Statistical Inference for Estimation of Entropy

Karina Marks

February 27, 2017

Abstract

What am I doing?

Contents

Chapter 1

Introduction

[20] [21] [5] [3] [22] [25] [14] [16] [9] [2]

1.1 Entropy

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.

1.1.1 Shannon Entropy

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

$$\begin{aligned} H &= -\mathbb{E}\{log(f(x))\} \\ &= -\int_{x:f(x)>0} f(x)log(f(x))dx \\ &= -\sum_{x\in\mathbb{R}^d} f(x)log(f(x)) \qquad (1.1) \end{aligned}$$

1.1.2 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 1.1. For a random vector $X \in \mathbb{R}^d$ with density function f, we define;

Rényi entropy

$$\begin{aligned} H_q^* &= \frac{1}{1-q} log\left(\int_{\mathbb{R}^d} f^q(x)dx\right) \qquad (q \neq 1) \qquad (1.2) \\ &= \frac{1}{1-q} log\left(\sum_{x\in\mathbb{R}^d} f^q(x)\right) \end{aligned}$$

Tsallis entropy

$$\begin{aligned} H_q &= \frac{1}{q-1}\left(1-\int_{\mathbb{R}^d} f^q(x)dx\right) \qquad (q \neq 1) \\ &= \frac{1}{q-1}\left(1-\sum_{x\in\mathbb{R}^d} f^q(x)\right) \end{aligned} \tag{1.3}$$

When the order of the entropy $q \to 1$, both the Rényi, (1.2), and Tsallis, (1.3), entropies tend to the Shannon entropy, (1.1), 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;

$$\begin{aligned} H_2^* &= -log\left(\int_{\mathbb{R}^d} f^2(x)dx\right) \\ &= -log\left(\sum_{x\in\mathbb{R}^d} f^2(x)\right) \end{aligned} \tag{1.4}$$

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 (1.4), H is the Shannon entropy (1.1) and d is the dimension of the distribution.

Chapter 2

Background

2.1 Properties of Entropy

Kapur and Kesavan's book on *Entropy Optimization Principles with Applications* [16], gives an account of some of the properties of the Shannon entropy H, which I will mention. First recall the definition of Shannon entropy (equation (1.1);

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

where f is the density of the distribution of x. Some properties are as follows;

- H is permutationally symmetric
- For $f(x)$ continuous on some interval; H is also continuous everywhere in the same interval
- Entropy doesn't change by the inclusion of an impossible event
- $H > 0$ for all circumstances unless if f is any of the N degenerate distributions; where $f(x_i) = 1$ if $i = k, k \in [1, N]$ otherwise $f(x_i) = 0$, then $H = 0$
- H is a concave function, *for Rényi and Tsallis entropies, when* $q > 0$ *they're concave; however, if* $q < 0$ *then they're convex*
- The maximum value of H is attained when f is the uniform distribution
- For two independent distributions ($f_X(x)$ and $f_Y(y)$), the entropy of their joint distribution ($f_{X,Y}(x, y)$) is just the sum of the entropies of the two distributions; $H(f_{X,Y}) = H(f_X) + H(f_Y)$

2.2 Applications of Entropy

Entropy began as a concept in thermodynamics, about the idea of that within any irreversible system, a small amount of heat energy is aways lost. Entropy has more recently found application in the field of information theory, where it describes a similar loss, this time of missing information or data in systems of information transmission. Thus, entropy has many applications across both these areas.

I will be concentrating on Shannon entropy - also mentioning Rényi and Tsallis entropies - which concern information theory; therefore the application thus considered are appropriate to this. I will give a short overview of some of its applications; however, this is not an exhaustive list, since the application of entropy are extensive.

Wikipedia gives an appropriate overview of the applications, that the estimation of Shannon entropy is useful in "various science/engineering applications, such as independent component analysis, image analysis, genetic analysis, speech recognition, manifold learning, and time delay estimation" [29].

Independent component analysis (ICA), in signal processing, is a computational method for decomposing large, often very complex, multivariate data to find underlying/hidden factors or components. The computation of ICA depends on knowing the entropy of the sample; and in most cases this must be estimated, as an exact entropy is not always known. Kraskov, Stögbauer and Grassberger [17] discussed how estimating the mutual information (MI) using entropy estimators is useful for assessing the independence of components from ICA. Learned-Miller and Fisher [19] also presented another example of how to use estimation of entropy to obtain a new algorithm for the ICA problem.

Image analysis is the investigation of an image and the extraction of useful information. Hero and Michel [13] first discuss the applications of Rényi entropy in image processing, then Neemuchwala, Hero and Carson [23] discuss how in image analysis an important task is that of image retrieval, which uses entropy estimation to compute entropic similarities that are used to match a reference image to another image. Moreover Du, Wang, Guo and Thouin, [7] considered the importance of entropy-based image thresholding; using both Shannon and relative entropy.

Genetic analysis is the study and research of genes and molecules to find information on biological systems. Statistical analysis of specific cells can help us understand how genomic entropy can help diagnose diseases and cancers. Wieringen and Vaart, [27] discuss how chromosomal disorganisation increases as cancer progresses, they mention how the K-L estimator can be used to help find this disorganisation/entropy; thus finding that "as cancer evolves, and the genomic entropy increases, the transcriptomic entropy is also expected to surge".

"Speech recognition (SR) is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers" [30]. Shen, Hung and Lee [24] discuss how an entropy based algorithm can conduct accurate SR in noisy environments. Moreover, Kuo and Gao [18] focus on a method where the probability of a state or word sequence given an observation sequence is computed directly from the maximum entropy direct model.

There are many more applications of entropy, it is important to note the statistical applications of entropy; there are some tests on goodness-of-fit established by the estimation of entropy. Vasicek explored the test for normality; that its entropy exceeds that of any other distributions with the same variance [28]. Dudewicz and van der Meulen [8] discussed the property mentioned in section 2.1, that the uniform distribution maximises the entropy. Moreover, others have explored different distributions and their entropic properties; see [10, 4]

2.3 Other Estimators of Entropy

There are several estimation methods for the nonparametric estimation of the Shannon entropy of a continuous random sample. The paper *Nonparametric Entropy Estimation: An Overview* (J.Beirlant, E.Dudewicz, L.Gyorfi, E.van der Muelen, 2001) [2], gives an overview of the properties of these various methods. Also, the paper *Causality detection based on information-theoretic approaches in time series analysis* (K.Hlaváčková, M.Paluš, M.Vejmelka, and J.Bhattacharya, 2007) gives a more detailed look into these different types of estimators. I will outline a summary below to the types of estimators, which will lead us to understand why we choose the Kozachenko-Leonenko estimator for entropy.

First, I must set out the types of consistency, so we can see more obviously how it compares to the K-L estimator, for $X_1, ..., X_N$ a i.i.d sample from the distribution $f(X)$, where H_N is the estimator of $H(f)$. Then we have (as $N \to \infty$);

- Weak Consistency
$$H_N \xrightarrow{p} H(f) \tag{2.1}$$
- Mean Square Consistency
$$\mathbb{E}\{(H_N - H(f))^2\} \to 0 \tag{2.2}$$
- Strong Consistency
$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{2.3}$$
$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to \sigma^2 \tag{2.4}$$

  This is the type of consistency shown with the K-L estimator in Theorem 2.

The types of nonparametric estimators can be split into 3 categories; plug-in estimates, estimates based on sample-spacings and estimates based on nearest neighbour distances. The latter is the Kozachenko-Leonenko estimator, which is the main focus of this paper and will be explored in more detail in section 3.2.

The plug-in estimates [2], [14] are based upon a consistent density estimate f_N, of density f, which depends on the sample $X_1, ..., X_N$, I will consider two of these; the most obvious estimator of this type if the integral estimate of entropy. Given by;

$$H_N = -\int_{A_N} f_N(x) log(f_N(x)) dx \tag{2.5}$$

where the set A_N excludes the tail values of f_N. When the sample is from a 1-dimensional distribution, Dmitriev and Tarasenko, [6] for $A_N = [-b_N, b_N]$ and f_N the kernel density estimator; proved a strong consistency for this estimator. However, if f_N is not estimated in this form, due to the numeric integration, for dimensions $d \geq 2$, Joe [15] points out that this estimator is not practical and thus proposed the next plug-in estimator for entropy - the resubstitution estimator.

The resubstitution estimate is of the form;

$$H_N = -\frac{1}{N} \sum_{i=1}^{N} log(f_N(X_i)) dx \tag{2.6}$$

which was first proposed in 1976, by Ahmad and Lin [1] who showed the mean-square consistency of this estimator, where f_N is a kernel density estimate. Joe [15] then went on to obtain the asymptotic bias and variance, and whilst satisfying certain conditions reduced the mean square error. Moreover, Hall and Morton [12] went on to say that under more restrictive conditions we have strong consistency for 1-dimensional distributions; however, when $d = 2$ the root-n consistent estimator will have significant bias.

There are also two more plug-in estimates discussed in this paper; the splitting data and cross-validation estimates. Where in the first estimator, strong consistency is shown for a general dimension d, under some conditions on f. And in the latter estimator, strong consistency holds for a kernel estimate of f and for other estimates of f under some conditions we have root-n consistency when $1 \leq d \leq 3$.

Thus, so far the estimates for entropy looked at are only consistent whilst under strong conditions on f and f_N and mostly for a 1-dimensional distribution. So it is important to look at the next category of estimates - estimates of entropy based on sample-spacings; namely the m-spacing estimate. Sometimes it in not practical to estimate f_N, so this estimate is found based on spacings between the sample observations.

This estimator is only defined for samples of 1-dimension, where we assume $X_1, ..., X_N$ are an i.i.d sample, and let $X_{N,1} \leq X_{N,2} \leq ... \leq X_{N,N}$ be the corresponding ordered sample, then $X_{N,i+m} - X_{N,m}$ is the m-spacing.

Firstly we look at this estimator of the form, with fixed m;

$$H_{m,N} = \frac{1}{N} \sum_{i=1}^{N-m} log\left(\frac{N}{m}(X_{N,i+m} - X_{n,i})\right) - \Psi(m) + log(m) \tag{2.7}$$

where $\Psi(x)$ is the digamma function - more detailed explanation in section 3.2. For a sample from a uniform distribution this estimator has been shown to be consistent; proved by Tarasenko [26]. Under some conditions on f, on its boundedness, the weak consistency and asymptotic normality was shown by Hall [11].

To decrease the asymptotic variance of the estimator, we consider the estimator when $m_N \to \infty$, which is defined slightly differently;

$$H_{m,N} = \frac{1}{N} \sum_{i=1}^{N-m_N} log \left(\frac{N}{m_N} (X_{N,i+m_N} - X_{n,i}) \right) \tag{2.8}$$

for this estimator the weak and strong consistencies are proved under the assumption that as $N \to \infty$, $m_N \to \infty$ and $\frac{m_N}{N} \to 0$, for densities with bounded support.

The last category of estimators discussed by Beirlant, Dudewicz, Gyorfi and Muelen are those based on nearest neighbour distances. The main focus of my paper is on the Kozachenko-Leonenko estimator for entropy; which is the estimator covered in this section of their paper. I will not go into detail for this estimator now; however, I will mention that strong consistency holds for dimension $d \leq 3$, but higher dimensions can cause problems. Thus, it is important to note that recently a new estimator has been proposed by Berrerrt, Samworth and Yuan [3], formed as a weighted average of k-nearest neighbour estimators for different values of k. This estimator has shown promising results in higher dimensions, where under the same assumptions as for the K-L estimator, the strong consistency condition holds.

Chapter 3

Kozachenko-Leonenko Estimator

3.1 History

This estimator was first introduced by L.Kozachenko and N.Leonenko, in 1987, where they first published the article *Sample Estimate of the Entropy of a Random Vector* [20]. Using the nearest neighbour method, they created a simple estimator for the Shannon entropy of an absolutely continuous random vector from a independent sample of observations, to then establish conditions under which we have asymptotic unbiasedness and consistency.

