Statistical Inference for Entropy

Karina Marks

November 22, 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

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

$$= \frac{1}{q-1}\left(1 - \sum_{x\in\mathbb{R}^d} f^q(x)\right)$$

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;

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

$$= -log\left(\sum_{x\in\mathbb{R}^d} f^2(x)\right)$$

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;

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

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

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

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

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

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);

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

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

$$= \int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) \left(log(\sqrt{(2\pi)}\sigma) + \frac{x^2}{2\sigma^2}\right)$$

$$= \frac{log(\sqrt{(2\pi)}\sigma)}{\sqrt{(2\pi)}\sigma} \int_{\mathbb{R}} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx + \frac{1}{2\sqrt{(2\pi)}\sigma} \int_{\mathbb{R}} \frac{x^2}{2\sigma^2} \exp\left(\frac{-x^2}{\sigma^2}\right) dx$$

$$= log(\sqrt{(2\pi)}\sigma) + \frac{1}{2}$$

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...

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

N	$\hat{H}_{N,1}$	$|Bias(\hat{H}_{N,1})|$	$Var(Bias(\hat{H}_{N,1}))$
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

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;

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

Since, I am first considering the 1-dimensional normal distribution, the Kozachenko-Leonenko estimator, (8), has $d = 1$. Thus, the volume of the 1-dimensional Euclidean ball is $V_1 = \frac{\pi^{\frac{1}{2}}}{\Gamma(\frac{3}{2})} = \frac{\sqrt{\pi}}{\frac{\sqrt{\pi}}{2}} = 2$. Hence, the estimator takes the form;

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

4.1.1   k=1

I will be considering k=1 for the estimator of entropy; thus, the estimator will take the form;

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

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 1.

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 $100 \rightarrow 50,000$. 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

N	a_1	c_1
5000	0.6545	0.1845
75000	0.5011	0.0462
10000	0.7699	0.3512
25000	0.5442	0.0620
50000	0.4594	0.0249

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 1, shows this relationship of $log|Bias(\hat{H}_{N,1})|$ against $log(N)$ with a fitted regression line. In this graph, I have considered the values of N up to $50,000$ at intervals of size 100, where each point is calculate 500 times and the average estimator is plotted. I have also found the corresponding coefficients a_1 and c_1 for the relationship shown in (16). These coefficients tend towards $a_1 = 0.4594$ and $c_1 = 0.0249$ for a large N;

Doesn't show what expected?? TODO..

$$|Bias(\hat{H}_{N,1})| \approx \frac{0.0249}{N^{0.4594}}$$

On their own these coefficients show that there is a negative relationship between $log(|Bias(\hat{H}_{N,1})|)$ and $log(N)$. This implies that the relationship between the Bias and N is such that as N increases the Bias decreases to 0. Hence, creating a consistent estimator for entropy, when considering the 1-dimensional normal distribution with k=1 in the Kozachenko-Leonenko estimator. However, to understand better the meaning of this relationship we must compare this to coefficients found of the regression relationships for different values of k, and for different distributions.

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 for k=1. Here the estimator will take the form;

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

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
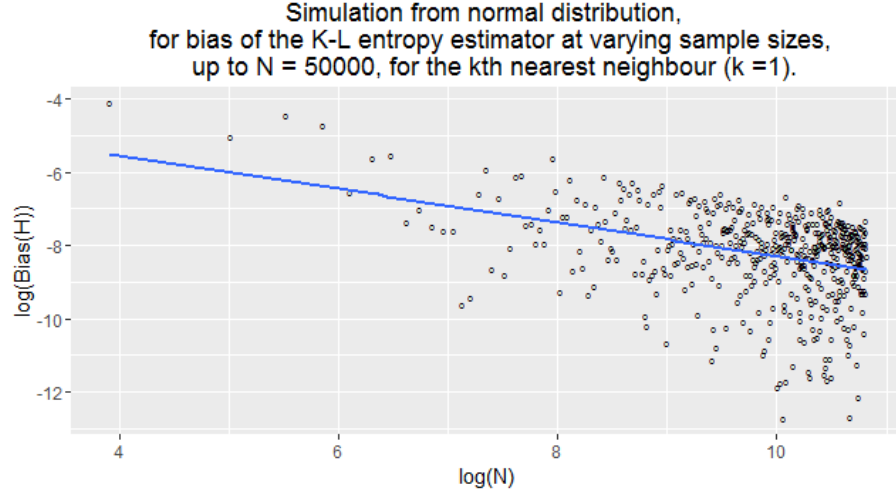
Figure 1: *Regression plot of* $\log|Bias(\hat{H}_{N,1})|$ *against* $\log(N)$

