1 Estimation of Entropy

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;

$$\rho_{(k),i} = \|X_{(k),i} - X_i\| \tag{1}$$

  denote the kth nearest neighbour or X_i.

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

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

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

$$\Psi(k) = -\gamma + \sum_{j=1}^{k-1} \frac{1}{j} \tag{3}$$

  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;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\rho_{(k),i}^d V_d (N-1)}{e^{\Psi(k)}} \right] \tag{4}$$

where, $\rho_{(k),i}^d$ is defined in (1), V_d is defined in (2) and $\Psi(k)$ is defined in (3).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 (4), 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;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{2\rho_{(k),i}(N-1)}{e^{\Psi(k)}} \right] \tag{5}$$

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;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\pi\rho^2_{(k),i}(N-1)}{e^{\Psi(k)}} \right] \tag{6}$$