Chapter 3 - Estimation of Entropy

Karina Marks

February 26, 2017

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.

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 of 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. Additionally, for $d = 1, 2, 3$, under some conditions on k, we are shown that the bias of the estimator acts in terms of $N^{-\frac{2}{d}}$; something which will also later be explored.

Lastly, also in 2016, S.Delattre and N.Fournier wrote the paper; *On the Kozachenko-Leonenko Entropy Estimator*, where they studied in detail the bias and variance of this estimator considering all 3 proposed values of k - $k = 1$, k fixed or k depends on N. The also provided a development for the bias of this estimator when $k = 1$, in dimensions $d = 1, 2, 3$, in terms of $O(N^{-\frac{1}{2}})$, and in higher dimensions, in terms of powers of $N^{\frac{-2}{d}}$. This is an idea that will be

considered in the focus of this paper; for $d = 1, 2$ to show how the bias acts for large N when $k = 1$.

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 negative Euler constant. Thus this estimator can be written if the form;

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

Under some conditions on the density function, this estimator is asymptotically unbiased and under stronger conditions it is also 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 Multidimensional 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 $\rho_{k,N-1}^{(i)}$ 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;

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

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

thus, under the latter conditions, provide consistent estimates of these entropies as $N \to \infty$ for $q > 1$.

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 proposed in 1987, 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 $\xi_{N,i,k} = (N-1)\exp[-\Psi(k)]V_d(\rho_{k,N-1}^{(i)})^d$, with V_d and $\rho_{k,N-1}^{(i)}$ defined as in the estimation of I_q and the digamma function $\Psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$. 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 and 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. This means that for large N, we have $\hat{H}_{N,k,1} \xrightarrow{L_2} H$; which implies that as $N \to \infty$, both $N^{\frac{1}{2}}(\hat{H}_{N,k,1} - H) \xrightarrow{d} N(0, \sigma^2)$ - it's asymptotically efficient - and $\mathbb{E}(\hat{H}_{N,k,1}) \to H$ - it's asymptotically unbiased.

1.1.3 Estimator with k dependent on N

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 1.1.2 - to be;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\rho_{(k),i}^{d} V_d (N-1)}{e^{\Psi(k)}} \right] \tag{6}$$

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;

$$\mathbb{E}(\hat{H}_N) - H = O\left(max \left\{ \frac{k^{\frac{\alpha}{\alpha+d} - \epsilon}}{N^{\frac{\alpha}{\alpha+d} - \epsilon}}, \frac{k^{\frac{\beta}{d}}}{N^{\frac{\beta}{d}}} \right\} \right) \qquad N \to \infty \tag{7}$$

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;

$$Var(\hat{H}_{N,k}) = \frac{\sigma^2}{N} + o(\frac{1}{N}) \tag{8}$$

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;

$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{9}$$

and

$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to \sigma^2 \tag{10}$$

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.

1.2 Focus of this Paper

I now wish to more explicitly introduce the Kozachenko-Leonenko estimator of the entropy H, in the form that I will be considering. 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;

  $$\rho_{(k),i} = \|X_{(k),i} - X_i\| \tag{11}$$

  denote the kth nearest neighbour or X_i.

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

  $$V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(1 + \frac{d}{2})} \tag{12}$$

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

  $$\Psi(k) = -\gamma + \sum_{j=1}^{k-1} \frac{1}{j} \tag{13}$$

  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;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\rho_{(k),i}^d V_d (N-1)}{e^{\Psi(k)}} \right] \tag{14}$$

where, $\rho_{(k),i}^d$ is defined in (11), V_d is defined in (12) and $\Psi(k)$ is defined in (13).

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;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{2\rho_{(k),i}(N-1)}{e^{\Psi(k)}}\right] \tag{15}$$

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;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\pi\rho^2_{(k),i}(N-1)}{e^{\Psi(k)}}\right] \tag{16}$$

I will be looking at the asymptotic bias and variance of the estimator for different values of k, the main theorems I will be working by are those from section 1.1.3, where we have the conditions 1, 2 and 3, which imply the results stated by Theorems 1 and 2.

(NB: these conditions and theorems have been tweaked slightly to only explicitly consider distributions of dimension $d = 1, 2$, since the only distributions being considered in this paper are of dimension 1 or 2)

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 1 (Asymptotic Unbiasedness) *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 $\beta \leq 2$ (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 $k \in \{1, ..., k^*\}$, as $N \to \infty$.

Theorem 2 (Efficiency and Consistency) *Assume that $d \leq 3$ and that condition 1 holds for $\beta = 2$ and condition 2 holds for some $\alpha > d$, then by condition 3 (where extra assumptions are made for $d = 3$), 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 $f(x)$. Thus, the estimator is asymptotically efficient and its asymptotic variance is the best attainable.

By the above, we can now say that $\hat{H}_{N,k}$ is an consistent and asymptotically unbiased estimator of exact entropy H; thus is a consistent estimator. This is due to using the central limit theorem, on the estimator for entropy $\hat{H}_{N,k}$, which states that;

$$\frac{\hat{H}_{N,k} - \mathbb{E}\hat{H}_{N,k}}{\sqrt{Var(\hat{H}_{N,k})}} \xrightarrow{d} N(0,1)$$

By section 1.1.3, we can assume that $Var(\hat{H}_{N,k}) = \frac{Var(\log f(x))}{N} + O(\frac{1}{N}) \approx \frac{\sigma^2}{N}$. Accordingly, the left side of the central limit theorem above can be written as;

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

So we can see that from Theorem 2; $\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2)$ as $N \to \infty$. Whilst from Theorem 1 we have $\mathbb{E}\hat{H}_{N,k} - H \to 0$ as $N \to \infty$. Thus as $N \to \infty$ this tends to the standard normal distribution, $N(0,1)$, and the central limit theorem holds.

I will be exploring the bias is more detail later to see which one of the two ideas show to be more true for the behaviors of the bias for a large value of N, in dimension $d = 1$ or $d = 2$.

- With a fixed k, by [**?**], for $\beta \in (0, 2] \cap (0, d]$, we choose $a \in (0, \frac{\beta}{d}]$, then;

$$|Bias(\hat{H}_{N,k})| = O\left(\frac{1}{N^a}\right) \tag{20}$$

- With k depending on N, by [**?**], for $\beta \in (0, 2]$, we again choose $a \in (0, \frac{\beta}{d}]$, then;

$$|Bias(\hat{H}_{N,k})| = O\left(\left(\frac{k}{N}\right)^a\right) \tag{21}$$