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

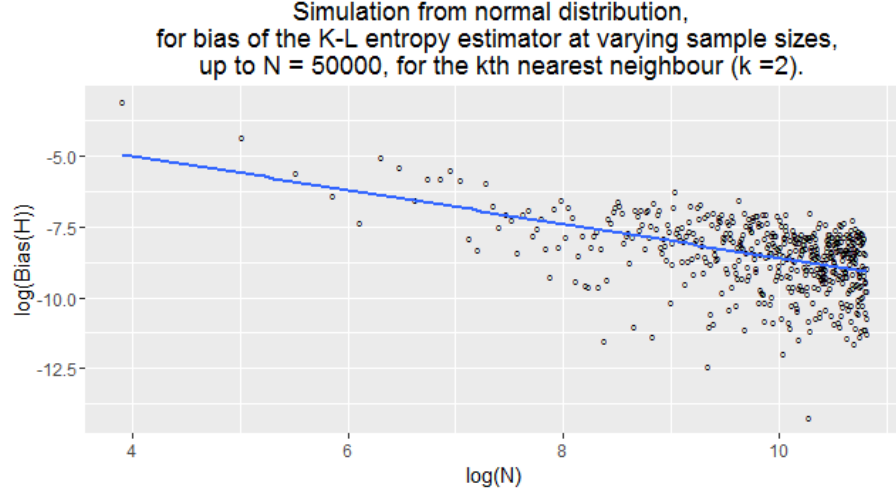| N	$\hat{H}_{N,2}$	$|Bias(\hat{H}_{N,2})|$	$Var(Bias(\hat{H}_{N,2}))$
100	1.408856	0.0100827948	0.01357417708
200	1.411165	0.0077730666	0.00688250329
500	1.419158	0.0002199163	0.00296693934
1000	1.415719	0.0032197158	0.00141616592
5000	1.418236	0.0007026416	0.00028872533
10000	1.418656	0.0002824567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

Figure 2: *Regression plot of* $\log|Bias(\hat{H}_{N,2})|$ *against* $\log(N)$

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 $N = 50,000$. This is showing that the consistency condition is being met 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 relation 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 $|Bias(\hat{H}_{N,2})| \approx 0.0101 \to 0.0002$, implying that from this analysis we cannot decide which value of k generates a better estimator.

We have found the coefficients for the equation (16), for $k = 2$, which are given by $a_2 = 0.5998$ and $c_2 = 0.0746$, thus;

$$|Bias(\hat{H}_{N,2})| \approx \frac{0.0746}{N^{0.5998}}$$

Again, this shows the relationship one would expect, that $|Bias(\hat{H}_{N,2})| \to 0$ as $N \to \infty$. This relationship for $k = 2$ is stronger than that for $k = 1$ since $a_1 \leq a_2$. A full comparison of the relationship of $|Bias(\hat{H}_{N,1})|$ to N and $|Bias(\hat{H}_{N,2})|$ to N is given in section 4.1.6.