Since then, there has been major developments in the estimator; firstly in 2007, N.Leonenko, L.Pronzato, V.Savani, proposed a similar alternative to this estimator in their paper *a Class of Renyi Information Estimators for Mulitdimensional densities* [21], this time using the k-nearest neighbour method, to consider estimators for the Rényi and Tsallis entropies. Then as the order of these entropies $q \to 1$, they defined the k-nearest neighbour estimator for the Shannon entropy, where k is fixed, and these estimators (under less rigorous conditions) are both consistent and asymptotically unbiased.

Moreover, in 2016, a new idea was proposed by T.Berrett, R.Samsworth and M.Yuan, written in *Efficient Mulitvariate Entropy Estimation via k-Nearest Neighbour Distances* [3]; that the value chosen for k, depends upon the sample size N. Also, this idea is then extended to a new estimator; "formed as a weighted average of Kozachenko-Leonenko estimators for different values of k". I will not be exploring this new estimator in depth; however, the understanding of the value of k depending on N will be examined in detail. Additionally, for $d = 1, 2, 3$, under some conditions on k, we are shown that the bias of the estimator acts in terms of $N^{-\frac{2}{d}}$; something which will also later be explored.

Lastly, also in 2016, S.Delattre and N.Fournier wrote the paper; *On the Kozachenko-Leonenko Entropy Estimator* [5], where they studied in detail the bias and variance of this estimator considering all 3 proposed values of k - $k = 1$,

k fixed or k depends on N. The also provided a development for the bias of this estimator when $k = 1$, in dimensions $d = 1, 2, 3$, in terms of $O(N^{-\frac{1}{2}})$, and in higher dimensions, in terms of powers of $N^{\frac{-2}{d}}$. This is an idea that will be considered in the focus of this paper; for $d = 1, 2$ to show how the bias acts for large N when $k = 1$.

3.1.1 Estimator with k=1

Firstly, I considered an article *On Statistical Estimation of Entropy of Random Vector* (N.Leonenko and L.Kozachenko, 1987) [20], which considers estimating the Shannon entropy of an absolutely continuous random sample of independent observations, with unknown probability density $f(x), x \in \mathbb{R}^d$. As $f(x)$ is unknown this is not easily estimated accurately for a random sample, and by just estimating the density $\hat{f}(x)$ to replace the actual density $f(x)$ in the formula for the entropy we get highly restrictive consistency conditions.

Therefore, the following estimator was proposed for the Shannon entropy of a random sample $X_1, X_2, ..., X_N$ of d-dimensional observations;

$$H_N = d\log(\bar{\rho}) + \log(c(d)) + \log(\gamma) + \log(N-1) \tag{3.1}$$

where $c(d) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$ is the volume of the d-dimensional unit ball, the Euler constant is $\log(\gamma) = \exp\left[-\int_0^\infty e^{-t}\log(t)dt\right] = -\Psi(1)$ and $\bar{\rho} = \left[\prod_{i=1}^N \rho_i\right]^{\frac{1}{N}}$, with ρ_i the nearest neighbour distance from X_i to another member of the sample X_j, $i \neq j$.

It is important to note that one can write the Euler constant $-\Psi(1) = \log(\exp(-\Psi(1))) = \log(\frac{1}{\exp(\Psi(1))})$, this notation is what is used in the latter papers, so it is useful to introduce it here. $\Psi(x)$ is the Digamma function, and when $x = 1$, this is just the negative Euler constant. Thus this estimator can be written if the form;

$$\begin{aligned}
H_N &= \log(\bar{\rho}^d) + \log(c(d)) - \Psi(1) + \log(N-1) \\
&= \log\left(\left[\prod_{i=1}^N \rho_i\right]^{\frac{d}{N}}\right)\log(c(d)(N-1)) + \log\left(\frac{1}{\exp(\Psi(1))}\right) \\
&= \frac{1}{N}\sum_{i=1}^N \log(\rho_i^d) + \log\left(\frac{c(d)(N-1)}{\exp(\Psi(1))}\right) \\
&= \frac{1}{N}\sum_{i=1}^N \log(\rho_i^d) + \frac{1}{N}\sum_{i=1}^N \log\left(\frac{c(d)(N-1)}{\exp(\Psi(1))}\right) \\
&= \frac{1}{N}\sum_{i=1}^N \log\left(\frac{\rho_i^d c(d)(N-1)}{\exp(\Psi(1))}\right) \qquad (3.2)
\end{aligned}$$

Under some conditions on the density function, this estimator is asymptotically unbiased and under stronger conditions it is also a consistent estimator for the Shannon entropy.

The estimator here is in a simple form, which is later developed into something more sophisticated, using the nearest neighbour method, but considering larger values of k (here $k = 1$). This estimator is developed so that the consistency and asymptotic unbias of the estimator holds under less constrained conditions.

3.1.2 Estimator with k fixed

The next paper I am exploring on estimation is *a Class of Renyi Information Estimators for Multidimensional densities* (N.Leonenko, L.Pronzato, V.Savani, 2007) [5], which looks at estimating the Rényi (H_q^*) and Tsallis (H_q) entropies, when $q \neq 1$, and the Shannon ($\hat{H}_{N,k,1}$) entropy. Where these are taken for a random vector $X \in \mathbb{R}^d$ with density function $f(x)$, by using the kth nearest neighbour method, with a fixed values of k.

For the Rényi and Tsallis entropies, this is achieved by considering the integral $I_q = \int_{\mathbb{R}^d} f^q(x)dx$, and generating its estimator, which is defined as $\hat{I}_{N,k,q} = \frac{1}{N}\sum_{i=1}^{N}(\zeta_{N,k,q})^{1-q}$. Where, $\zeta_{N,k,q} = (N-1)C_k V_d(\rho_{k,N-1}^{(i)})^d$, $V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$ is the volume of d-dimensional unit ball, $C_k = \left[\frac{\Gamma(k)}{\Gamma(k+1-q)}\right]^{\frac{1}{1-q}}$ and $\rho_{k,N-1}^{(i)}$ is the kth nearest neighbour distance from the observation X_i to some other X_j.

The estimator $\hat{I}_{N,k,q}$, provided $q > 1$ and I_q exists - and for any $q \in (1, k+1)$ if f is bounded - is thus found to be an asymptotically unbiased estimator for I_q. Also, provided $q > 1$ and I_{2q-1} exists - and for any $q \in (1, \frac{k+1}{2})$, when $k \geq 2$ if f is bounded - $\hat{I}_{N,k,q}$ is thus a consistent estimator for I_q.

Moreover, by simple formulas both the Rényi and Tsallis entropies can be written in terms of this estimated value;

$$\hat{H}_q^* = \frac{1}{1-q} log(\hat{I}_{N,k,q}) \tag{3.3}$$

$$\hat{H}_q = \frac{1}{q-1}(1 - \hat{I}_{N,k,q}) \tag{3.4}$$

thus, under the latter conditions, provide consistent estimates of these entropies as $N \to \infty$ for $q > 1$.

Furthermore, this paper goes on to discuss an estimator for the Shannon entropy, H_1 by taking the limit of the estimator for the Tsallis entropy, $\hat{H}_{N,k,q}$ as $q \to 1$, again with a fixed value of k. This estimator is similar to that proposed in 1987, equation 3.2; however, it is now extended from the nearest neighbour to the kth nearest neighbour;

$$\hat{H}_{N,k,1} = \frac{1}{N}\sum_{i=1}^{N}\log(\xi_{N,i,k}) \tag{3.5}$$

where $\xi_{N,i,k} = (N-1)\exp[-\Psi(k)]V_d(\rho^{(i)}_{k,N-1})^d$, with V_d and $\rho^{(i)}_{k,N-1}$ defined as in the estimation of I_q and the digamma function $\Psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$. The digamma function at $k = 1$ is given by $\Psi(1) = -\log(\gamma)$, the Euler constant, which was used for the $k = 1$ version of this estimator. Under the following less restrictive conditions; f is bounded and I_{q_1} exists for some $q_1 > 1$; then H_1 exists and the estimator $\hat{H}_{N,k,1}$ is a consistent estimator for the Shannon entropy. This means that for large N, we have $\hat{H}_{N,k,1} \overset{L_2}{\to} H$; which implies that as $N \to \infty$, both $N^{\frac{1}{2}}(\hat{H}_{N,k,1} - H) \overset{d}{\to} N(0, \sigma^2)$ - it's asymptotically efficient - and $\mathbb{E}(\hat{H}_{N,k,1}) \to H$ - it's asymptotically unbiased.

3.1.3 Estimator with k dependent on N

The last main paper, whose results I will be exploring is *Efficient Multivariate Entropy Estimation via k-Nearest Neighbour Distances* (T.Berrett, R.Samworth, M.Yuan, 2016) [3], which initially studies the K-L estimator, and the conditions under which it is efficient and asymptotically unbiased (for a value of k depending on the sample size N).

Considering dimensions $d \leq 3$, and a sample size N from distribution with density $f(x)$, they defined the k-nearest neighbour estimator of entropy - just as in section 3.1.2 - to be;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\rho^d_{(k),i}V_d(N-1)}{e^{\Psi(k)}}\right] \tag{3.6}$$

where $\rho_{(k),i}$, V_d and $\Psi(k)$ are all defined as in the 2007 paper. However, the difference here is in the conditions under which the estimator is consistent and asymptotically unbiased.

Here, some conditions on the finiteness of the α moment of f and the continuity and differentaibility of f are proposed, with $k \in \{1, ..., O(N^{1-\epsilon})\}$, for some $\epsilon > 0$, we have asymptotic unbias of the estimator; where the bias can be expressed as;

$$\mathbb{E}(\hat{H}_N) - H = O\left(max\left\{\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{N^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{\frac{\beta}{d}}}{N^{\frac{\beta}{d}}}\right\}\right) \qquad N \to \infty \tag{3.7}$$

Also, they considered the asymptotic normality of the estimator, given the α moment of f is finite (for $\alpha > d$), and some conditions on the continuity and differentaibility of f hold and with $k \in \{k_0, ..., k_1\}$. Then the variance of the estimator is given by;

$$Var(\hat{H}_{N,k}) = \frac{\sigma^2}{N} + o(\frac{1}{N}) \tag{3.8}$$

as $N \to \infty$, where $\sigma^2 = Var(log(f(x))$, and we define k_0, k_1 such that $\frac{k_0}{log^5(N)} \to \infty$ and $k_1 = O(N^\tau)$, where $\tau < \min\left\{\frac{2\alpha}{5\alpha+3d}, \frac{\alpha-d}{2\alpha}, \frac{4}{4+3d}\right\}$.

Moreover, T.Berrett, R.Samsworth and M.Yuan also go on to show that a consequence of the variance, given the dimension of the sample $d \leq 3$, with the same conditions, we have the asymptotic normality;

$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{3.9}$$

and

$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to \sigma^2 \tag{3.10}$$

where the estimator is asymptotically efficient and the asymptotic variance here is the best possible.

It is important to note that for higher dimensions ($d > 3$), these results do not necessarily hold; since I am just considering the specific dimensions $d = 1$ and $d = 2$, there is no need to detail this. However, they do then go on to discuss a more appropriate estimator for higher dimensions, given sufficient smoothness, which is efficient in arbitrary dimensions, which was previously mentioned in section **??**.

3.2 Focus of this Paper

I now wish to more explicitly introduce the Kozachenko-Leonenko estimator of the entropy H, in the form that I will be considering. 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;

  $$\rho_{(k),i} = \|X_{(k),i} - X_i\| \tag{3.11}$$

  denote the kth nearest neighbour or X_i.

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

  $$V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(1 + \frac{d}{2})} \tag{3.12}$$

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

  $$\Psi(k) = -\gamma + \sum_{j=1}^{k-1} \frac{1}{j} \tag{3.13}$$

  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;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\rho^d_{(k),i}V_d(N-1)}{e^{\Psi(k)}}\right] \tag{3.14}$$

where, $\rho^d_{(k),i}$ is defined in (3.11), V_d is defined in (3.12) and $\Psi(k)$ is defined in (3.13).

