1 Kozachenko-Leonenko Estimator

1.1 History

This estimator was first introduced by L.Kozachenko and N.Leonenko, in 1987, where they first published the article *Sample Estimate of the Entropy of a Random Vector*, in the paper *Problems of Information Transmission.* Using the nearest neighbour method, they created a simple estimator for the Shannon entropy of an absolutely continuous random vector from a independent sample of observations, to then establish conditions under which we have asymptotic unbiasedness and consistency.

Since then, there has been major developments in the estimator; firstly in 2007, N.Leonenko, L.Pronzato, V.Savani, proposed a similar alternative to this estimator in their paper *a Class of Renyi Information Estimators for Mulitdimensional densities*, this time using the k-nearest neighbour method, to consider estimators for the Rényi and Tsallis entropies. Then as the order of these entropies $q \to 1$, they defined the k-nearest neighbour estimator for the Shannon entropy, where k is fixed, and these estimators (under less rigorous conditions) are both consistent and asymptotically unbiased.

Also, the use of a fixed k has been backed up by a more recent paper in 2016, by S.Delattre and N.Fournier, *On the Kozachenko-Leonenko Entropy Estimator*, which is a detailed study of the bias and variance of this estimator, using a fixed k. Subsequently finding that, in higher dimensions, the bias can be expressed in terms of $N^{-\frac{2}{d}}$; thus, leading to the development of explicit asymptotic confidence intervals.

Moreover, in 2016, a new idea was proposed by T.Berrett, R.Samsworth and M.Yuan, written in *Efficient Mulitvariate Entropy Estimation via k-Nearest Neighbour Distances*; that the value chosen for k, depends upon the sample size N. Also, this idea is then extended to a new estimator; "formed as a weighted average if Kozachenko-Leonenko estimators for different values of k". I will not be exploring this new estimator in depth; however, the understanding of the value of k depending on N will be examined in detail.

1.1.1 Estimator with k=1

Firstly, I considered an article *On Statistical Estimation of Entropy of Random Vector* (N.Leonenko and L.Kozachenko, 1987), which considers 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;

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

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 $\log(\gamma) = \exp\left[-\int_0^\infty e^{-t}\log(t)dt\right] = -\Psi(1)$ 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, $i \neq j$.

It is important to note that one can write the Euler constant $-\Psi(1) = \log(\exp(-\Psi(1))) = \log(\frac{1}{\exp(\Psi(1))})$, this notation is what is used in the latter papers, so it is useful to introduce it here. $\Psi(x)$ is the Digamma function, and when $x = 1$, this is just the Euler constant. Thus this estimator can be written if the form;

$$
\begin{aligned}
H_N &= \log(\bar{\rho}^d) + \log(c(d)) - \Psi(1) + \log(N-1) \\
&= \log\left(\left[\prod_{i=1}^N \rho_i\right]^{\frac{d}{N}}\right) \log(c(d)(N-1)) + \log\left(\frac{1}{\exp(\Psi(1))}\right) \\
&= \frac{1}{N}\sum_{i=1}^N \log(\rho_i^d) + \log\left(\frac{c(d)(N-1)}{\exp(\Psi(1))}\right) \\
&= \frac{1}{N}\sum_{i=1}^N \log(\rho_i^d) + \frac{1}{N}\sum_{i=1}^N \log\left(\frac{c(d)(N-1)}{\exp(\Psi(1))}\right) \\
&= \frac{1}{N}\sum_{i=1}^N \log\left(\frac{\rho_i^d c(d)(N-1)}{\exp(\Psi(1))}\right) \qquad (2)
\end{aligned}
$$

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.

1.1.2 Estimator with k fixed

The next paper I am exploring on estimation is *a Class of Renyi Information Estimators for Mulitdimensional densities* (N.Leonenko, L.Pronzato, V.Savani, 2007), which looks at estimating the Rényi (H_q^*) and Tsallis (H_q) entropies, when $q \neq 1$, and the Shannon ($\hat{H}_{N,k,1}$) entropy. Where these are taken for a random vector $X \in \mathbb{R}^d$ with density function $f(x)$, by using the kth nearest neighbour method, with a fixed values of k.

For the Rényi and Tsallis entropies, 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 $\rho^{(i)}_{k,N-1}$ 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;

$$\hat{H}^*_q = \frac{1}{1-q} log(\hat{I}_{N,k,q}) \tag{3}$$

$$\hat{H}_q = \frac{1}{q-1}(1 - \hat{I}_{N,k,q}) \tag{4}$$

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 similar to that discussed before by Leonenko in his 2004 paper, equation 2; however, it is now extended from the nearest neighbour to the kth nearest neighbour;

$$\hat{H}_{N,k,1} = \frac{1}{N} \sum_{i=1}^{N} \log(\xi_{N,i,k}) \tag{5}$$

where $\xi_{N,i,k} = (N-1)\exp[-\Psi(k)]V_d(\rho^{(i)}_{k,N-1})^d$, with V_d and $\rho^{(i)}_{k,N-1}$ defined as in the estimation of I_q and the digamma function $\Psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$. The digamma function at $k = 1$ is given by $\Psi(1) = \log(\gamma)$, the Euler constant, which was used for the $k = 1$ version of this estimator. Under the following less restrictive conditions; f is bounded, I_{q_1} exists for some $q_1 > 1$; then H_1 exists and the estimator $\hat{H}_{N,k,1}$ is a consistent estimator for the Shannon entropy.

