$ and $\|y - x\| \leq r_*$, we have;*

$$\|f^{(t)}(x)\| \leq g_*(x) f(x)$$

,

$$\|f^{(m)}(y) - f^{(m)}(x)\| \leq g_*(x) f(x) \|y - x\|^{\eta}$$

and $sup_{x:f(x)\geq\delta} g_*(x) = o(\delta^{-\epsilon})$ as $\delta \downarrow 0$, for each $\epsilon > 0$.

Condition 2 *(α) For density $f(x)$ and dimension d, we have;*

$$\int_{\mathbb{R}^d} \|x\|^{\alpha} f(x) dx < \infty$$

Condition 3 *Assume that condition 1 holds for $\beta = 2$ and condition 2 holds for some $\alpha > d$. Let $k_0^* = k_{0,N}^*$ and $k_1^* = k_{1,N}^*$ denote two deterministic sequences of positive integers with $k_0^* \leq k_1^*$, with $\frac{k_0^*}{\log^5 N} \to \infty$ and with $k_1^* = O(N^{\tau})$, where*

$$\tau < \min\left\{\frac{2\alpha}{5\alpha + 3d}, \frac{\alpha - d}{2\alpha}, \frac{4}{4 + 3d}\right\}$$

Theorem 6 *Assume that conditions 1 and 2 hold for some $\beta, \alpha > 0$. Let $k^* = k_N^*$ denote a deterministic sequence of positive integers with $k^* = O(N^{1-\epsilon})$ as $N \to \infty$ for some $\epsilon > 0$. Then, for $d \leq 2$ (or $d \geq 3$) with either $\beta \leq 2$ or $\alpha \in (0, \frac{2d}{d-2})$, then for every $\epsilon > 0$ we have;*

$$\mathbb{E}(\hat{H}_N) - H = O\left(max\left\{\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{N^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{\frac{\beta}{d}}}{N^{\frac{\beta}{d}}}\right\}\right) \tag{17}$$

uniformly for $k \in \{a, ..., k^*\}$, as $N \to \infty$.

Theorem 7 *Assume that condition 1 holds for $\beta = 2$ and condition 2 holds for some $\alpha > d$, then by condition 3, for the estimator $\hat{H}_{N,k}$ we have;*

$$Var(\hat{H}_{N,k}) = \frac{Var(\log f(x))}{N} + o(\frac{1}{N})$$

as $N \to \infty$, uniformly for $k \in \{k_0^*, ..., k_1^*\}$.

Theorem 8 *Assume that $d \leq 3$ and the conditions of Theorem 7 are satisfied (where if $d = 3$, we additionally assume $k_1^* = o(n^{\frac{1}{4}})$. Then, for the estimator $\hat{H}_{N,k}$ we have;*

$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{18}$$

and

$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to \sigma^2 \tag{19}$$

as $N \to \infty$ uniformly for $k \in \{k_0^*, ..., k_1^*\}$, where $\sigma^2 = Var(log(f(x))$, for density function $f(x)$.

(This combines Theorem 1 with stronger conditions; hence we can now say that $\hat{H}_{N,k}$ is an consistent and asymptotically unbiased estimator of exact entropy H.)

2.6.1 Summary - paper 4

The last main paper, whose results I will be exploring is *Efficient Multivariate Entropy Estimation via k-Nearest Neighbour Distances* (T.Berrett, R.Samworth, M.Yuan, 2016), which initially studies the K-L estimator, and the conditions under which it is efficient and asymptotically unbiased (for a value of k depending on the sample size N).

Considering dimensions $d \leq 3$, and a sample size N from distribution with density $f(x)$, they defined the k-nearest neighbour estimator of entropy - just as in section **??** - to be;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\rho_{(k),i}^d V_d (N-1)}{e^{\Psi(k)}} \right] \tag{20}$$

where $\rho_{(k),i}$, V_d and $\Psi(k)$ are all defined as in the 2007 paper. However, the difference here is in the conditions under which the estimator is consistent and asymptotically unbiased.

Here, some conditions on the finiteness of the α moment of f and the continuity and differentaibility of f are proposed, with $k \in \{1, ..., O(N^{1-\epsilon})\}$, for some $\epsilon > 0$, we have asymptotic unbias of the estimator; where the bias can be expressed as;

$$\mathbb{E}(\hat{H}_N) - H = O\left(max \left\{ \frac{k^{\frac{\alpha}{\alpha+d} - \epsilon}}{N^{\frac{\alpha}{\alpha+d} - \epsilon}}, \frac{k^{\frac{\beta}{d}}}{N^{\frac{\beta}{d}}} \right\} \right) \qquad N \to \infty \tag{21}$$

Also, they considered the asymptotic normality of the estimator, given the α moment of f is finite (for $\alpha > d$), and some conditions on the continuity and

differentaibility of f hold and with $k \in \{k_0, ..., k_1\}$. Then the variance of the estimator is given by;

$$Var(\hat{H}_{N,k}) = \frac{\sigma^2}{N} + o(\frac{1}{N}) \tag{22}$$

as $N \to \infty$, where $\sigma^2 = Var(log(f(x))$, and we define k_0, k_1 such that $\frac{k_0}{log^5(N)} \to \infty$ and $k_1 = O(N^\tau)$, where $\tau < \min\left\{\frac{2\alpha}{5\alpha+3d}, \frac{\alpha-d}{2\alpha}, \frac{4}{4+3d}\right\}$.

Moreover, T.Berrett, R.Samsworth and M.Yuan also go on to show that a consequence of the variance, given the dimension of the sample $d \leq 3$, with the same conditions, we have the asymptotic normality;

$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{23}$$

and

$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to \sigma^2 \tag{24}$$

where the estimator is asymptotically efficient and the asymptotic variance here is the best possible.

It is important to note that for higher dimensions ($d > 3$), these results do not necessarily hold; since I am just considering the specific dimensions $d = 1$ and $d = 2$, there is no need to detail this. However, they do then go on to discuss a more appropriate estimator for higher dimensions, given sufficient smoothness, which is efficient in arbitrary dimensions, which was previously mentioned in section TODO.

2.7 Aug 2016 - Demystifying Fixed k-Nearest Neighbour Information Estimators (W.Gao, S.Oh, P.Viswanath) 9

Now considering k is independent of n, the kl estimator has bias $O(N^{\frac{-1}{d}})$