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;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{2\rho_{(k),i}(N-1)}{e^{\Psi(k)}}\right] \tag{3.15}$$

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;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{\pi\rho^2_{(k),i}(N-1)}{e^{\Psi(k)}}\right] \tag{3.16}$$

I will be looking at the asymptotic bias and variance of the estimator for different values of k, the main theorems I will be working by are those from section 3.1.3, where we have the conditions 1, 2 and 3, which imply the results stated by Theorems 1 and 2.

(NB: these conditions and theorems have been tweaked slightly to only explicitly consider distributions of dimension $d = 1, 2$, since the only distributions being considered in this paper are of dimension 1 or 2)

Condition 1 *(β) For density f bounded, denoting $m := \lfloor\beta\rfloor$ and $\eta := \beta - m$, we have that f is m times continuously differentiable and there exists $r_* > 0$ and a Borel measurable function g_* such that for each $t = 1, 2, ..., m$ and $\|y - x\| \leq r_*$, we have;*

$$\|f^{(t)}(x)\| \leq g_*(x)f(x)$$

,

$$\|f^{(m)}(y) - f^{(m)}(x)\| \leq g_*(x)f(x)\|y - x\|^\eta$$

and $sup_{x:f(x)\geq\delta}g_*(x) = o(\delta^{-\epsilon})$ as $\delta \downarrow 0$, for each $\epsilon > 0$.

Condition 2 *(α) For density $f(x)$ and dimension d, we have;*

$$\int_{\mathbb{R}^d} \|x\|^\alpha f(x)dx < \infty$$

Condition 3 *Assume that condition 1 holds for $\beta = 2$ and condition 2 holds for some $\alpha > d$. Let $k_0^* = k_{0,N}^*$ and $k_1^* = k_{1,N}^*$ denote two deterministic sequences of positive integers with $k_0^* \leq k_1^*$, with $\frac{k_0^*}{\log^5 N} \to \infty$ and with $k_1^* = O(N^\tau)$, where*

$$\tau < \min\left\{\frac{2\alpha}{5\alpha + 3d}, \frac{\alpha - d}{2\alpha}, \frac{4}{4 + 3d}\right\}$$

Theorem 1 (Asymptotic Unbiasedness) *Assume that conditions 1 and 2 hold for some $\beta, \alpha > 0$. Let $k^* = k_N^*$ denote a deterministic sequence of positive integers with $k^* = O(N^{1-\epsilon})$ as $N \to \infty$ for some $\epsilon > 0$. Then, for $d \leq 2$ (or $d \geq 3$) with $\beta \leq 2$ (or $\alpha \in (0, \frac{2d}{d-2})$), then for every $\epsilon > 0$ we have;*

$$\mathbb{E}(\hat{H}_N) - H = O\left(max\left\{\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{N^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{\frac{\beta}{d}}}{N^{\frac{\beta}{d}}}\right\}\right) \tag{3.17}$$

uniformly for $k \in \{1, ..., k^*\}$, as $N \to \infty$.

Theorem 2 (Efficiency and Consistency) *Assume that $d \leq 3$ and that condition 1 holds for $\beta = 2$ and condition 2 holds for some $\alpha > d$, then by condition 3 (where extra assumptions are made for $d = 3$), for the estimator $\hat{H}_{N,k}$ we have;*

$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{3.18}$$

and

$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to \sigma^2 \tag{3.19}$$

as $N \to \infty$ uniformly for $k \in \{k_0^*, ..., k_1^*\}$, where $\sigma^2 = Var(log(f(x))$, for density function $f(x)$. Thus, the estimator is asymptotically efficient and its asymptotic variance is the best attainable.

By the above, we can now say that $\hat{H}_{N,k}$ is an consistent and asymptotically unbiased estimator of exact entropy H; thus is a consistent estimator. This is due to using the central limit theorem, on the estimator for entropy $\hat{H}_{N,k}$, which states that;

$$\frac{\hat{H}_{N,k} - \mathbb{E}\hat{H}_{N,k}}{\sqrt{Var(\hat{H}_{N,k})}} \xrightarrow{d} N(0, \sigma^2)$$

where $\sigma^2 = Var(\log f(x))$. By section 3.1.3, we can assume that $Var(\hat{H}_{N,k}) = \frac{Var(\log f(x))}{N} + O(\frac{1}{N}) \approx \frac{1}{N}$, as for large N, the variance of the logarithm of the density function stays constant. Accordingly, the left side of the central limit theorem above can be written as;

$$\begin{aligned}\frac{\hat{H}_{N,k} - \mathbb{E}\hat{H}_{N,k}}{\sqrt{Var(\hat{H}_{N,k})}} &= \sqrt{N}(\hat{H}_{N,k} - \mathbb{E}\hat{H}_{N,k}) \\ &= \sqrt{N}[(\hat{H}_{N,k} - H) - (\mathbb{E}\hat{H}_{N,k} - H)] \\ &= \sqrt{N}(\hat{H}_{N,k} - H) - \sqrt{N}Bias(\hat{H}_{N,k})\end{aligned}$$

So we can see that from Theorem 2; $\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2)$ as $N \to \infty$. Whilst from Theorem 1 we have $Bias\hat{H}_{N,k} \to 0$ as $N \to \infty$. Thus as $N \to \infty$ this tends to the normal distribution, $N(0, \sigma^2)$, and the central limit theorem holds.

I will be exploring the bias is more detail later to see which one of the two ideas show to be more true for the behaviors of the bias for a large value of N, in dimension $d = 1$ or $d = 2$.

- With a fixed k, by [5], for $\beta \in (0, 2] \cap (0, d]$, we choose $a \in (0, \frac{\beta}{d}]$, then;

$$|Bias(\hat{H}_{N,k})| = O\left(\frac{1}{N^a}\right) \tag{3.20}$$

- With k depending on N, by [3], for $\beta \in (0, 2]$, we again choose $a \in (0, \frac{\beta}{d}]$, then;

$$|Bias(\hat{H}_{N,k})| = O\left(\left(\frac{k}{N}\right)^a\right) \tag{3.21}$$

Chapter 4

Monte-Carlo Simulations

In this chapter I will explore simulations of the bias of estimator (3.14) in comparison to the size of the sample estimated from, with respect to different values of k; by exploring 1-dimensional distributions and then progressing onto 2-dimensional. Firstly, the distributions considered will be analysed to determine if they satisfy the conditions 1, 2 and 3 stated for Theorems 1 and 2 to hold. Then, I will explore the estimator of entropy for simulations of samples from certain distributions, for different values of k.

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. Using Theorem 1, we can write that the bias of the estimator approaches 0 as $N \to \infty$. This is because we can write $Bias(\hat{H}_{N,k}) = \mathbb{E}(\hat{H}_{N,k}) - H$, which in equation (3.17) implies $Bias(\hat{H}_{N,k}) \to 0$ as $N \to \infty$. Thus, there must be a type of inverse relationship between the modulus of the bias of the estimator, $|Bias(\hat{H}_{N,k})|$, and N. We believe this relationship is of the form;

$$|Bias(\hat{H}_{N,k})| = \frac{c}{N^a} \tag{4.1}$$

for $a, c > 0$. By taking the logarithm of this, we can generate a linear relationship, which is easier to analyse, and is given by;

$$log|Bias(\hat{H}_{N,k})| \approx log(c) - a[log(N)] + \epsilon \tag{4.2}$$

where $\epsilon > 0$ is some small error term. I will investigate the consistency of this estimator for a sample from the normal distribution, dependent on the value of k, this mean finding the optimum value of k for which $|Bias(\hat{H}_{N,k})| \to 0$ for $N \to \infty$. For the relationship in equation (4.1), this will happen for larger values of a and relatively small c, as $N \to \infty$. As previously mentioned, there is evidence supporting that the bias becomes either of order $(\frac{1}{N})^a$ (equation (3.20)) or $(\frac{k}{N})^a$ (equation (3.21)). This leads to also examining the dependence 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 (1.1) 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 (1.1);

$$\begin{aligned}
H &= -\int_{x:f(x)>0} f(x) log(f(x))dx \\
&= -\int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) log\left[\frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)\right] dx \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right) \left(log(\sqrt{(2\pi)}\sigma) + \frac{x^2}{2\sigma^2}\right) \\
&= \frac{log(\sqrt{(2\pi)}\sigma)}{\sqrt{(2\pi)}\sigma} \int_{\mathbb{R}} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx + \frac{1}{2\sqrt{(2\pi)}\sigma} \int_{\mathbb{R}} \frac{x^2}{2\sigma^2} \exp\left(\frac{-x^2}{\sigma^2}\right) dx \\
&= log(\sqrt{(2\pi)}\sigma) + \frac{1}{2}
\end{aligned}$$

Thus the exact entropy for the normal distribution is given by

$$H = log(\sqrt{(2\pi e)}\sigma) \tag{4.3}$$

I will first explore samples from 1-dimensional standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$, $N(0, 1)$, to consider the behavior of the Kozachenko-Leonenko estimator. The exact entropy of this distribution is given by equation (4.3), with $\sigma^2 = 1$;

$$H = log(\sqrt{(2\pi e)}) \approx 1.418939 \tag{4.4}$$

Since, I am first considering the 1-dimensional normal distribution, the estimator takes the form in equation (3.15), which is given by;

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

4.1.1 Estimator Conditions

The normal distribution satisfies Theorem **??**, using that the density function is given by $f(x) = C \exp\left(\frac{-x^2}{2\sigma^2}\right)$ for $x \in \mathbb{R}$, given $\mu = 0$ and where $C := \frac{1}{\sqrt{(2\pi)}\sigma} >$

0. Then by equation (**??**), considering that here $d = 1$, taking some $\epsilon > 0$, the first condition of the Theorem is satisfied;

$$\begin{aligned}\int_{\mathbb{R}} |log(f(x))|^{1+\epsilon} f(x)dx = C \int_{\mathbb{R}} \left| log\,(C) - \left(\frac{x^2}{2\sigma^2}\right) \right|^{1+\epsilon} \exp\left(\frac{-x^2}{2\sigma^2}\right)dx \\ < \int_{\mathbb{R}} \frac{|x|^{1+\epsilon}}{\exp x^2}dx \\ < \int_{-\infty}^{\infty} \frac{|x|}{\exp x^2}dx < \infty\end{aligned}$$

Also, the second condition, equation (**??**), of Theorem **??** is satisfied;

$$\begin{aligned}\int_{(\mathbb{R})^2} |log(\|x-y\|)|^{1+\epsilon} f(x)f(y)dxdy = C^2 \int_{(\mathbb{R})^2} |log(\|x-y\|)|^{1+\epsilon} \exp\left(\frac{-(x^2+y^2)}{2\sigma}\right)dxdy \\ < \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{|log(\|x\|+\|y\|)|^{1+\epsilon}}{\exp\,(x^2+y^2)}dxdy \\ < \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{|log(\|x\|+\|y\|)|}{\exp\,(x^2+y^2)}dxdy < \infty\end{aligned}$$

Thus we can say that for the normal distribution, $\hat{H}_{N,k}$ is an asymptotically unbiased estimator for entropy.

Moreover, the normal distribution satisfies Theorem **??**, as it fulfills the first condition shown in equation (**??**);

$$\begin{aligned}\int_{\mathbb{R}} |log(f(x))|^{2+\epsilon} f(x)dx = C \int_{\mathbb{R}} \left| log\,(C) - \left(\frac{x^2}{2\sigma^2}\right) \right|^{2+\epsilon} \exp\left(\frac{-x^2}{2\sigma^2}\right)dx \\ < \int_{\mathbb{R}} \frac{|x|^{2+\epsilon}}{\exp x^2}dx \\ < \int_{-\infty}^{\infty} \frac{|x|^2}{\exp x^2}dx < \infty\end{aligned}$$

and the second condition, equation (**??**);

$$\begin{aligned}\int_{(\mathbb{R})^2} |log(\|x-y\|)|^{2+\epsilon} f(x)f(y)dxdy = C^2 \int_{(\mathbb{R})^2} |log(\|x-y\|)|^{2+\epsilon} \exp\left(\frac{-(x^2+y^2)}{2\sigma}\right)dxdy \\ < \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{|log(\|x\|+\|y\|)|^{2+\epsilon}}{\exp\,(x^2+y^2)}dxdy \\ < \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{|log(\|x\|+\|y\|)|^2}{\exp\,(x^2+y^2)}dxdy < \infty\end{aligned}$$

Henceforth, the estimator $\hat{H}_{N,k}$, for a sample from the normal distribution is a consistent estimator by Theorem **??**.

