Statistical Inference for Entropy

Karina Marks

November 12, 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...

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;

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

Table 1: 1-dimensional normal distribution, $k = 1$

N	$\hat{H}_{N,1}$	$\lvert Bias(\hat{H}_{N,1})\rvert$	$Var(Bias(\hat{H}_{N,1}))$
100	1.405890	0.01304883282	0.0275142086
200	1.411070	0.00786872927	0.0128689734
500	1.416666	0.00227293433	0.0051416433
1000	1.419401	0.00046261516	0.0028127916
5000	1.418469	0.00046981107	0.0005147810
10000	1.417998	0.00094067533	0.0002472848
25000	1.418877	0.00006147045	0.0001088641
50000	1.419286	0.00034705584	0.0000496450

4.1.1 k=1

I will be considering k=1 for the estimator of entropy; thus, 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, shown in table 1. 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, shown in **??**.

Considering table 1, we can see for a larger value of N, the Bias of the estimator becomes much smaller; the bias decreases from ≈ 0.0130 to ≈ 0.0003 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 $|Bias(\hat{H}_{N,1})|$ of the estimator and the size of the sample N, can be computed for these sample sizes. Figure **??**, shows this relationship of $log|Bias(\hat{H}_{N,1})|$ 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 $|Bias(\hat{H}_{N,1})|$ 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.

Table 2: 1-dimensional normal distribution, $k = 2$

N	$\hat{H}_{N,2}$	$\lvert Bias(\hat{H}_{N,2})\rvert$	$Var(Bias(\hat{H}_{N,2}))$
100	1.408856	0.0100827948	0.01357417708
200	1.411165	0.0077730666	0.00688250329
500	1.419158	0.0002199163	0.00296693934
1000	1.415719	0.0032197158	0.00141616592
5000	1.418236	0.0007026416	0.00028872533
10000	1.418656	0.0002824567	0.00014348493
25000	1.418376	0.0005620780	0.00005791073
50000	1.418681	0.0002574343	0.00002956529

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 table 2.

We can see that, as expected, the Bias of the estimator decreases from ≈ 0.0100 when $N = 100$ to ≈ 0.0002 when $N = 50,000$, showing clearly that the consistency condition is being met here. This is true since 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). We also have that the variance of the bias of these estimators decrease as $N \to \infty$, as expected. In comparison to $k = 1$, we can see that the bias of this estimator for $k = 2$ decreases at a similar pace as $N \to \infty$; $|Bias(\hat{H}_{N,1})| \approx 0.0130 \to 0.0003$ and $|Bias(\hat{H}_{N,2})| \approx 0.0101 \to 0.0002$.

CONSIDER THE GRAPH...

4.1.3 k=3

Again, for $k = 3$, I will examine 500 samples of size N from the standard normal distribution considered before; with estimator of the form;

$$\hat{H}_{N,3} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\rho^d_{(3),i}V_d(N-1)}{e^{\Psi(3)}}\right] = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\rho^d_{(3),i}V_d(N-1)}{e^{-\gamma+1+\frac{1}{2}}}\right] \qquad (22)$$

Table 3: 1-dimensional normal distribution, $k = 3$

N	$\hat{H}_{N,3}$	$\lvert Bias(\hat{H}_{N,3})\rvert$	$Var(Bias(\hat{H}_{N,3}))$
100	1.398784	0.0201546812	0.01210622150
200	1.412908	0.0060302660	0.00530612702
500	1.414035	0.0049035937	0.00223855589
1000	1.416105	0.0028340080	0.00107754839
5000	1.420184	0.0012459298	0.00022320970
10000	1.418351	0.0005874791	0.00011630350
25000	1.419115	0.0001760980	0.00004286406
50000	1.418853	0.0000851863	0.00002257717

Table 4: 1-dimensional normal distribution, $k = 5$

N	$\hat{H}_{N,5}$	$\lvert Bias(\hat{H}_{N,5})\rvert$	$Var(Bias(\hat{H}_{N,5}))$
100	1.391834	0.02710439666	0.00807261026
200	1.405356	0.01358205942	0.00425419382
500	1.411436	0.00750282472	0.00168848112
1000	1.415091	0.00384740080	0.00091927735
5000	1.418150	0.00078877480	0.00018941496
10000	1.418648	0.00029099525	0.00008767553
25000	1.418879	0.00005917171	0.00003243503
50000	1.418644	0.00029451951	0.00001705529

The results are displayed in table 3.

This shows again, that the Kozachenko-Leonenko estimator for entropy tends to 0, $\hat{H}_{N,3} \to 0$, as $N \to \infty$. However, comparing these results to those for $k = 1, 2$, which had similar bias as $N \to \infty$, we can see that for $k = 3$, $|Bias(\hat{H}_{N,3})| \approx 0.0202 \to 0.00009$. So for larger N, the estimator with $k = 1 or 2$ would be less appropriate to use, since the bias is slightly larger than for the estimator using $k = 3$.

CONSIDER THE GRAPH...

4.1.4 k=5

This analysis is done again, but for k=5, and the results are shown in table 4

Here, the $|Bias(\hat{H}_{N,5})|$ decreases as N goes from $100 \to 25,000$, but at $50,000$ this jumps to a larger number. Up to $25,000$ indicates that the estimator is becoming closer to the actual value, the jump at $50,000$ could be due to a number of reasons.

