Statistical Inference for Entropy

Karina Marks

November 6, 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;

$$\begin{aligned} H &= -\mathbb{E}\{log(f(x))\} \\ &= -\int_{x:f(x)>0} f(x)log(f(x))dx \\ &= -\sum_{x\in\mathbb{R}^d} f(x)log(f(x)) \end{aligned} \tag{1}$$

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

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

Tsallis entropy

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

When the order of the entropy $q \to 1$, both the Rényi, (2), and Tsallis, (3), entropies tend to the Shannon entropy, (1), this is a special case for when $q = 1$. There are also other special cases, sometimes the Rényi entropy is considered for the special case, $q = 2$, and known as the quadratic Rényi entropy;

$$H_2^* = -log\left(\int_{\mathbb{R}^d} f^2(x)dx\right) \tag{4}$$
$$= -log\left(\sum_{x\in\mathbb{R}^d} f^2(x)\right)$$

As $q \to \infty$, the limit of the Rényi entropy exists, and is defined as the minimum entropy, since it's the smallest possible value of H_q^*;

$$H_\infty^* = -\log \sup_{x\in\mathbb{R}^d} f(x)$$

Thus, it follows that; $H_\infty^* \leq H_2^* \leq 2H_\infty^*$.

There is also an approximate relationship between the Shannon entropy and the quadratic Rényi entropy;

$$H_2^* \leq H \leq \log(d) + \frac{1}{d} - e^{-H_2^*}$$

where H_2^* is the quadratic Rényi entropy (4), H is the Shannon entropy (1) and d is the dimension of the distribution.

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 ρ, defined as;

  $$\rho_{(k),i} = \|X_{(k),i} - X_i\| \tag{5}$$

  denote the kth nearest neighbour or X_i.

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

$$V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(1+\frac{d}{2})} \tag{6}$$

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

$$\Psi(k) = -\gamma + \sum_{j=1}^{k-1} \frac{1}{j} \tag{7}$$

  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;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\rho_{(k),i}^{d} V_d (N-1)}{e^{\Psi(k)}} \right] \tag{8}$$

where, $\rho_{(k),i}^{d}$ is defined in (5), V_d is defined in (6) and $\Psi(k)$ is defined in (7).This estimator for entropy, when $d \leq 3$, under a wide range of k and some regularity conditions, satisfies;

$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to 0 \quad (N \to \infty) \tag{9}$$

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(9) no longer holds but the estimator $\hat{H}_{N,k}$, defined by (8), 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, with estimator fo the form;

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

and the 2-dimensional samples, with estimator of the form;

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

3.1.1 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;

$$Bias(\hat{H}_{N,k}) = \mathbb{E}(\hat{H}_{N,k}) - H = \mathbb{E}(\hat{H}_{N,k} - H) \tag{10}$$

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*

$$\int_{\mathbb{R}^d} |log(f(x))|^{1+\epsilon} f(x)dx < \infty \tag{11}$$

and

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |log(\|x-y\|)|^{2+\epsilon} f(x)f(y)dxdy < \infty \tag{12}$$

Then

$$\lim_{N\to\infty} \mathbb{E}(\hat{H}_{N,k}) = H \tag{13}$$

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

$$\int_{\mathbb{R}^d} |log(f(x))|^{2+\epsilon} f(x)dx < \infty \tag{14}$$

and

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

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

3.2 Other Estimators

- The estimator for H_2^* from paper 5
- The estimator for higher dimensions d, from paper 4

4 Monte-Carlo Simulations

In this section I will explore simulations of the bias of estimator (8) 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. Throughout this

analysis we will be considering the absolute value of this bias, since when considering its logarithm, we need a positive value. We believe the relationship between these two variables is of the form;

$$|Bias(\hat{H}_{N,k})| = \frac{c}{N^a} \tag{16}$$

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

$$log|Bias(\hat{H}_{N,k})| = log(c) - a[log(N)] \tag{17}$$

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 (16), 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.