For Theorems **??** and **??** to be satisfied by the estimators generated by samples from the normal distribution, this distribution must meet the Conditions 1, 2 and 3. Firstly, to satisfy Condition 1, for density function $f(x) = \frac{1}{\sqrt{(2\pi)}} \exp\left(\frac{-x^2}{2}\right)$ for $x \in \mathbb{R}$, given $\mu = 0$ and $\sigma^2 = 1$, it must be such that;

- f is bounded - obvious, since for any probability distribution we always have $f(x) \geq 0$, additionally for the normal distribution we have that $f(x) = \frac{1}{\sqrt{(2\pi)}} \exp\left(\frac{-x^2}{2}\right) < 0.4$, $\forall x \in \mathbb{R}$. Hence, f is bounded above and below; so bounded.

- f is m-times differentiable - using Hermite polynomials, defined as;

  $$H_m(x) = (-1)^m e^{\frac{x^2}{2}} \frac{d^m}{dx^m}\left(e^{\frac{-x^2}{2}}\right)$$

  multiplying this by the coefficient in the distribution of $f(x)$, $\frac{1}{\sqrt{(2\pi}}$, we then get;

  $$\begin{aligned} \frac{d^m}{dx^m} f(x) &= \frac{H_m(x)}{(-1)^m} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \\ &= \frac{H_m(x)}{(-1)^m} f(x) \end{aligned}$$

  where $\frac{H_m(x)}{(-1)^m}$ is a polynomial; thus f is m-times differentiable.

- $\exists r_* > 0$ and a Borel measurable function g_*, with $\|y - x\| \leq r_*$ so that $\|f^{(t)}(x)\| \leq g_*(x)f(x)$ and $\|f^{(m)}(x) - f^{(m)}(x)\| \leq g_*(x)f(x)\|y - x\|^\eta$, for some g_* such that $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$.

  Since we are considering a 1-dimensional distribution, we can write the norms $\|\cdot\|$ as $|\cdot|$. Moreover, considering that for Theorems **??** and **??**, we have the value of $\beta = 2$ and since $m = \lfloor\beta\rfloor = \lfloor 2\rfloor = 2 = \beta$ and $\eta = \beta - m$, we have that $\eta = 0$. Thus we need $|f^{(t)}(x)| \leq g_*(x)f(x)$, which is obvious by above, in view of writing $|\frac{d^t}{dx^t} f(x)| = g_*(x)f(x)$, where we choose $g_*(x) = |\frac{H_t(x)}{(-1)^t}| = |H_t(x)|$, for $t = 1, 2, ..., m$, and $|f(x)| = f(x)$, since $f(x) > 0$. Also, g_* is a polynomial and is hence Borel measurable over $\mathbb{R}$, and for any polynomial we obviously have $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$. Additionally, we need $|f^{(m)}(x) - f^{(m)}(x)| \leq g_*(x)f(x)|y - x|^0 = g_*(x)f(x)$. We currently have;

  $$\begin{aligned} |f^{(m)}(x) - f^{(m)}(x)| &= \left|\frac{H_m(x)}{(-1)^m} f(x) - \frac{H_m(y)}{(-1)^m} f(y)\right| \\ &\leq \left|\frac{H_m(x)}{(-1)^m} f(x)\right| + \left|\frac{H_m(y)}{(-1)^m} f(y)\right| \\ &= g_*(x)f(x) + g_*(y)f(y) \\ &\leq g_*(x)f(x) \end{aligned}$$

  since we know that $f(x) > 0$ for all $x \in \mathbb{R}$, and $g_*(x) = |H_m(x)| > 0$, which is similar to the g_* before; thus satisfies the conditions for it.

Next, to satisfy Condition 2, for the density function f of the normal distribution, must fulfill that;

- The α-moment of f must be finite, so $\int_{\mathbb{R}^d} \|x\|^\alpha f(x)dx < \infty$ - this is always true for the normal distribution, all of its moments are finite, since they are defined with respect to σ^n, for some n, and $\sigma < \infty$.

Lastly, to satisfy Condition 3, we must find the values of k for which the estimator provides a uniform convergence for Theorems **??** and **??**. To do this we must have, for some $\alpha > d = 1$, let k_0^* and k_1^* denote two deterministic sequences of positive integers with $k_0^* \leq k_1^*$. Taking $\alpha := 2$, we must have;

- $k_1^* = O(N^\tau)$, where $\tau < \min\left\{\frac{2\alpha}{5\alpha+3d}, \frac{\alpha-d}{2\alpha}, \frac{4}{4+3d}\right\} = \min\left\{\frac{4}{13}, \frac{1}{4}, \frac{4}{7}\right\} = \frac{1}{4}$, so we can choose $\tau := \frac{2}{9} < \frac{1}{4}$ so that we have $k_1^* = O(N^{\frac{2}{9}})$
- $\frac{k_0^*}{\log^5 N} \to \infty$ - for this to be true we need to choose $k_0^* := N^A$ for some $A > 0$. Considering that $k_0^* \leq k_1^*$ and $k_1^* = O(N^{\frac{2}{9}})$, thus $A \in (0, \frac{2}{9})$. So we can choose $A := \frac{1}{16}$, which gives that $k_0^* = O(N^{\frac{1}{16}})$.

Thus, on account of the values of N being considered in the simulations; $N = 100, 200, ..., 50000$, we have that for the smallest $N = 100$, the values of k for which Theorem **??** and **??** both hold for are $k \in \{1, 2, 3\}$. Moreover, for the largest $N = 50,000$, we must consider $k \in \{2, 3, ..., 11\}$.

Overall, due to the above conditions for Theorems **??**, **??**, **??** and **??** being met, we can say that the Kozachenko-Leonenko estimator, of a sample from the 1-dimensional normal distribution is an asymptotically unbiased and consistent estimator for entropy, for certain values of k. We will explore values of $k \in \{1, 2, ..., 11\}$, more specifically focusing on $k = 1, 2, 3, 5, 10$ and exploring the other values in this range if needed to help make a conclusion.

4.1.2 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 4.1. We will then consider the relationship show in (4.2) for each sample and again work out the average for the values of a and c, shown in Figure 4.1.

Considering table 4.1, 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 be asymptotically unbiased, shown in Theorem **??**. 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

Table 4.1: *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

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 4.1, 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 $a_1 \approx 0.5054$ and $c_1 \approx 0.0433$ for the relationship shown in (4.1). Thus we have the relationship between the bias and N, for $k = 1$, to be of the form;

$$|Bias(\hat{H}_{N,1})| \approx \frac{0.0433}{N^{0.5054}}$$

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.

4.1.3 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;

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

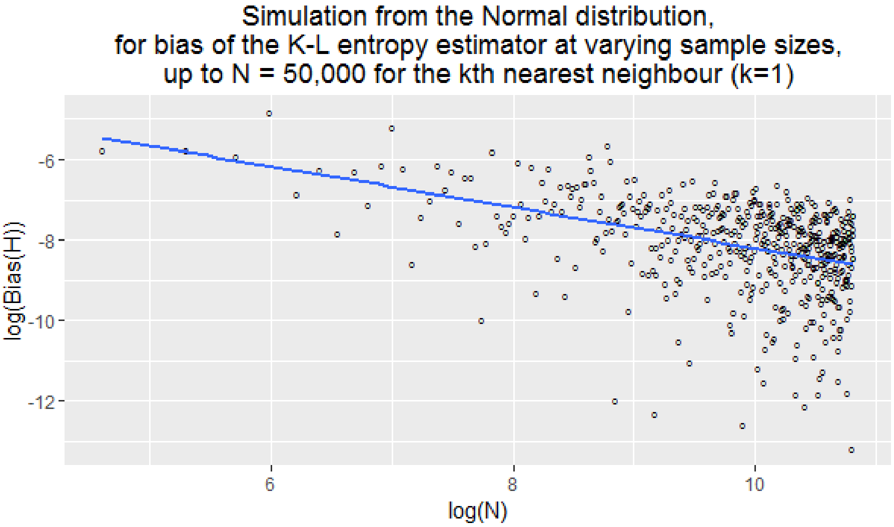

Figure 4.1: *Regression plot of* $\log|Bias(\hat{H}_{N,1})|$ *against* $\log(N)$

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 4.2.

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 **??**. 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 (4.1), for $k = 2$, which are given by $a_2 = 0.5490$ and $c_2 = 0.0459$, thus;

$$|Bias(\hat{H}_{N,2})| \approx \frac{0.0459}{N^{0.5490}}$$

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 4.1.10.

Table 4.2: *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

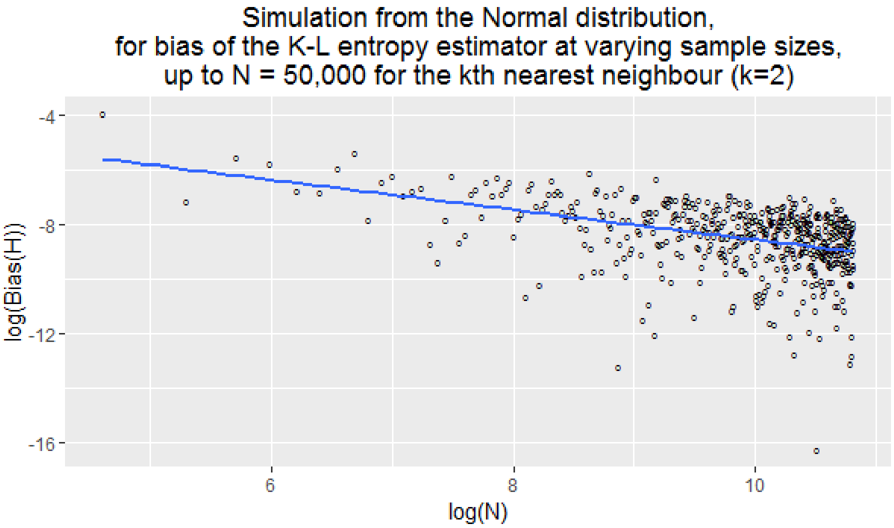

Figure 4.2: *Regression plot of* $\log|Bias(\hat{H}_{N,2})|$ *against* $\log(N)$

Table 4.3: *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

4.1.4 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 4.3. 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 4.2 is shown in figure 4.3. The have found the coefficients for the formula 4.1, for the graph shown with $k = 3$ are given by $a_3 = 0.6169$ and $c_3 = 0.0894$, thus;

$$|Bias(\hat{H}_{N,3})| \approx \frac{0.0894}{N^{0.6169}}$$

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 4.1.10.