Table 5: 1-dimensional normal distribution, $k = 10$

N	$\hat{H}_{N,10}$	$\lvert Bias(\hat{H}_{N,10})\rvert$	$Var(Bias(\hat{H}_{N,10}))$
100	1.375699	0.0432399931	0.00678770166
200	1.391934	0.0270050257	0.00293164825
500	1.407625	0.0113137866	0.00148669638
1000	1.411684	0.0072549983	0.00067990485
5000	1.417306	0.0016322988	0.00013650841
10000	1.418196	0.0007429215	0.00006783354
25000	1.418356	0.0005825702	0.00003162161
50000	1.418790	0.0001488755	0.00001318863

Firstly, this could indicate that for $k = 5$, this estimator becomes less efficient, and doesn't satisfy the property ... as strongly as smaller values of k have done so far.

Secondly,this could just be an error in the data for $|Bias(\hat{H}_{50000,5})|$; since we are only considering a relative small number of samples, 500, and are taking the average of this, we could have just found an outlier.

Lastly, there could be an error in the previous two data points, $|Bias(\hat{H}_{25000,5})|$ and $|Bias(\hat{H}_{10000,5})|$, causing us to either believe it is decreasing, when it isn't - satisfyin the first reason. Or, this could mean that the second condition isn't correct.... YADA YADA

To determine the reason for this jump of Bias in the wrong direction, I will examine $|Bias(\hat{H}_{50000,5})|$ and $Var(Bias(\hat{H}_{50000,5}))$ for 5,000 samples and see if this is consistent with the previous findings. I have found this number to be;

$$|Bias(\hat{H}_{50000,5})| \approx ... \tag{23}$$

Thus, YADA YADA

4.1.5 k=10

4.1.6 Comparison of k

The above analysis, sections ... to ..., is done to examine the difference in the bias of the estimator for different values of k. Considering the above samples, for $N = 25,000$ and $N = 50,000$, we can create a table to compare the different k;

The results shown in table 6 are inconclusive in determining if larger/smaller values of k generate better estimators, with smaller bias. However, these results are consistent in showing that for the larger value of N, the smaller the variance in the estimator; thus, we know that .... The results for the Bias are not conclusive because for when $N = 25,000$ we can see that for $k = 1, 5$ and possibly $k = 3$ have a slight smaller bias than the others. However, when $N = 50,000$ we

Table 6: 1-dimensional normal distribution, comparison of k

k	$\lvert Bias(\hat{H}_{25000,k})\rvert$	$Var(Bias(\hat{H}_{25000,k}))$	$\lvert Bias(\hat{H}_{50000,k})\rvert$	$Var(Bias(\hat{H}_{50000,k}))$
1	0.00006147045	0.0001088641	0.00034705584	0.0000496450
2	0.0005620780	0.00005791073	0.0002574343	0.00002956529
3	0.0001760980	0.00004286406	0.0000851863	0.00002257717
5	0.00005917171	0.00003243503	0.00029451951	0.00001705529
10	0.0005825702	0.00003162161	0.0001488755	0.00001318863

This table is comparing the values of $|Bias(\hat{H}_{N,k})|$ for the values of k explored in tables 1, 2, 3, 4 and 5 with $N = 25,000$ and $N = 50,000$, when the estimator is taken over 500 samples

find that for $k = 3, 10$ we have the smallest values of bias. These are inconsistent with one and other. To further examine this, I will now generate a table for values $k = 1, 2, 3, 5 and 10$ with $N = 50,000$ in all cases. Moreover, this time I will consider $3,000$ samples of this size, not the 500 considered before, and will find the mean and variance of the bias of this estimator.

From the results in table 7, we can see that $|Bias(\hat{H}_{N,k})|$ is the smallest, for sample size $N = 50,000$, when $k = 3$, which is consistent with the results found in table 6. So from these simulations, we can conclude that for large N, the consistency condition is best satisfied when $k = 3$. Interestingly, the $Var(Bias(\hat{H}_{50000,k})) \to 0$ for $k \to 10$, but this is to be expected, as by the definition of the estimator using the nearest neighbour method. Taking a larger k in the nearest neighbour method will produce less varied results, this is because more smoothing takes place for a larger k, eventually - if k is made large enough - the output will be constant and the variance negligible regardless of the inputted values. Thus, considering the variance of the bias of the estimator is not necessarily informative.

4.2 1-dimensional Uniform distribution

I will now explore the entropy of samples from the 1-dimensional uniform distribution, $U[ab]$. This distribution also has an exact formula to work out the entropy for. We can find this formula by considering the density function, f, from the uniform distribution, which is given by;

$$f(x) = ... \quad a \leq x \leq b \tag{24}$$

Table 7: 1-dimensional normal distribution, comparison of k

k	$\lvert Bias(\hat{H}_{50000,k})\rvert$	$Var(Bias(\hat{H}_{50000,k}))$
1	0.00013495546	0.00005116758
2	0.00012647214	0.00002868082
3	0.00003478968	0.00002299754
5	0.00006034936	0.00001733369
10	0.00022455715	0.00001409080

This table is comparing the values of $|Bias(\hat{H}_{N,k})|$ for the values of k explored before now with only $N = 50,000$ and the estimator being taken over $3,000$ samples

Using the definition of Shannon entropy given in equation (1), we can work this out;

$$H = ... \tag{25}$$

Thus, the actual value of entropy for the uniform disrtibution is given by;

$$H = \tag{26}$$

In our samples we will be consider the uniform distribution $U[0, 100]$; this is because, using the standard uniform, $U[0, 1]$, would fail since taking $N = 50,000$ samples between 0 and 1 would generate problems since the pdf would be $f(x) = \frac{1}{50000} \approx 0.00002$, which would incur working on a very small scale.