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

N	$\hat{H}_{N,3}$	$|Bias(\hat{H}_{N,3})|$	$Var(Bias(\hat{H}_{N,3}))$
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

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;

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

The results of the comparison between the actual value and the estimated value of entropy, for different values of N, are displayed in table 3. This shows again, that the Kozachenko-Leonenko estimator for entropy, $\hat{H}_{N,3} \to 0$, as $N \to \infty$, as more specifically $\hat{H}_{50000,3} \approx 0.00009$. 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 $k = 2$ would be less appropriate to use, since the bias is slightly larger than for the estimator using $k = 3$.

The graph showing the relationship given by **??** is shown in figure 3. The have found the coefficients for the formula 16, for the graph shown with $k = 3$ are given by $a_3 = 0.6443$ and $c_3 = 0.1156$, thus;

$$|Bias(\hat{H}_{N,3})| \approx \frac{0.1156}{N^{0.6443}}$$

So for $N = 50,000$, this implies that $|Bias(\hat{H}_{50000,3})| \approx 0.000176$, which is close to the value found in table 3, that $|Bias(\hat{H}_{50000,3})| \approx 0.000085$. We here have that $a_3 \geq a_2 \geq a_1$, hence, according to this analysis, when $k = 3$ we have a stronger negative relationship between the bias and N. A full comparison of the regression analysis for each k is conducted in section 4.1.6.
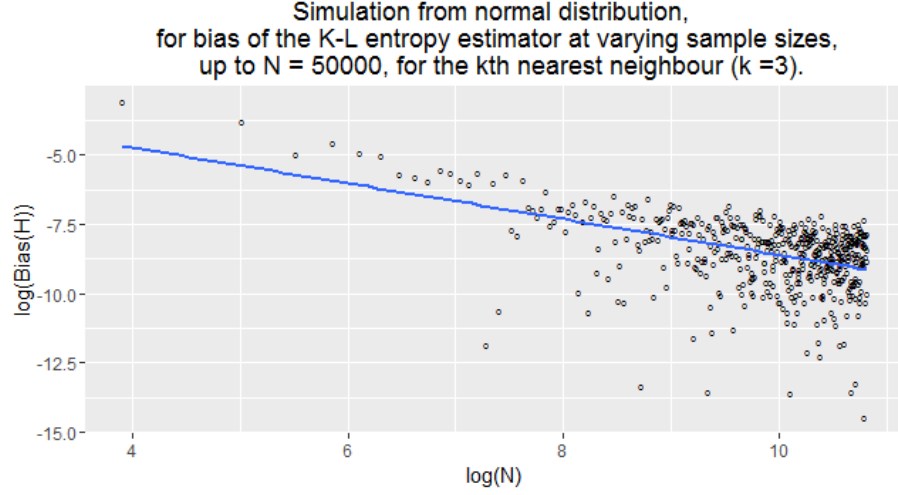
Simulation from normal distribution,
for bias of the K-L entropy estimator at varying sample sizes,
up to N = 50000, for the kth nearest neighbour (k =3).

Figure 3: *Regression plot of* $\log|Bias(\hat{H}_{N,3})|$ *against* $\log(N)$

4.1.4   k=5

Now we consider the estimator, shown in equation 8, for k=5. This takes the form;

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

Comparing this to the exact entropy for the standard normal distribution, (19), gives table 4. Here, the $|Bias(\hat{H}_{N,5})|$ decreases as N goes from $100 \rightarrow 25,000$, 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 $50,000$

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

N	$\hat{H}_{N,5}$	$|Bias(\hat{H}_{N,5})|$	$Var(Bias(\hat{H}_{N,5}))$
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