4.1.5 k=4

Again, for $k = 4$, a similar analysis will be done with estimator of the form;

$$\hat{H}_{N,4} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{2\rho_{(4),i}(N-1)}{e^{\Psi(4)}}\right] = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{2\rho_{(4),i}(N-1)}{e^{-\gamma+1+\frac{1}{2}+\frac{1}{3}}}\right]$$

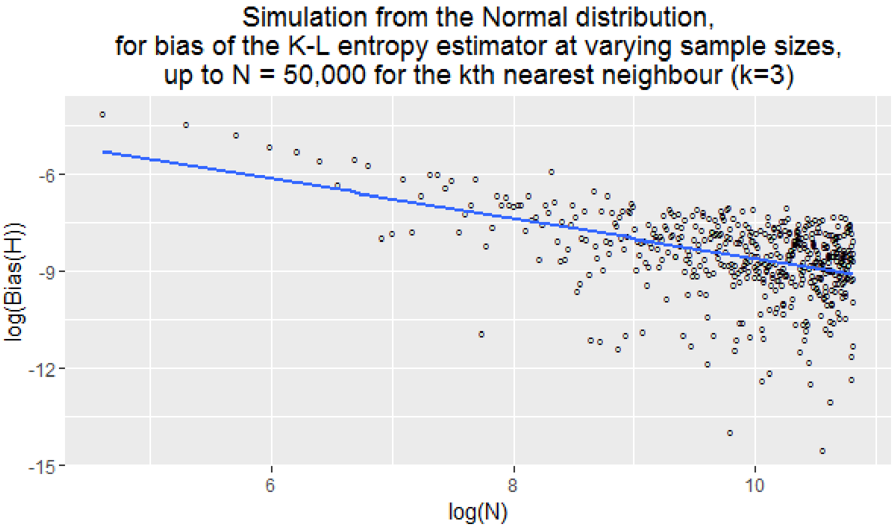

Figure 4.3: *Regression plot of* $\log|Bias(\hat{H}_{N,3})|$ *against* $\log(N)$

Table 4.4: *1-dimensional normal distribution,* $k = 4$

N	$\hat{H}_{N,4}$	$\lvert Bias(\hat{H}_{N,4})\rvert$	$Var(Bias(\hat{H}_{N,4}))$
TODO			

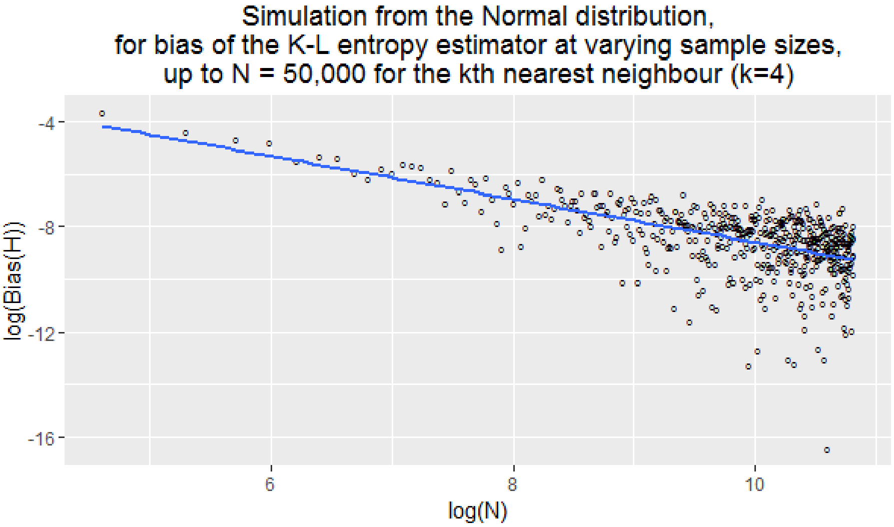

Figure 4.4: *Regression plot of* $\log|Bias(\hat{H}_{N,4})|$ *against* $\log(N)$

The results of the comparison between the actual value and the estimated value of entropy, for different values of N, are displayed in table 4.4. This shows again, that the Kozachenko-Leonenko estimator for entropy, $\hat{H}_{N,4} \to 0$, as $N \to \infty$, as more specifically $\hat{H}_{50000,4} \approx TODO$. However, comparing these results to those for $k = 1, 2, 3$, which had similar bias as $N \to \infty$, we can see that for larger N, the estimator with $k = 1, 2$ or 3 would be less appropriate to use, since the bias is slightly larger than for the estimator using $k = 4$.

Figure 4.4 contains the graph showing the relationship given by 4.2. The coefficients for the formula 4.1, for the graph shown with $k = 4$ are given by $a_4 = 0.8181$ and $c_3 = 0.6690$, thus;

$$|Bias(\hat{H}_{N,3})| \approx \frac{0.6690}{N^{0.8181}}$$

We here have that $a_4 \geq a_3 \geq a_2 \geq a_1$, hence, according to this analysis, when $k = 4$ we have yet again a stronger negative relationship between the bias and N. Considering the values of c_k for $k = 1, 2, 3, 4$, there has been an increase each time, however it is not a uniform increase, as there is a large jump between c_3 and c_4. This will be further discussed later in section 4.1.10.

Table 4.5: *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

4.1.6 k=5

Now we consider the estimator, shown in equation 3.14, 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, (4.4), gives table 4.5. 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 $50,000$ 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 4.5, 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 $k = 1, 2, 3, 4$.

However, if we consider the graph in figure 4.5, we can see an obvious negative relationship between the logarithm of $|Bias(\hat{H}_{N,5})|$ and the logarithm of

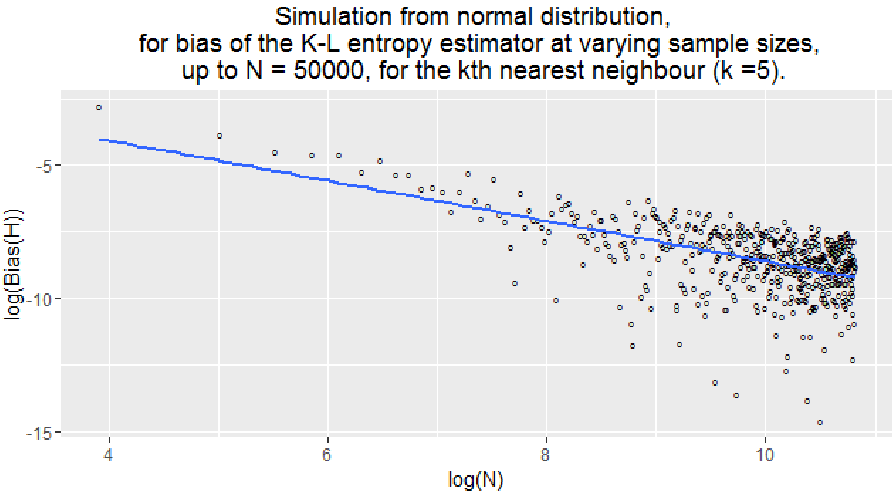

Figure 4.5: *Regression plot of* $\log|Bias(\hat{H}_{N,5})|$ *against* $\log(N)$

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 is asymptotically unbiased, as required for Theorem **??**.

This graph, in figure 4.5, gives the coefficients for the formula 4.1, for $k = 5$, which are $a_5 = 0.8486$ and $c_5 = 0.8235$, thus;

$$|Bias(\hat{H}_{N,5})| \approx \frac{0.8235}{N^{0.8486}}$$

For these coefficients we have that $a_5 \geq a_4 \geq a_3 \geq a_2 \geq a_1$, thus according to this analysis, we have a stronger consistency of our estimator for $k = 5$, in comparison to $k = 1, 2, 3, 4$. This comparison will be considered in more detail in section 4.1.10.

4.1.7 k=6

Again, for $k = 6$, we will again consider samples from the normal distribution with estimator of the form;

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

The results of the comparison between the actual value and the estimated value of entropy, for different values of N, are displayed in table 4.6. This shows that TODO...

Figure 4.6 contains the graph showing the relationship given by 4.2. The coefficients for the formula 4.1, for the graph shown with $k = 6$ are given by

Table 4.6: *1-dimensional normal distribution,* $k = 6$

N	$\hat{H}_{N,6}$	$\lvert Bias(\hat{H}_{N,6})\rvert$	$Var(Bias(\hat{H}_{N,6}))$
TODO			

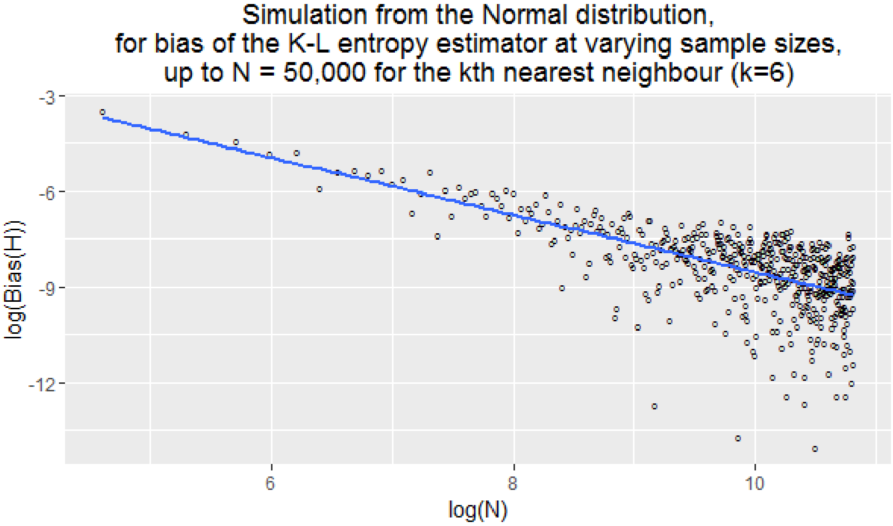

Figure 4.6: *Regression plot of* $\log|Bias(\hat{H}_{N,6})|$ *against* $\log(N)$

Table 4.7: *1-dimensional normal distribution,* $k = 7$

N	$\hat{H}_{N,6}$	$\vert Bias(\hat{H}_{N,6})\vert$	$Var(Bias(\hat{H}_{N,6}))$
TODO			

$a_6 = 0.8976$ and $c_3 = 1.5514$, thus;

$$|Bias(\hat{H}_{N,3})| \approx \frac{1.5514}{N^{0.8976}}$$

We here have that $a_6 \geq a_5 \geq ... \geq a_1$, again showing that the larger the k, the stronger the slope of the regressional relationship. This will be further discussed later in section 4.1.10.

4.1.8 k=7

For $k = 7$, the estimator is now of the form;

$$\hat{H}_{N,7} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{2\rho_{(7),i}(N-1)}{e^{\Psi(7)}}\right]$$

The results of the comparison between the actual value and the estimated value of entropy, for different values of N, are displayed in table 4.7. This shows that TODO...

Figure 4.7 contains the graph showing the relationship given by 4.2. The coefficients for the formula 4.1, for the graph shown with $k = 7$ are given by $a_7 = 0.9464$ and $c_3 = 2.3576$, thus;

$$|Bias(\hat{H}_{N,3})| \approx \frac{2.3576}{N^{0.9464}}$$

We here have that $a_7 \geq a_6 \geq ... \geq a_1$, showing that for $k = 7$, there is a steeper negative slope of the regressional relationship between $log|Bias(\hat{H}_{N,7})|$ and $log(N)$. This will be further discussed later in section 4.1.10.

TODO k=8 k=9

4.1.9 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;

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

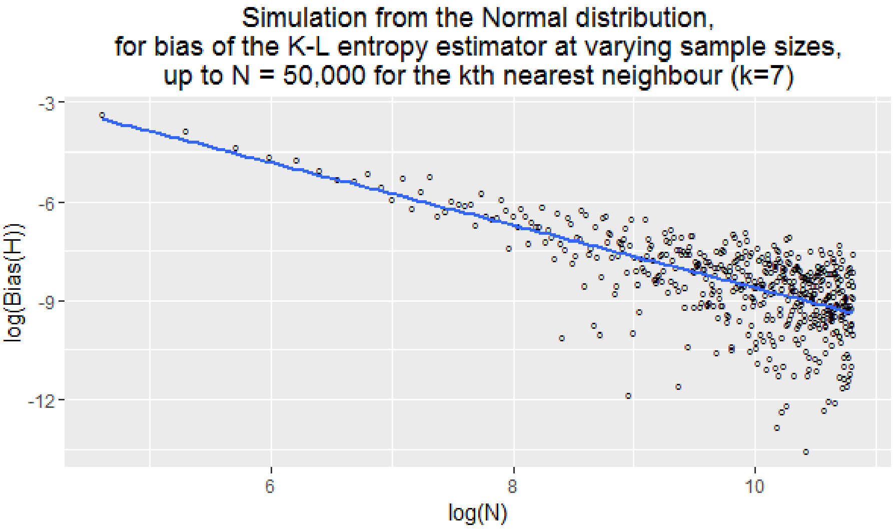

Figure 4.7: *Regression plot of* $\log|Bias(\hat{H}_{N,7})|$ *against* $\log(N)$

Table 4.8: *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

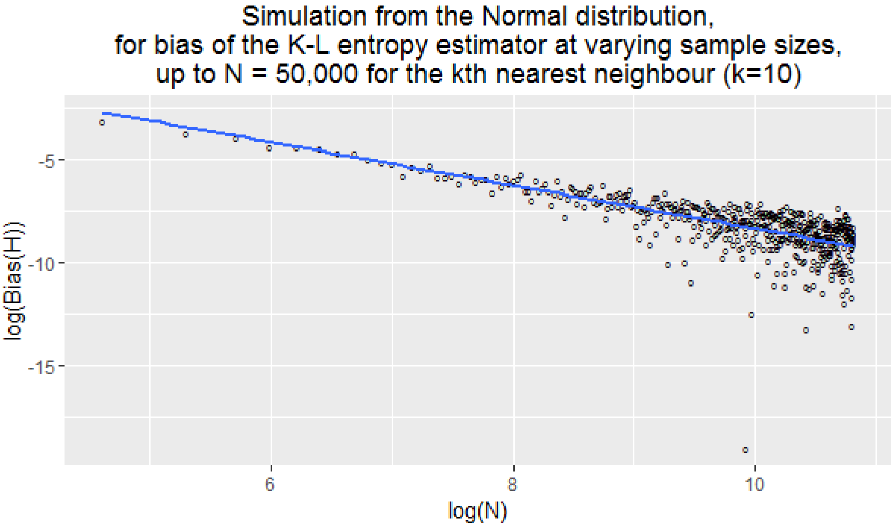

Figure 4.8: *Regression plot of* $\log|Bias(\hat{H}_{N,10})|$ *against* $\log(N)$

The results for the comparison between this estimator and 4.4 are displayed in table 4.8.