As I wish to consider the difference in accuracy of the estimator when using different values of k, let us denote the approximate values for a and c dependent on k as a_k and c_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 σ^2 as needed. The normal distribution has an exact formula to work out the entropy, given the variance σ^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, using equation (1);

$$\begin{aligned}
H &= -\int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) log\left[\frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)\right] dx \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) \left(log(\sqrt{(2\pi)}\sigma) + \frac{x^2}{2\sigma^2}\right) \\
&= \frac{log(\sqrt{(2\pi)}\sigma)}{\sqrt{(2\pi)}\sigma} \int_{\mathbb{R}} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx + \frac{1}{2\sqrt{(2\pi)}\sigma} \int_{\mathbb{R}} \frac{x^2}{2\sigma^2} \exp\left(\frac{-x^2}{\sigma^2}\right) dx \\
&= log(\sqrt{(2\pi)}\sigma) + \frac{1}{2}
\end{aligned}$$

Thus the exact entropy for the normal distribution is given by

$$H = log(\sqrt{(2\pi e)}\sigma) \tag{18}$$

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

N	$\hat{H}_{N,1}$	$\lvert Bias(\hat{H}_{N,1})\rvert$	Variance of $\lvert Bias(\hat{H}_{N,1})\rvert$
100	$\hat{H}_{100,1} \approx 1.4005$	$\lvert Bias(\hat{H}_{100,1})\rvert \approx 0.1277576$	$Var(\lvert Bias(\hat{H}_{100,1})\rvert) \approx 0.0103719$
200	$\hat{H}_{200,1} \approx 1.40468$	$\lvert Bias(\hat{H}_{200,1})\rvert \approx 0.0955115$	$Var(\lvert Bias(\hat{H}_{200,1})\rvert) \approx 0.005237138$
500	$\hat{H}_{500,1} \approx 1.416559$	$\lvert Bias(\hat{H}_{500,1})\rvert \approx 0.05513945$	$Var(\lvert Bias(\hat{H}_{500,1})\rvert) \approx 0.00185589$
1,000	$\hat{H}_{1000,1} \approx 1.41675$	$\lvert Bias(\hat{H}_{1000,1})\rvert \approx 0.04126499$	$Var(\lvert Bias(\hat{H}_{1000,1})\rvert) \approx 0.0008685067$
5,000	$\hat{H}_{5000,1} \approx 1.418289$	$\lvert Bias(\hat{H}_{5000,1})\rvert \approx 0.000649703$	$Var(\lvert Bias(\hat{H}_{5000,1})\rvert) \approx 0.0005185365$
10,000	$\hat{H}_{10000,1} \approx 1.418916$	$\lvert Bias(\hat{H}_{10000,1})\rvert \approx 0.01327136$	$Var(\lvert Bias(\hat{H}_{10000,1})\rvert) \approx 0.0001016049$
25,000	$\hat{H}_{25000,1} \approx 1.419647$	$\lvert Bias(\hat{H}_{25000,1})\rvert \approx 0.008195638$	$Var(\lvert Bias(\hat{H}_{25000,1})\rvert) \approx 0.00003700078$
50,000	$\hat{H}_{50000,1} \approx 1.419416$	$\lvert Bias(\hat{H}_{50000,1})\rvert \approx 0.005903308$	$Var(\lvert Bias(\hat{H}_{50000,1})\rvert) \approx 0.00002064492$

Figure 1: summary for simulations from the normal distribution with $k = 1$, and varying N

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;

$$H = log(\sqrt{(2\pi e)}) \approx 1.418939 \tag{19}$$

and, since I am considering k=1, the estimator will take the form;

$$\hat{H}_{N,1} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\rho^d_{(1),i} V_d(N-1)}{e^{\Psi(1)}}\right] = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\rho^d_{(1),i} V_d(N-1)}{e^{-\gamma}}\right] \tag{20}$$

I will consider 500 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 (17) for each sample and again work out the average for the values of a and c. For varying values of N we get the table in Figure 1. As we can see for a larger value of N, the Bias of the estimator becomes much smaller; the bias decreases from ≈ 0.1278 to ≈ 0.0059 as N increases from $100 \to 50,000$. This result is to be expected for an estimator to satisfy the consistency condition (**??**). We can also see that the variance of the bias is decreasing as N increases implying that, not only is the average of the estimator getting closer to the actual value of entropy, also the variability between the estimator of different samples is decreasing, making it a consistent and asymptotically unbiased estimator in practice, as well as in theory.