11

**Simulation from normal distribution,
for bias of the K-L entropy estimator at varying sample sizes,
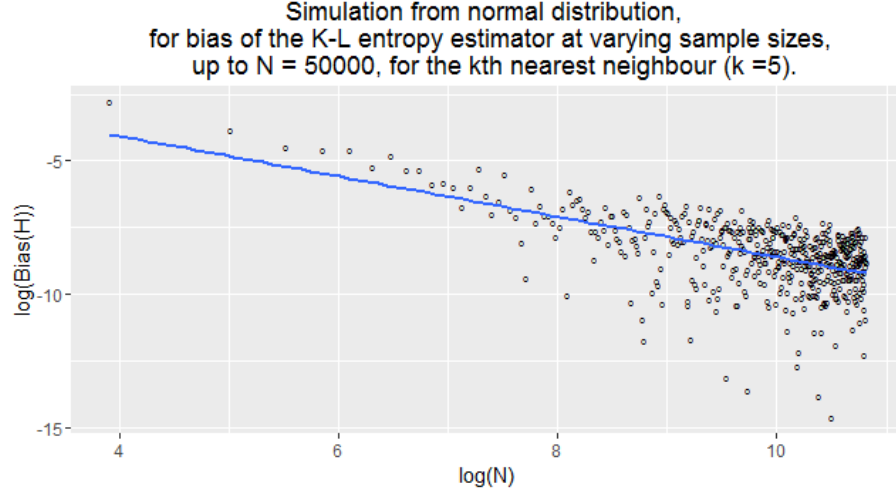up to N = 50000, for the kth nearest neighbour (k =5).**

Figure 4: *Regression plot of* $\log |Bias(\hat{H}_{N,5})|$ *against* $\log(N)$

could be due to a number of reasons.

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 just have 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.

To determine the reason for this jump of Bias in the wrong direction, I will examine $|Bias(\hat{H}_{50000,5})|$ for 3,000 samples and see if this is consistent with the previous findings. I have found this number to be;

$$|Bias(\hat{H}_{50000,5})| \approx 0.00006034936$$

This gives us a much smaller bias than $|Bias(\hat{H}_{50000,5})|$ shown in table 4, however, this is still not smaller than the value of $|Bias(\hat{H}_{25000,5})|$ shown in the same table. This could mean that for $k = 5$, the estimator doesn't satisfy the consistency condition as strongly as the previous estimators for $k = 1, 2, 3$.

However, if we consider the graph in figure 4, we can see an obvious negative relationship between the logarithm of $|Bias(\hat{H}_{N,5})|$ and the logarithm of N. Henceforth, I would expect that the numbers above are within the standard error range, so that for $k = 5$, we do have an estimator which satisfies the consistency condition **??**.

By plotting the graph in figure 4, we have found the coefficients for the

12

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

N	$\hat{H}_{N,10}$	$|Bias(\hat{H}_{N,10})|$	$Var(Bias(\hat{H}_{N,10}))$
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

formula 16, for $k = 5$, which are given by $a_5 = 0.7568$ and $c_5 = 0.3557$, thus;

$$|Bias(\hat{H}_{N,5})| \approx \frac{0.3557}{N^{0.7568}}$$

For these coefficients we have that $a_5 \geq a_3 \geq a_2 \geq a_1$, thus according to this analysis, we have a stronger consistency of our estimator for $k = 5$, in comparison to $k = 1, 2, 3$. This comparison will be considered in more detail in section 4.1.6.

4.1.5   k=10

The last estimator for the entropy of a sample from the standard normal distribution that I wish to explore is that for $k = 10$. Here, the estimator takes the form;

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

The results for the comparison between this estimator and 19 are displayed in table 5.

Here, we can again see that this estimator is satisfying the consistency condition **??**, as $\hat{H}_{N,10} \approx 0.0432 \rightarrow 0.0001$ as $N \approx 100 \rightarrow 50,000$. Comparing this to previous values of k, we can see that the bias changes decreases in a similar manner to that for $k = 1, 2, 3$.

CONSIDER GRAPH...