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

From the graph in Figure 4.8, we have the relationship between the Bias and N taking the form;

$$|Bias(\hat{H}_{N,5})| \approx \frac{8.5402}{N^{1.0454}}$$

So we have the coefficients of the regression formula (4.1) as $a_{10} = 1.0454$ and $c_k = 8.5402$. I will explore the meaning of these coefficients in more details in section 4.1.10.

TODO k=11

4.1.10 Comparison of k

The above analysis, sections 4.1.2 to 4.1.9 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.

TODO... what does this show?

The results shown in table 4.9 are inconclusive in determining which value of k generates the best estimator, with the smallest bias. However, these results

Table 4.9: *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
4				
5	0.00005917171	0.00003243503	0.00029451951	0.00001705529
6				
7				
8				
9				
10	0.0005825702	0.00003162161	0.0001488755	0.00001318863
11				

This table is comparing the values of $|Bias(\hat{H}_{N,k})|$ for the values of k explored in tables 4.1, 4.2, 4.3, 4.4, 4.5 4.6, 4.7, **??**, **??**, *4.8 and* **??** *with $N = 25,000$ and $N = 50,000$, when the estimator is taken over 500 samples*

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 $k = 1,2,3,5,10$ 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.

TODO .. what it now shows? The results in table 4.10, consider the scenario set out above, and 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 4.9. 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 in comparison to k is not necessarily informative. However, considering the variance of the bias of the estimator in comparison to N is informative. Theorem **??** states that the

Table 4.10: *1-dimensional normal distribution, comparison of k*

k	$\vert Bias(\hat{H}_{50000,k})\vert$	$Var(Bias(\hat{H}_{50000,k}))$
1	0.00013495546	0.00005116758
2	0.00012647214	0.00002868082
3	0.00003478968	0.00002299754
4		
5	0.00006034936	0.00001733369
6		
7		
8		
9		
10	0.00022455715	0.00001409080
11		

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*

variance of the estimator becomes $\approx \frac{Var(log(f(x))}{N}$ as $N \to \infty$, and previously I have stated that, if $Var(log(f(x))$ is constant, this becomes 0 for large N. This is consistent with the results found considering the variance of the bias, $Var|\hat{H}_{N,k} - H|$, instead of just variance of the estimator, $Var(\hat{H}_{N,k})$. But, since variance is not linear; $Var(X + a) = Var(X)$ for some $a \in \mathbb{R}$, we can say that $Var(\hat{H}_{N,k} - H) = Var(\hat{H}_{N,k})$, thus $Var|\hat{H}_{N,k} - H| \leq Var(\hat{H}_{N,k} - H) = Var(\hat{H}_{N,k})$. So if $Var(\hat{H}_{N,k}) \to 0$ as $N \to \infty$, then we also must have $Var|\hat{H}_{N,k} - H| \to 0$ as $N \to \infty$, which we have confirmed to be true numerically throughout this analysis.

Table 4.11, shows that as k runs from $1 \to 11$, we have that a_k and c_k both increase, with smooth values of a_k and a large jump between $k = 3$ and 4, and $k = 9$ and 10, in the value of c_k. 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 $|Bias(\hat{H}_{N,k})| \to 0$ 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 (4.1). Thus, considering a large sample size, say $N = 100,000$, we can find the bias of the Kozachenko-Leonenko estimator according to the

Table 4.11: *Comparison of coefficients of regression a_k and c_k from equation 4.1, for 1-dimensional normal distribution*

k	a_k	c_k
1	0.5054	0.0433
2	0.5490	0.0459
3	0.6169	0.0894
4	0.8181	0.6690
5	0.8486	0.8235
6	0.8976	1.5514
7	0.9464	2.3576
8	0.9574	3.2021
9	0.9883	4.4558
10	1.0454	8.5402
11	1.0386	8.7457

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

$$|Bias(\hat{H}_{100000,1})| \approx \frac{0.0433}{100000^{0.5054}} \approx 0.00012867314$$
$$|Bias(\hat{H}_{100000,2})| \approx \frac{0.0459}{100000^{0.5490}} \approx 0.00008256818$$
$$|Bias(\hat{H}_{100000,3})| \approx \frac{0.0894}{100000^{0.6169}} \approx 0.00007359317$$
$$|Bias(\hat{H}_{100000,4})| \approx \frac{0.6690}{100000^{0.8181}} \approx 0.00005431579$$
$$|Bias(\hat{H}_{100000,5})| \approx \frac{0.8235}{100000^{0.8486}} \approx 0.00004706127$$
$$|Bias(\hat{H}_{100000,6})| \approx \frac{1.5514}{100000^{0.8976}} \approx 0.00005043404$$
$$|Bias(\hat{H}_{100000,7})| \approx \frac{2.3576}{100000^{0.9464}} \approx 0.00004369886$$
$$|Bias(\hat{H}_{100000,8})| \approx \frac{3.2021}{100000^{0.9574}} \approx 0.00005229196$$
$$|Bias(\hat{H}_{100000,9})| \approx \frac{4.4558}{100000^{0.9883}} \approx 0.00005098304$$
$$|Bias(\hat{H}_{100000,10})| \approx \frac{8.5402}{100000^{1.0454}} \approx 0.00005063701$$
$$|Bias(\hat{H}_{100000,11})| \approx \frac{8.7457}{100000^{1.0386}} \approx 0.00005607827$$

This shows that the bias decreases slightly faster for a higher values of k. I have also compared these relationships through a graph of the regression lines found from plotting the simulations above, shown in Figure 4.9. From this we

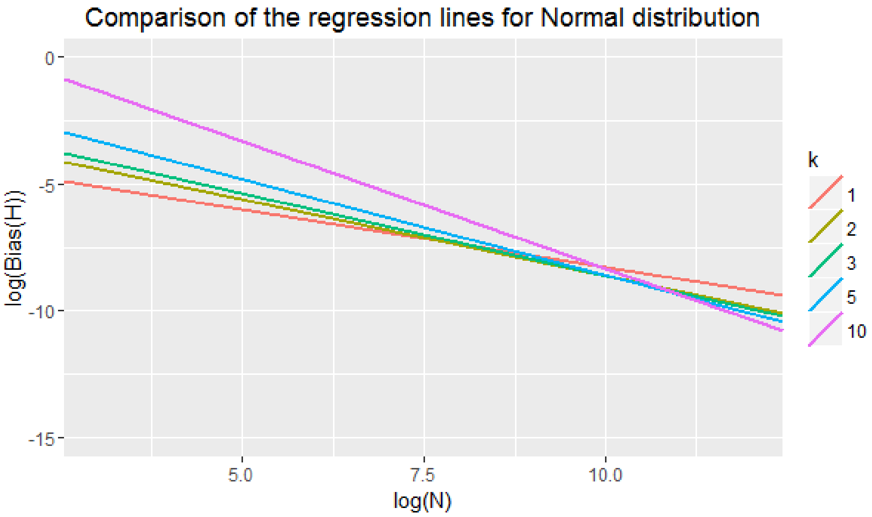

Figure 4.9: *Plot of regression lines for* $\log|Bias(\hat{H}_{N,k})|$ *against* $\log(N)$*, for* $k = 1, 2, 3, 5, 10$*, for samples from the normal distribution*

can see obviously that for $k = 10$, the regression line is steepest, indicating the strongest relationship between the logarithm of the bias and the logarithm of N. This implies that for a larger value of k, the estimator is stronger, to explore this I will plot the regression lines for some larger values of k to see if this is true.

I have also considered a plot of the values of a_k and c_k against k, to see if there is a clear relationship between the value of k and the coefficients, a_k and c_k. This relationship is depicted in figure **??**.

what does this show? ... TODO

4.2 1-dimensional Uniform distribution

I will now explore the entropy of samples from the 1-dimensional uniform distribution, $U[a, b]$. 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;

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

Using the definition of Shannon entropy given in equation (1.1), we can find the exact entropy for the uniform distribution;

$$\begin{aligned}
H &= -\int_{x:f(x)>0} f(x)log(f(x))dx \\
&= -\int_a^b \frac{1}{b-a} log\left[\frac{1}{b-a}\right] dx \\
&= -\frac{1}{b-a} log\left[\frac{1}{b-a}\right] \int_a^b dx \\
&= -log\left[\frac{1}{b-a}\right]
\end{aligned}$$

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

$$H = log[b-a] \tag{4.5}$$

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 of equation (3.15);

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

Moreover, the samples considered will not be from the standard uniform, but from the 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 as the pdf would be $f(x) = 1, \quad 0 \le x \le 1$, 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, I will be using the pdf $f(x) = 0.01, \quad 0 \le x \le 100$, which is from the $U[0, 100]$ distribution and gives the exact entropy to be;

$$H = log(100) \approx 4.605170 \tag{4.6}$$

4.2.1 Estimator Conditions

The uniform distribution satisfies Theorem **??**, firstly by considering equation (**??**) with $d = 1$ and the distribution $f(x) = C$ for $a \le x \le b$, where $C = \frac{1}{b-a} > 0$, we have for some $\epsilon > 0$;

$$\begin{aligned}
\int_{\mathbb{R}} |log(f(x))|^{1+\epsilon} f(x)dx &= C\int_a^b |log(C)|^{1+\epsilon}dx \\
&< \int_a^b |\hat{C}|^{1+\epsilon}dx < \infty
\end{aligned}$$

Also, the second condition, equation (**??**), of Theorem **??** is satisfied;

$$\int_{(\mathbb{R})^2} |log(\|x-y\|)|^{1+\epsilon} f(x)f(y)dxdy = C^2 \int_a^b \int_a^b |log(\|x-y\|)|^{1+\epsilon} dxdy$$
$$< \int_a^b \int_a^b |log(\|x\| + \|y\|)| dxdy < \infty$$

Thus we can say that for the normal distribution, $\hat{H}_{N,k}$ is an asymptotically unbiased estimator for entropy.

Moreover, the uniform distribution also satisfies Theorem **??**, as it fulfills the first condition shown in equation (**??**);

$$\int_{\mathbb{R}} |log(f(x))|^{2+\epsilon} f(x)dx = C \int_a^b |log(C)|^{2+\epsilon} dx$$
$$< \int_a^b \hat{C}^{2+\epsilon} dx < \infty$$

and the second condition, equation (**??**);

$$\int_{(\mathbb{R})^2} |log(\|x-y\|)|^{2+\epsilon} f(x)f(y)dxdy = C^2 \int_a^b \int_a^b |log(\|x-y\|)|^{2+\epsilon} dxdy$$
$$< \int_a^b \int_a^b |log(\|x\| + \|y\|)|^2 dxdy < \infty$$

For Theorems **??** and **??** to be satisfied by the estimators generated by samples from the uniform distribution, this distribution must meet the Conditions 1, 2 and 3. Firstly, to satisfy Condition 1, for the density function $f(x) = 0.01$ for $0 \leq x \leq 100$, it must be such that;

- f is bounded - obviously, since the density function for the normal distribution is constant for $x \in [a, b]$ and 0 otherwise; hence is bounded.
- f is m-times differentiable - TODO
- $\exists r_* > 0$ and a Borel measurable function g_*, with $\|y - x\| \leq r_*$ so that $\|f^{(t)}(x)\| \leq g_*(x)f(x)$ and $\|f^{(m)}(x) - f^{(m)}(x)\| \leq g_*(x)f(x)\|y - x\|^\eta$, for some g_* such that $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$

  TODO

Next, to satisfy Condition 2, for the density function f of the uniform distribution, must fulfill that;

- The α-moment of f must be finite, so $\int_{\mathbb{R}^d} \|x\|^\alpha f(x)dx < \infty$ - this is true, since for the 1-dimensional uniform distribution, $f(x)$ is constant; thus we would be integrating a polynomial $|x|^\alpha$, over a finite interval $a \leq x \leq b$, which is always finite.