This relationship between the bias $\lvert Bias(\hat{H}_{N,1})\rvert$ of the estimator and the size of the sample N, can be computed for these sample sizes. Figure **??**, shows this relationship of $log\lvert Bias(\hat{H}_{N,1})\rvert$ against $log(N)$ for the samples above, with a fitted regression line. I have also found the corresponding coefficients a and c for the relationship shown in (16); $a =$ and $c =$. On their own these coefficients show that there is a negative relationship between $\lvert Bias(\hat{H}_{N,1})\rvert$ and N, but for them to have more meaning we must compare them to coefficients of the regression relationship for different values of k, and for different distributions.

N	$\hat{H}_{N,2}$	$\lvert Bias(\hat{H}_{N,2})\rvert$	Variance of $\lvert Bias(\hat{H}_{N,2})\rvert$
100	$\hat{H}_{100,2} \approx 1.404102$	$\lvert Bias(\hat{H}_{100,2})\rvert \approx 0.09696776$	$Var(\lvert Bias(\hat{H}_{100,2})\rvert) \approx 0.005354366$
200	$\hat{H}_{200,2} \approx 1.416022$	$\lvert Bias(\hat{H}_{200,2})\rvert \approx 0.07164508$	$Var(\lvert Bias(\hat{H}_{200,2})\rvert) \approx 0.003037111$
500	$\hat{H}_{500,2} \approx 1.4415159$	$\lvert Bias(\hat{H}_{500,2})\rvert \approx 0.04286876$	$Var(\lvert Bias(\hat{H}_{500,2})\rvert) \approx 0.001083868$
1,000	$\hat{H}_{1000,2} \approx 1.416211$	$\lvert Bias(\hat{H}_{1000,2})\rvert \approx 0.03056964$	$Var(\lvert Bias(\hat{H}_{1000,2})\rvert) \approx 0.0005165147$
5,000	$\hat{H}_{5000,2} \approx 1.418589$	$\lvert Bias(\hat{H}_{5000,2})\rvert \approx 0.01322525$	$Var(\lvert Bias(\hat{H}_{5000,2})\rvert) \approx 0.00009693512$
10,000	$\hat{H}_{10000,2} \approx 1.418323$	$\lvert Bias(\hat{H}_{10000,2})\rvert \approx 0.009492$	$Var(\lvert Bias(\hat{H}_{10000,2})\rvert) \approx 0.00005560607$
25,000	$\hat{H}_{25000,2} \approx 1.418986$	$\lvert Bias(\hat{H}_{25000,2})\rvert \approx 0.006037769$	$Var(\lvert Bias(\hat{H}_{25000,2})\rvert) \approx 0.00002110012$
50,000	$\hat{H}_{50000,2} \approx 1.419212$	$\lvert Bias(\hat{H}_{50000,2})\rvert \approx 0.004504959$	$Var(\lvert Bias(\hat{H}_{50000,2})\rvert) \approx 0.00001236801$

Figure 2: summary for simulations from the normal distribution with k=2 and varying N

4.1.2 k=2

I am now going to examine the case where k=2 in the Kozachenko-Leonenko estimator, to compare the results of simulations from this estimator, with that of (20) for k=1. Here the estimator will take the form

$$\hat{H}_{N,2} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\rho^d_{(2),i}V_d(N-1)}{e^{\Psi(2)}}\right] = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\rho^d_{(2),i}V_d(N-1)}{e^{-\gamma+1}}\right] \qquad (21)$$

I wish to explore, in a similar manner as for k=1, the changes in the bias of the estimator depending on a change in N. Additionally, later I will make the comparison between the regression coefficients for different values of k.

I will again consider 500 samples of size N from the 1-dimensional standard normal distribution $N(0,1)$, the results from the analysis is shown in the table in Figure (2).

We can see that, as expected, the Bias of the estimator decreases from ≈ 0.0969 when $N = 100$ to ≈ 0.0045 when $N = 50,000$, showing clearly that the consistency condition is being met here; as $N \to \infty$ we have $|Bias(\hat{H}_{N,2})| \to 0$, which is equivalent to saying $\lim_{N\to\infty} \mathbb{E}(\hat{H}_{N,k}) = H$, Theorem (1).

