1 Literature Review

1.1 Applications of Entropy

1.2 Estimation of Entropy

1.2.1 Estimator with k=1

Firstly, I considered an article *On Statistical Estimation of Entropy of Random Vector* (N.Leonenko, 2004), 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;

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

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.

1.2.2 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 $\rho_{k,N-1}^{(i)}$ 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 $N \to \infty$.

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 $\xi_{N,i,k} = (N-1)\exp[-\Psi(k)]V_d(\rho_{k,N-1}^{(i)})^d$, 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, 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.

1.2.3 Estimator with k=k(n)