# Statistical Inference for Entropy

Karina Marks

November 1, 2016

## 1 Introduction

## 2 Entropies and Properties

Entropy can be thought of as a representation of the average information content of an observation; sometimes referred to as a measure of unpredictability or disorder.

### 2.1 Shannon Entropy

The Shannon entropy of a random vector X with density function f is given by;

$$H = -\mathbb{E}\{log(f(x))\} = -\int_{x:f(x)>0} f(x)log(f(x))dx \tag{1}$$

### 2.2 Rényi and Tsallis Entropy

These entropies are for the order $q \neq 1$ and the construction of them relies upon the generalisation of the Shannon entropy 1. For a random vector $X \in \mathbb{R}^d$ with density function f, we define;

Rényi entropy

$$H_q^* = \frac{1}{1-q} log\left(\int_{\mathbb{R}^d} f^q(x)dx\right) \qquad (q \neq 1) \tag{2}$$

Tsallis entropy

$$H_q = \frac{1}{q-1}\left(1 - \int_{\mathbb{R}^d} f^q(x)dx\right) \qquad (q \neq 1) \tag{3}$$

When the order of the entropy $q \to 1$, both the Renyi, (2), and Tsallis, (3), entropies tend to the Shannon entropy, (1).

# 3 Estimation of Entropy

## 3.1 Kozachenko-Leonenko Estimator

We now wish to 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 $\rho$, defined as;

$$\rho_{(k),i} = \|X_{(k),i} - X_i\| \tag{4}$$

  denote the kth nearest neighbour or $X_i$.

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

$$V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(1 + \frac{d}{2})} \tag{5}$$

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

$$\Psi(k) = -\gamma + \sum_{j=1}^{k-1} \frac{1}{j} \tag{6}$$

  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;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[ \frac{\rho_{(k),i}^d V_d (N-1)}{e^{\Psi(k)}} \right] \tag{7}$$

This estimator for entropy, when $d \leq 3$, under a wide range of k and some regularity conditions, satisfies;

$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to 0 \quad (N \to \infty) \tag{8}$$

so $\hat{H}_{N,k}$ is efficient in the sense that the asymptotic variance is the best attainable; $N^{\frac{1}{2}}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, Var[log(f(x))])$, the normal distribution with 0 mean and variance as shown.

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 $d = 4$, equation( 8) no longer holds but the estimator $\hat{H}_{N,k}$, defined by (7), 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, explored in ...SOMEONE... . Moreover, this will not be examined here as this paper focuses only on the 1-dimensional samples;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\rho_{(k),i}V_1(N-1)}{e^{\Psi(k)}}\right] \tag{9}$$

and the 2-dimensional;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\rho^2_{(k),i}V_2(N-1)}{e^{\Psi(k)}}\right] \tag{10}$$

## 3.2 Bias of the K-L estimator

$\hat{H}_{N,k}$ is approximately an unbiased estimator for H; we wish to explore how approximate this is, by considering the bias of the estimator for entropy;

$$Bias(\hat{H}_{N,k}) = \mathbb{E}(\hat{H}_{N,k}) - H = \mathbb{E}(\hat{H}_{N,k} - H) \tag{11}$$

To do this we consider the consistency and asymptotic bias of the estimator $\hat{H}_{N,k}$, ...SOMEONE... has explored this in detail thus the following theorems hold.

DON'T KNOW WHAT TO DO HERE?

**Theorem 1** *For some $\epsilon > 0$, let*

$$\int_{\mathbb{R}^d} |log(f(x))|^{1+\epsilon} f(x)dx < \infty \tag{12}$$

*and*

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |log(\|x-y\|)|^{2+\epsilon} f(x)f(y)dxdy < \infty \tag{13}$$

*Then*

$$\lim_{N\to\infty} \mathbb{E}(\hat{H}_{N,k}) = H \tag{14}$$

**Theorem 2** *For some $\epsilon > 0$, let*

$$\int_{\mathbb{R}^d} |log(f(x))|^{2+\epsilon} f(x)dx < \infty \tag{15}$$

*and*

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |log(\|x-y\|)|^{2+\epsilon} f(x)f(y)dxdy < \infty \tag{16}$$

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

# 4 Monte-Carlo Simulations

In this section I will explore simulations of the bias of estimator (7) in comparison to the size of the sample estimated from, with respect to different values of k; firstly exploring 1-dimensional distributions and then progressing onto 2-dimensional.