4.1.3 k=3

4.1.4 k=5

4.1.5 k=10

4.1.6 Comparison of k

N	$\hat{H}_{N,3}$	$\lvert Bias(\hat{H}_{N,3})\rvert$	Variance of $\lvert Bias(\hat{H}_{N,3})\rvert$
100	$\hat{H}_{100,3} \approx 1.396819$	$\lvert Bias(\hat{H}_{100,3})\rvert \approx 0.08063608$	$Var(\lvert Bias(\hat{H}_{100,3})\rvert) \approx 0.003750528$
200	$\hat{H}_{200,3} \approx 1.407153$	$\lvert Bias(\hat{H}_{200,3})\rvert \approx 0.05938714$	$Var(\lvert Bias(\hat{H}_{200,3})\rvert) \approx 0.002145324$
500	$\hat{H}_{500,3} \approx 1.418022$	$\lvert Bias(\hat{H}_{500,3})\rvert \approx 0.0373923$	$Var(\lvert Bias(\hat{H}_{500,3})\rvert) \approx 0.0007464306$
1,000	$\hat{H}_{1000,3} \approx 1.418596$	$\lvert Bias(\hat{H}_{1000,3})\rvert \approx 0.02636523$	$Var(\lvert Bias(\hat{H}_{1000,3})\rvert) \approx 0.000432176$
5,000	$\hat{H}_{5000,3} \approx 1.41799$	$\lvert Bias(\hat{H}_{5000,3})\rvert \approx 0.01165431$	$Var(\lvert Bias(\hat{H}_{5000,3})\rvert) \approx 0.00007221805$
10,000	$\hat{H}_{10000,3} \approx 1.418454$	$\lvert Bias(\hat{H}_{10000,3})\rvert \approx 0.009006306$	$Var(\lvert Bias(\hat{H}_{10000,3})\rvert) \approx 0.00004969628$
25,000	$\hat{H}_{25000,3} \approx 1.418685$	$\lvert Bias(\hat{H}_{25000,3})\rvert \approx 0.005304324$	$Var(\lvert Bias(\hat{H}_{25000,3})\rvert) \approx 0.00001522065$
50,000	$\hat{H}_{50000,3} \approx 1.418678$	$\lvert Bias(\hat{H}_{50000,3})\rvert \approx 0.003799854$	$Var(\lvert Bias(\hat{H}_{50000,3})\rvert) \approx 0.000008950053$

Figure 3: summary for simulations from the normal distribution with k=3 and varying N

N	$\hat{H}_{N,5}$	$\lvert Bias(\hat{H}_{N,5})\rvert$	Variance of $\lvert Bias(\hat{H}_{N,5})\rvert$
100	$\hat{H}_{100,5} \approx$	$\lvert Bias(\hat{H}_{100,5})\rvert \approx$	$Var(\lvert Bias(\hat{H}_{100,5})\rvert) \approx$
200	$\hat{H}_{200,5} \approx$	$\lvert Bias(\hat{H}_{200,5})\rvert \approx$	$Var(\lvert Bias(\hat{H}_{200,5})\rvert) \approx$
500	$\hat{H}_{500,5} \approx$	$\lvert Bias(\hat{H}_{500,5})\rvert \approx$	$Var(\lvert Bias(\hat{H}_{500,5})\rvert) \approx$
1,000	$\hat{H}_{1000,5} \approx$	$\lvert Bias(\hat{H}_{1000,5})\rvert \approx$	$Var(\lvert Bias(\hat{H}_{1000,5})\rvert) \approx$
5,000	$\hat{H}_{5000,5} \approx$	$\lvert Bias(\hat{H}_{5000,5})\rvert \approx$	$Var(\lvert Bias(\hat{H}_{5000,5})\rvert) \approx$
10,000	$\hat{H}_{10000,5} \approx$	$\lvert Bias(\hat{H}_{10000,5})\rvert \approx$	$Var(\lvert Bias(\hat{H}_{10000,5})\rvert) \approx$
25,000	$\hat{H}_{25000,5} \approx$	$\lvert Bias(\hat{H}_{25000,5})\rvert \approx$	$Var(\lvert Bias(\hat{H}_{25000,5})\rvert) \approx$
50,000	$\hat{H}_{50000,5} \approx$	$\lvert Bias(\hat{H}_{50000,5})\rvert \approx$	$Var(\lvert Bias(\hat{H}_{50000,5})\rvert) \approx$

Figure 4: summary for simulations from the normal distribution with k=5 and varying N