Table 4.12: *1-dimensional uniform distribution, $k = 1$*

N	$\hat{H}_{N,1}$	$\|Bias(\hat{H}_{N,1})\|$	$Var(Bias(\hat{H}_{N,1}))$
100	4.600031	0.0051387799	0.02082714731
200	4.610892	0.0057214665	0.01051097236
500	4.606304	0.0011339740	0.00464501181
1000	4.604414	0.0007562145	0.00197979118
5000	4.606068	0.0008976195	0.00041910068
10000	4.604871	0.0002993139	0.00021349464
25000	4.605529	0.0003587332	0.00008599342
50000	4.605547	0.0003764863	0.00004437685

Lastly, to satisfy Condition 3, we must find the values of k for which the estimator provides a uniform convergence for Theorems **??** and **??**. These values are independent of the distribution that the sample is from, and only depends on the size of the sample, the dimension of the distribution that sample is taken from and the value chosen for α, where we have chosen $\alpha = 2$. Thus, the values of k found in section 4.1.1 are $\{1, 2, ..., 11\}$.

Due to the above conditions for Theorems **??**, **??**, **??** and **??** being met, we can say that the Kozachenko-Leonenko estimator, of a sample from the uniform distribution is an asymptotically unbiased and consistent estimator for entropy.

4.2.2 k=1

We will begin by considering 500 samples from the uniform distribution $U[0, 100]$, of size $N = 100 \to 50,000$ 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 (4.6) for different samples sizes N gives Table 4.12.

From table 4.12 there is an obvious decrease in value of the Bias, for larger N, with $|Bias(\hat{H}_{N,1})|$ decreasing from ≈ 0.005 to ≈ 0.0004. This decrease is similar to that considered in the normal distribution for $k = 1$, where $|Bias(\hat{H}_{N,1})|$ went from ≈ 0.01 to ≈ 0.0003; however, the difference is that for this distribution, the estimator seems to be more accurate for smaller N. Another thing to notice from this table is that, even through there is a general decrease, considering each rows in the table in comparison to the next does not necessarily show a decrease. The smallest value of bias occurs for $N = 10,000$, which does not correspond with the findings for the normal distribution. This could indicate a number of features;

- The values in table 4.12 contains outliers - this could be the case, since the numbers seem to jump around more ion this occasion than any others seen before. However, for $k = 1$ the bias decreases from $0.0057 \rightarrow 0.00029$ in the uniform distribution and decreases from $0.013 \rightarrow 0.00034$ in the normal distribution (not necessarily as N gets larger); these values of bias are not too dissimilar from one and other. Also, the tables are made from considering 500 samples of each N, finding the estimator in all cases, then taking the average of these as the actual estimator; this makes an outlier much less likely, as they would have been smoothed out from the averaging process.

- The actual value of entropy is significantly smaller itself, so the bias of the estimator is accordingly small - this is again unlikely, since the actual value of entropy for the uniform distribution is given by ≈ 4.605170 (4.6), and the for the normal distribution is ≈ 1.418939 (4.4). These values are not significantly different to one and other, also under this reasoning, one would expect the normal distribution to have an accordingly smaller bias than the uniform - but we are experiencing values the other way around.

- The estimator works better for samples from uniform than the normal distributions - this should not be true since the uniform distribution satisfies the conditions under which this estimator can be used in exactly the same manner as the normal distribution, so one would not expect samples from a specific distribution to yield a more accurate estimator for entropy. However, this is the most likely reason for the difference occurring between the two distributions. This is due to, as mentioned before, the nature of the uniform distribution, that by using $U[0, 100]$, for $N = 50,000$ each sample would be ≈ 0.002 distance apart. So using the nearest neighbour method; all of the data in the samples will have close neighbours, which could be the reason for the unreliable values shown in table 4.12.

To understand what is occurring in table 4.12, I have plotted the results of the approximate correlation between the bias of the estimator against N, as shown in equation 4.2. This is shown in Figure 4.10.

From this analysis, I have found the coefficients of the regression to be $a_1 = 0.3698$ and $c_1 = 0.0103$. On their own these coefficients show that the bias of the Koazchenko-Leonenko estimator for entropy has the following relationship with N;

$$|Bias(\hat{H}_{N,2})| \approx \frac{0.0103}{N^{0.3698}}$$

This is not what we would like to see, as for the estimator to be asymptotically unbiased, we would like to have $a_k > 0.5$, which here is not true. This could possibly be due to outliers in the samples, or could be due to the fact that for $k = 1$, the estimator for the entropy of a sample from the uniform distribution, is not as strong as expected. If we compare this to the values from the normal distribution; $a_1 = 0.4594$ and $c_1 = 0.0249$,we can see that for the normal distribution, the value for a_k is closer to that desired, but they have only an

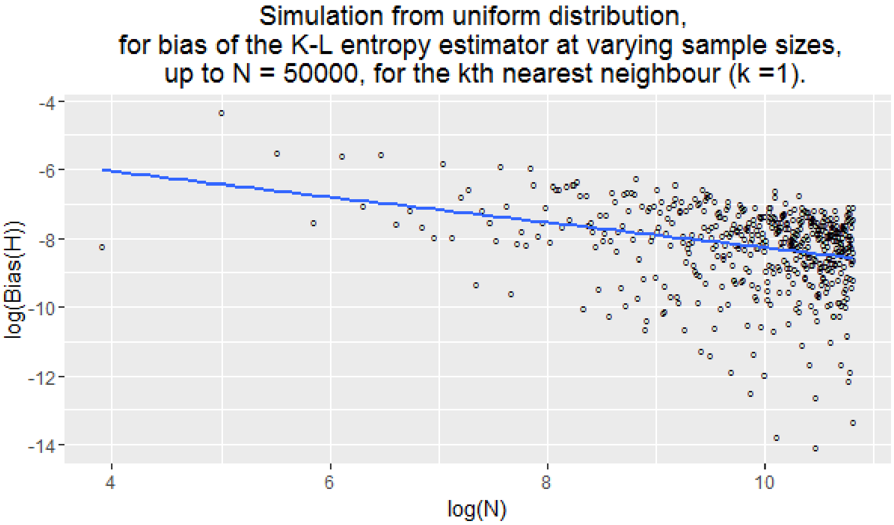

Figure 4.10: *Regression plot of* $\log|Bias(\hat{H}_{N,1})|$ *against* $\log(N)$

≈ 0.09 difference between them. To further understand the meaning of these coefficients we must compare them with those for higher values of k, which will be done in more detail in section 4.2.7.

4.2.3 k=2

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

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

Using the same parameters as before; taking 500 samples of size $N = 100 \to 50,000$ from the uniform, $U[0,100]$, distribution, we get the results in Table 4.13. These show that there is a general decrease in bias for a larger N from $|Bias(\hat{H}_{500,2})| \approx 0.006430$ to $|Bias(\hat{H}_{50000,2})| \approx 0.00031$; however, if we look closely, in a similar fashion to for $k = 1$, the bias is not always decreasing as N gets larger. This is shown in it increasing for the first 3 data points; $N = 100, 200, 500$, then decreasing for a bit, then increasing again for $N = 10000, 25000$. The smallest value of Bias actually occurs at $N = 5000$, which does not correspond with the results one would expect from this analysis. Moreover, the size of the bias does begin smaller for this case than it has done previously for values from the normal distribution, section 4.1.

To understand better what is occurring in table 4.13, I have plotted the results of the approximate correlation between the bias of the estimator against N, as shown in equation 4.1. This is shown in Figure 4.11.

Table 4.13: *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

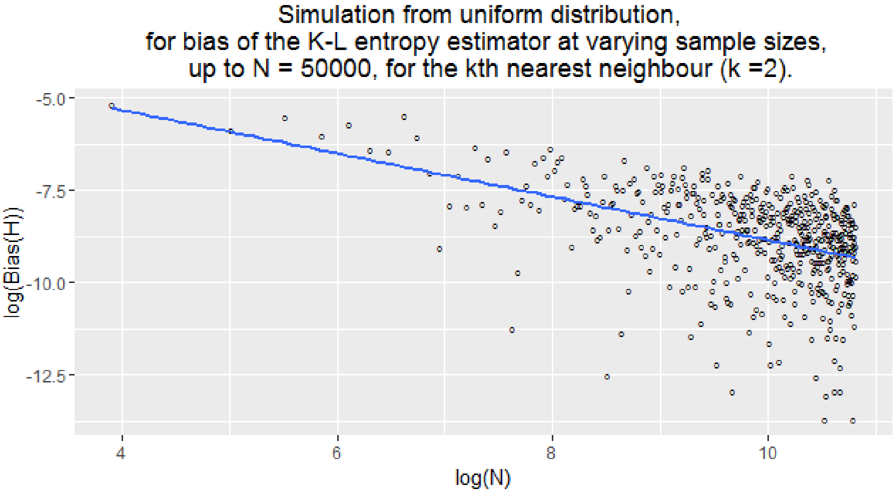

Figure 4.11: *Regression plot of* $\log|Bias(\hat{H}_{N,2})|$ *against* $\log(N)$

Table 4.14: *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

This graph has the regression line plotted of the form 4.2, with $a_2 \approx 0.5857$ and $c_2 \approx 0.00503$. From this analysis, I would expect the bias of the Koazchenko-Leonenko estimator for entropy to have the following relationship with N;

$$|Bias(\hat{H}_{N,2})| \approx \frac{0.0503}{N^{0.5857}}$$

As we can see from the graph, the relationship is obviously a negative correlation, and the values around the line are sparsely located. So I believe the reason for table 4.13 not looking as expected, is just due to bad luck in the values of N that I have chosen to be numerically represented in it. The graph plotted and the regression coefficients, align well with the normal distribution; whose coefficients $a_2 \approx 0.5998$ and $c_2 \approx 0.0746$ were found in section 4.1.3. Thus removing any uncertainty that we have about the estimator of entropy for a sample from the uniform distribution acting differently to that from the normal distribution. A comparison of the values of a_2 and c_2, with other values of k, will be further explored in section 4.2.7.

4.2.4 k=3

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

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

Using the same parameters as before; taking 500 samples of size $N = 100 \rightarrow 50,000$ from the uniform, $U[0, 100]$, distribution, we get the results in Table 4.14. These results are again inconclusive, to showing a the consistency condition that $|Bias(\hat{H}_{N,3})| \rightarrow 0$ as $N \rightarrow \infty$, since the numbers jump around and increase between, say $N = 500$ and 1000, which we would not expect to happen.