4.1.6   Comparison of k

The above analysis, sections 4.1.1 to 4.1.5 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 values of the bias of the estimator for the different values of k considered;

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

k	$|Bias(\hat{H}_{25000,k})|$	$Var(Bias(\hat{H}_{25000,k}))$	$|Bias(\hat{H}_{50000,k})|$	$Var(Bias(\hat{H}_{50000,k}))$
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

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. The results for the Bias are not conclusive; because, for $N = 25,000$ we can see that with $k = 1, 5$ and possibly $k = 3$ have a slight smaller bias than the others. However, when $N = 50,000$ we 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 $Var(Bias(\hat{H}_{50000,k})) \rightarrow 0$ for $k \rightarrow 10$, 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.

Table 8, shows that as k runs from $1 \rightarrow 5$, we have that a_k and c_k both increase. The higher the value of a_k, the stronger the negative relationship is between the two variables in question, so for a larger values of a_k, we have that $|Bias(\hat{H}_{N,k})| \rightarrow 0$ for large N faster than smaller values of a_k. This is due to the relationship between $|Bias(\hat{H}_{N,k})|$ and a_k shown in equation (16). Thus, considering a large sample size, say $N = 100,000$, we can find the bias of the

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

k	$|Bias(\hat{H}_{50000,k})|$	$Var(Bias(\hat{H}_{50000,k}))$
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

Table 8: *Comparison of coefficients of regression a_k and c_k from equation 16, for 1-dimensional normal distribution*

k	a_k	c_k
1	0.4594	0.0249
2	0.5998	0.0746
3	0.6443	0.1156
5	0.7568	0.3557
10		

15

Kozachenko-Leonenko estimator according to the regressional relationship for each k; $|Bias(\hat{H}_{N,k})| = \frac{c_k}{N^{a_k}}$. We find that;

$$|Bias(\hat{H}_{100000,1})| \approx \frac{0.0249}{100000^{0.4594}} \approx 0.00012566$$

$$|Bias(\hat{H}_{100000,2})| \approx \frac{0.0746}{100000^{0.5998}} \approx 0.00007477$$

$$|Bias(\hat{H}_{100000,3})| \approx \frac{0.1156}{100000^{0.6443}} \approx 0.00006942$$

$$|Bias(\hat{H}_{100000,5})| \approx \frac{0.3557}{100000^{0.7568}} \approx 0.00005949$$

This shows that the bias decreases slightly faster for a higher values of $k = 1, 2, 3, 5$.

What does c_k show?? TODO..

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;

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & otherwise \end{cases}$$

Using the definition of Shannon entropy given in equation (1), we can find the exact entropy for the uniform distribution;

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

$$= -\int_a^b \frac{1}{b-a}log\left[\frac{1}{b-a}\right]dx$$

$$= -\frac{1}{b-a}log\left[\frac{1}{b-a}\right]\int_a^b dx$$

$$= -log\left[\frac{1}{b-a}\right]$$

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

$$H = log[b-a] \tag{20}$$

The uniform distribution automatically satisfies the conditions .... because ...

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 $N =$

Table 9: *1-dimensional uniform distribution, $k = 1$*

N	$\hat{H}_{N,1}$	$|Bias(\hat{H}_{N,1})|$	$Var(Bias(\hat{H}_{N,1}))$
100	4.596921	0.00824926770	0.0222674976
200	4.616316	0.01114602105	0.0100263188
500	4.601683	0.00348766754	0.0040618438
1000	4.610342	0.00517161041	0.0017659932
5000	4.605226	0.00005536281	0.0003974555
10000	-Inf	Inf	NaN
25000	-Inf	Inf	NaN
50000	-Inf	Inf	NaN

$50,000$ samples between 0 and 1 would generate problems as the pdf would be $f(x) = 1, \quad 0 \leq x \leq 1$, which would incur working on a very small scale; i.e taking a points with distance between them as ≈ 0.00002 along the x-direction. Thus using the pdf $f(x) = 0.01, \quad 0 \leq x \leq 100$, which gives the exact entropy to be;

$$H = log(100) \approx 4.605170 \tag{21}$$

Similarly to the 1-dimensional normal distribution, we have for the 1-dimensional uniform distribution that $d = 1$ so $V_1 = 2$, thus our estimator takes the form;

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

4.2.1 k=1

We will begin by considering 500 samples from the uniform distribution $U[0, 100]$ and finding the estimator for $k = 1$, which is of the form;

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

Considering the bias for this estimator against the actual value (21) for different samples sizes N gives table 9.