Extend this to also include paper 3

1.1.3 Estimator with k=k(n)

Paper 4

1.2 Focus of this Paper

I now wish to more explicitly introduce the Kozachenko-Leonenko estimator of the entropy H. Let $X_1, X_2, ..., X_N$, $N \geq 1$ be independent and identically distributed random vectors in $\mathbb{R}^d$, and denote $\|.\|$ the Euclidean norm on $\mathbb{R}^d$.

- For $i = 1, 2, ..., N$, let $X_{(1),i}, X_{(2),i}, .., X_{(N-1),i}$ denote an order of the X_k for $k = \{1, 2, ..., N\} \setminus \{i\}$, such that $\|X_{(1),i} - X_i\| \leq \cdots \leq \|X_{(N-1),i} - X_i\|$.

Let the metric ρ, defined as;

$$\rho_{(k),i} = \|X_{(k),i} - X_i\| \tag{6}$$

denote the kth nearest neighbour or X_i.

- For dimension d, the volume of the unit d-dimensional Euclidean ball is defined as;

$$V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(1+\frac{d}{2})} \tag{7}$$

- For the kth nearest neighbour, the digamma function is defined as;

$$\Psi(k) = -\gamma + \sum_{j=1}^{k-1} \frac{1}{j} \tag{8}$$

where $\gamma = 0.577216$ is the Euler-Mascheroni constant (where the digamma function is chosen so that $\frac{e^{\Psi(k)}}{k} \to 1$ as $k \to \infty$).

Then the Kozachenko-Leonenko estimator for entropy, H, is given by;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\rho^d_{(k),i} V_d (N-1)}{e^{\Psi(k)}} \right] \tag{9}$$

where, $\rho^d_{(k),i}$ is defined in (6), V_d is defined in (7) and $\Psi(k)$ is defined in (8).This estimator for entropy, when $d \leq 3$, under a wide range of k and some regularity conditions, satisfies some theorems.

Theorem **??** holds, according to the central limit theorem, on the estimator for entropy $\hat{H}_{N,k}$;

$$\frac{\hat{H}_{N,k} - \mathbb{E}\hat{H}_{N,k}}{\sqrt{Var(\hat{H}_{N,k})}} \xrightarrow{d} N(0, \sigma^2)$$

By Theorem **??**, we can assume that $Var(\hat{H}_{N,k}) = \frac{Var(\log f(x))}{N} \approx \frac{1}{N}$, as for large N, the variance of the logarithm of the density function stays constant. Thus, the left side of the central limit theorem above can be written as;

$$\begin{aligned} \frac{\hat{H}_{N,k} - \mathbb{E}\hat{H}_{N,k}}{\sqrt{Var(\hat{H}_{N,k})}} &= \sqrt{N}(\hat{H}_{N,k} - \mathbb{E}\hat{H}_{N,k}) \\ &= \sqrt{N}[(\hat{H}_{N,k} - H) + (H - \mathbb{E}\hat{H}_{N,k})] \\ &= \sqrt{N}(\hat{H}_{N,k} - H) + \sqrt{N}(H - \mathbb{E}\hat{H}_{N,k}) \end{aligned}$$

and as $N \to \infty$ this tends to the normal distribution, $N(0, \sigma^2)$. So we can say that $\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2)$ while $\sqrt{N}(H - \mathbb{E}\hat{H}_{N,k}) \to \sigma^2$, which is equivalent to the properties stated in Theorem **??**.

Later, I will further discuss this estimator for the specific dimensions $d = 1$ and $d = 2$; however, it is important to note that for larger dimensions this estimator is not accurate. When $d = 4$, equations (**??**) and (**??**) no longer hold but the estimator $\hat{H}_{N,k}$, defined by (9), is still root-N consistent, provided k is bounded. Also, when $d \geq 5$ there is a non trivial bias, regardless of the choice of k. There is a new proposed estimator, formed as a weighted average of $\hat{H}_{N,k}$ for different values of k, where k depends on the choice of N, explored in PAPER 4 (TODO reference).

Moreover, this paper focuses only on distributions for $d \leq 3$, more specifically, I will first be considering samples from 1-dimensional distributions, $d = 1$. Therefore, the volume of the 1-dimensional Euclidean ball is given by $V_1 = \frac{\pi^{\frac{1}{2}}}{\Gamma(\frac{3}{2})} = \frac{\sqrt{\pi}}{\frac{\sqrt{\pi}}{2}} = 2$. Hence the Kozachenko-Leonenko estimator is of the form;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{2\rho_{(k),i}(N-1)}{e^{\Psi(k)}} \right] \tag{10}$$

Later, I will be considering samples from 2-dimensional distributions; thus, $d = 2$ and the volume of the 2-dimensional Euclidean ball is given by $V_2 = \frac{\pi^{\frac{2}{2}}}{\Gamma(2)} = \frac{\pi}{1} = \pi$. Hence, the estimator takes the form;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\pi\rho^2_{(k),i}(N-1)}{e^{\Psi(k)}} \right] \tag{11}$$