The motivation for these simulations is to explore the consistency of this estimator for different values of $k$; the relationship between the size of the bias of the estimator $\hat{H}_{N,k}$, $Bias(\hat{H}_{N,k})$, and the sample size, $N$. We believe this relationship is of the form;

$$|Bias(\hat{H}_{N,k})| = \frac{c}{N^a} \tag{17}$$

for $a, c > 0$. By taking the logarithm of this, we can see that this relationship is in fact linear;

$$log|Bias(\hat{H}_{N,k})| = log(c) - a[log(N)] \tag{18}$$

I will investigate the consistency of this estimator for a sample from the normal distribution, dependent on the value of $k$. I wish to find the optimum value of $k$ for which $|Bias(\hat{H}_{N,k})| \to 0$ for $N \to \infty$. For the relationship in (17), this will happen for large values of $a$ and relatively small $c$. I will also examine the dependence of the value of $c$ on the value of $k$.

## 4.1 1-dimensional Normal Distribution

I will begin by exploring entropy of samples from the normal distribution $N(0, \sigma^2)$, where without loss of generality we can use the mean $\mu = 0$ and change the variance $\sigma^2$ as needed. The normal distribution has an exact formula to work out the entropy, given the variance $\sigma^2$. Using equation (1) and the density function for the normal distribution $f(x) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)$ for $x \in \mathbb{R}$, given $\mu = 0$, we can write the exact entropy for the normal distribution;

$$H = log(\sqrt{(2\pi e)}\sigma) \tag{19}$$

The normal distribution has the properties which automatically satisfy the conditions above.... condition 1 since ... condition 2 since...

### 4.1.1 k=1

I will first explore the 1-dimensional standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$, $N(0, 1)$. The exact entropy of this distribution is given by;

$$H = log(\sqrt{(2\pi e)}) \approx 1.418939 \tag{20}$$

I consider 1000 samples of size $N$ from this distribution, find the estimator in each case and take the average of these estimators to find our entropy estimator. We will then consider the relationship show in (18) for each sample and again work out the average for the values of a and c.

| N | $\hat{H}_{N,1}$ | $\lvert Bias(\hat{H}_{N,1})\rvert$ | Variance of $\hat{H}_{N,1}$ |
|---|---|---|---|
| 5,000 | $\hat{H}_{5000,1} \approx 1.418289$ | $\lvert Bias(\hat{H}_{5000,1})\rvert \approx 0.000649703$ | $Var(\hat{H}_{5000,1}) \approx 0.0005185365$ |
| 10,000 | $\hat{H}_{10000,1} \approx$ | $\lvert Bias(\hat{H}_{10000,1})\rvert \approx$ | $Var(\hat{H}_{10000,1}) \approx$ |
| 50,000 | $\hat{H}_{50000,1} \approx$ | $\lvert Bias(\hat{H}_{50000,1})\rvert \approx$ | $Var(\hat{H}_{50000,1}) \approx$ |
| 75,000 | $\hat{H}_{75000,1} \approx$ | $\lvert Bias(\hat{H}_{75000,1})\rvert \approx$ | $Var(\hat{H}_{75000,1}) \approx$ |
| 100,000 | $\hat{H}_{100000,1} \approx$ | $\lvert Bias(\hat{H}_{100000,1})\rvert \approx$ | $Var(\hat{H}_{100000,1}) \approx$ |
| 150,000 | $\hat{H}_{150000,1} \approx$ | $\lvert Bias(\hat{H}_{150000,1})\rvert \approx$ | $Var(\hat{H}_{150000,1}) \approx$ |
| 200,000 | $\hat{H}_{200000,1} \approx$ | $\lvert Bias(\hat{H}_{200000,1})\rvert \approx$ | $Var(\hat{H}_{200000,1}) \approx$ |
| 250,000 | $\hat{H}_{250000,1} \approx$ | $\lvert Bias(\hat{H}_{250000,1})\rvert \approx$ | $Var(\hat{H}_{250000,1}) \approx$ |

Figure 1: summary for simulations from the normal distribution with k=1

By considering the situation outline above, with $N = 5000$ I have worked out the estimator for entropy;

$$\hat{H}_{5000,1} \approx 1.41856 \tag{21}$$

Thus,

$$\lvert Bias(\hat{H}_{5000,1})\rvert = \lvert \hat{H}_{5000,1} - H\rvert \approx 0.0003782577 \tag{22}$$

For varying values of N we get the following table;

As we can see for the larger value of N, the Bias of the estimator becomes much smaller, as one would expect for an estimator to satisfy the consistent condition (8)

#### 4.1.2 k=2