We can see from this table that for $N \geq 10,000$ the estimator cannot be computed. This is due to the same reason, that we decided to use the distribution $U[0, 100]$ instead of the standard normal $U[0, 1]$. However, even considering $U[0, 100000]$ only gives non infinite values for $N \leq 10,000$. Hence, from now on, we will only consider samples of size $N \leq 5,000$ and we will use the distribution $U[0, 100]$. From the above table, we can see obviously, that for the largest non-infinite value of N, we have $|Bias(\hat{H}_{5000,1})| \approx 0.00006$, which is significantly smaller than $|Bias(\hat{H}_{100,1})| \approx 0.00825$, confirming the consistency condition.

CONSIDER GRAPHS..

Table 10: *1-dimensional uniform distribution, $k = 2$*

N	$\hat{H}_{N,2}$	$|Bias(\hat{H}_{N,2})|$	$Var(Bias(\hat{H}_{N,2}))$
100	4.606725	0.0015545396	0.00851906912
200	4.610785	0.0056145269	0.00456819154
500	4.611599	0.0064290000	0.00189878458
1000	4.603534	0.0016366244	0.00095980274
5000	4.604978	0.0001921979	0.00017282038
10000	4.605488	0.0003180407	0.00009978981
25000	4.604753	0.0004176853	0.00004003515
50000	4.605480	0.0003095010	0.00001737807

Table 11: *1-dimensional uniform distribution, $k = 3$*

N	$\hat{H}_{N,3}$	$|Bias(\hat{H}_{N,3})|$	$Var(Bias(\hat{H}_{N,3}))$
100	4.610386	0.00521552744	0.00697968570
200	4.611047	0.00587685467	0.00316173901
500	4.605035	0.00013506468	0.00121563168
1000	4.606418	0.00124808620	0.00054151215
5000	4.605270	0.00009934301	0.00011448612
10000	4.604869	0.00030102035	0.00007028042
25000	4.605341	0.00017121288	0.00002543334
50000	4.605123	0.00004761182	0.00001155187

4.2.2   k=2

We now wish to consider the estimator for $k = 2$, which takes the form;

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

4.2.3   k=3

We now wish to consider the estimator for $k = 3$, which takes the form;

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

Table 12: *1-dimensional uniform distribution, k = 5*

N	$\hat{H}_{N,5}$	$|Bias(\hat{H}_{N,5})|$	$Var(Bias(\hat{H}_{N,5}))$
100	4.621001	0.0158306351	0.003899909770
200	4.612364	0.0071935784	0.001744887954
500	4.609793	0.0046227604	0.000768266967
1000	4.607499	0.0023291487	0.000381576396
5000	4.605980	0.0008102069	0.000068805987
10000	4.606053	0.0008829085	0.000035434958
25000	4.605339	0.0001689148	0.000015449454
50000	4.605333	0.0001629615	0.000007555981

Table 13: *1-dimensional uniform distribution, k = 10*

N	$\hat{H}_{N,10}$	$|Bias(\hat{H}_{N,10})|$	$Var(Bias(\hat{H}_{N,10}))$
100	4.639476	0.03430601671	0.002521145015
200	4.621455	0.01628474750	0.000951343553
500	4.611200	0.00602956738	0.000380048902
1000	4.609219	0.00404887162	0.000186210819
5000	4.605507	0.00033726494	0.000037485008
10000	4.605341	0.00017035096	0.000018679424
25000	4.605138	0.00003191065	0.000007044257
50000	4.605190	0.00002025184	0.000003429152