I believe that the best way to show this consistency condition is to just consider the graphical representation, and to not worry about the tabulated

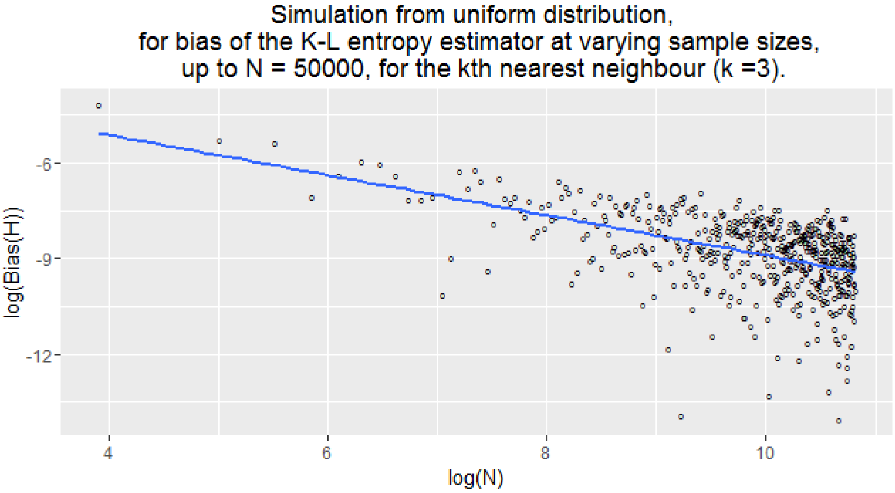

Figure 4.12: *Regression plot of* $\log |Bias(\hat{H}_{N,3})|$ *against* $\log(N)$

values, as they're inconsistent due to the reasons stated in section 4.2.3. From this plot, Figure 4.12, I have found that $a_3 \approx 0.6291$ and $c_3 \approx 0.0737$. This implies the relationship;

$$|Bias(\hat{H}_{N,3})| \approx \frac{0.0737}{N^{0.6291}}$$

In comparison to this, for $k = 2$, we had that $a_2 \approx 0.5857$ and $a_1 \approx 0.3698$, so $a1 < a_2 < a_3$, which implies that the relationship for $k = 3$ is stronger than that for $k = 1, 2$. We can also compare the values found in section 4.1.4, where samples from the normal distribution were considered and an estimator for $k = 3$ was found. The coefficients here were given by $a_3 \approx 0.6443$ and $c_3 \approx 0.1156$, which is a ≈ 0.015 difference in a_3, and a ≈ 0.04 difference in c_3, in comparison to the uniform distribution. This implies a very similar relationship is shown for each distribution; moreover, a more detailed comparison will be further explored in section 4.2.7.

4.2.5 k=5

Now we consider the estimator, shown in equation 3.14, 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 estimator to the exact value of entropy, shown in equation (4.6), for 500 samples of size N, as before, we get an inconclusive result from table 4.15. This is again due to the fact that taking a larger number of samples

Table 4.15: *1-dimensional uniform distribution,* $k = 5$

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

around a relatively small interval, will give similar results; hence, the relationship is again not as obvious from this.

Furthermore, I have considered a plot to represent equation (4.2), shown in Figure 4.13, where I have also found the coefficients of regression to be; $a_5 \approx 0.7501$ and $c_5 \approx 0.1889$.

Thus, here we get the relationship;

$$|Bias(\hat{H}_{N,5})| \approx \frac{0.1889}{N^{0.7501}}$$

In comparison to the coefficients from the previous value of k, we find that $a_1 < a_2 < a_3 < a_5$, indicating that as k increases the strength of the relationship between the $|Bias(\hat{H}_{N,k})|$ and N also increases. This implies that $|Bias(\hat{H}_{N,k})| \to 0$ as $N \to \infty$ faster for $k = 5$, over a smaller k. We can also see that c_k, so far, has also been increasing for larger $k \leq 5$. We can also compare this to the information found form the analysis of the normal distribution, with $k = 5$, explored in section 4.1.6. Here we found that $a_5 = 0.7568$ and $c_5 = 0.3557$, which shows an ≈ 0.06 difference in the value of a_5 and a ≈ 0.17 difference in the values of c_5. These show small differences between the two distributions, which will be further explored in section 4.2.7.

4.2.6 k=10

The last estimator for the entropy of a sample from the 1-dimensional uniform 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 4.6 are displayed in table 4.16.

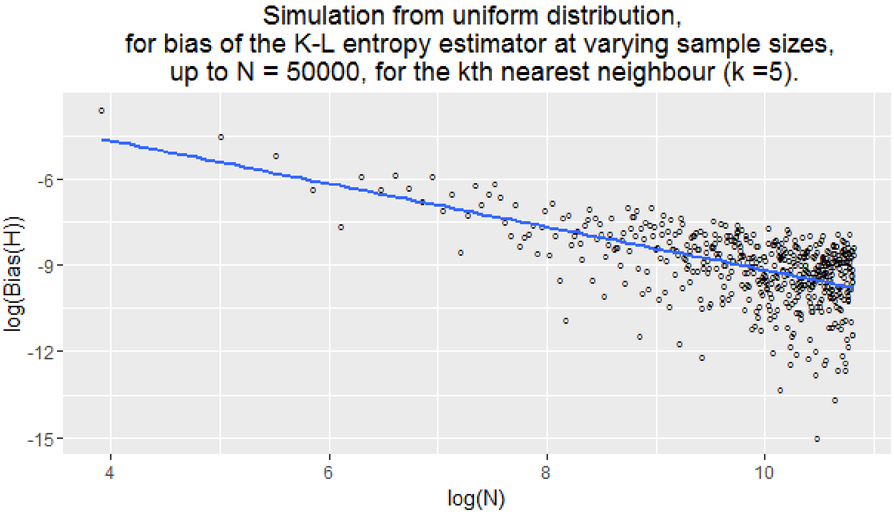

Figure 4.13: *Regression plot of* $\log|Bias(\hat{H}_{N,5})|$ *against* $\log(N)$

Table 4.16: *1-dimensional uniform distribution,* $k = 10$

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

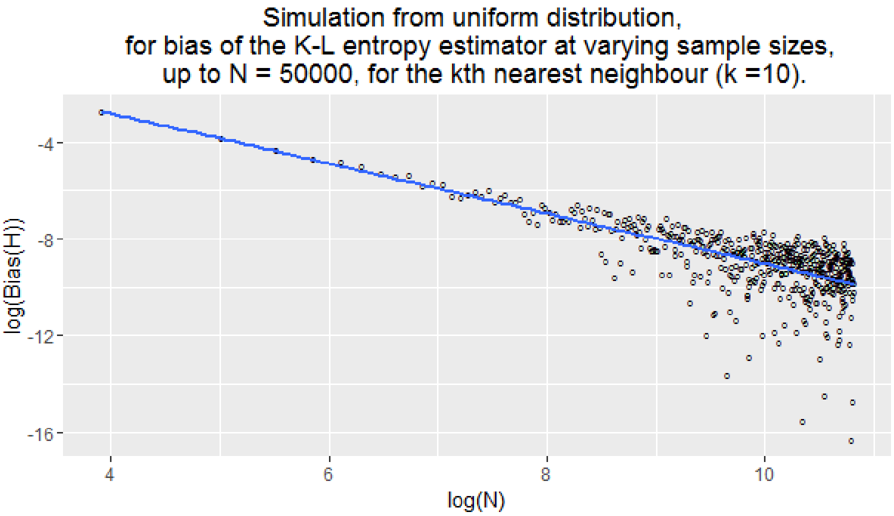

Figure 4.14: *Regression plot of* $\log|Bias(\hat{H}_{N,10})|$ *against* $\log(N)$

In comparison to what we have seen before in tables 4.13, 4.14 and 4.15, we here have more of a obvious comparison, shown numerically, between the size of the sample, N, and the size of the bias. This fits in with the condition of asymptotic unbiasedness from Theorem **??**, that $|Bias(\hat{H}_{N,10})| \to 0$ as $N \to \infty$.

From plotting the logarithm of the bias of the estimator against the logarithm of N I have found the coefficients of regression to be $a_{10} = 1.0357$ and $c_{10} = 3.8217$. Again these values are both larger than recorded previously for smaller values of k, for samples from the uniform distribution. Most importantly; $a_{10} > a_5 > a_3 > a_2 > a_1$. These coefficients imply a relationship between the bias of the estimator and N of the form;

$$|Bias(\hat{H}_{N,10})| \approx \frac{3.8217}{N^{1.0357}}$$

This implies that the value of $|Bias(\hat{H}_{N,10})| \to 0$ for large N, in a faster manner than that for smaller k, which again is fitting with Theorem **??** on the asymptotic unbiasedness of the estimator. I can also make the comparison between the regressional relationship between the uniform and the normal distribution for $k = 10$, where for the normal distribution I found the coefficients to be $a_{10} = 1.0055$ and $c_{10} = 5.5942$, in section 4.1.9. We can see that these regression lines would be close to one and other, which is to be expected. For both distributions the value of a_{10} is only ≈ 0.03 in difference, but the value of c_{10} has a larger difference of ≈ 1.8, which is a much larger distance and gives the impression of the two lines being parallel to one and other; the uniform regression line is just shifted down slightly from the normal. A more detailed comparison of this will be shown in section 4.2.7.

Table 4.17: *1-dimensional uniform distribution, a comparison of k*

k	a_k	c_k
1	0.3698	0.0103
2	0.5857	0.0503
3	0.6291	0.0737
5	0.7501	0.1889
10	1.0357	3.8217

4.2.7 Comparison of k

In sections 4.2.2 to 4.2.6, I have explored the Koazchenko-Leonenko estimator for samples from the 1-dimensional uniform distribution. In the most part, the tables of information from this estimator were inconsistent and inconclusive, due to the nature of the uniform distribution; that the samples are very close to one and other. Because of this, I will not be going into more detail of that comparison; thus, will be focusing solely on the relationship shown in equation (4.1);

$$|Bias(\hat{H}_{N,k})| = \frac{c_k}{N^{a_k}}$$

The results from the investigation above have been condensed into table 4.17, showing the change in values of a_k and c_k for different k.

From table 4.17, we can see that as k in creases from 1 to 10, that both a_k and c_k also increase. The value of a_k increasing implies an increase in strength of the asymptotic unbiasedness of the estimator, Theorem **??**; $N(H - \mathbb{E}\hat{H}_{N,k})^2 \to 0 \quad (N \to \infty)$, which is equivalent to saying that $|Bias(\hat{H}_{N,k})| \to 0 \quad (N \to \infty)$. Thus, considering a large sample size, say $N = 100,000$, we can find the bias of the Kozachenko-Leonenko estimator according to the regressional relationships found in sections 4.2.2 to 4.2.6 for each k;

$$|Bias(\hat{H}_{100000,1})| \approx \frac{0.0103}{100000^{0.3698}} \approx 0.000145826$$
$$|Bias(\hat{H}_{100000,2})| \approx \frac{0.0503}{100000^{0.5857}} \approx 0.000059301$$
$$|Bias(\hat{H}_{100000,3})| \approx \frac{0.0737}{100000^{0.6291}} \approx 0.000052719$$
$$|Bias(\hat{H}_{100000,5})| \approx \frac{0.1889}{100000^{0.7501}} \approx 0.000033553$$
$$|Bias(\hat{H}_{100000,10})| \approx \frac{3.8217}{100000^{1.0357}} \approx 0.000025337$$

These values confirm our original thoughts that the larger value of $k \leq 10$ gives a more consistent estimator. Moreover, we can compare these values to those found for the standard normal distribution, along with comparing the values of

Table 4.18: *Comparison between 1-dimensional Uniform and Normal distribution*

—	Normal			Uniform		
k	a_k	c_k	$\|Bias(\hat{H}_{100000,k})\|$	a_k	c_k	$\|Bias(\hat{H}_{100000,k})\|$
1	0.4594	0.0249	0.00012566	0.3698	0.0103	0.000145826
2	0.5998	0.0746	0.00007477	0.5857	0.0503	0.000059301
3	0.6443	0.1156	0.00006942	0.6291	0.0737	0.000052719
5	0.7568	0.3557	0.00005949	0.7501	0.1889	0.000033553
10	1.0055	5.5942	0.00005251	1.0357	3.8217	0.000025337

a_k and c_k, to see if the estimator has a similar accuracy for both distributions, this comparison is shown in table 4.18.

In this comparison we can see that, the values of a_k and $|Bias(\hat{H}_{100000,k})|$ are similar for both distributions, with a_k varying by less that ≈ 0.015 for $k = 2, 3, 5, 10$ and $|Bias(\hat{H}_{100000,k})|$ varying by less than ≈ 0.00002 for all k. Both of these confirm that the approximation to the relationship between the estimator and the actual value of entropy, is a good approximation to make; one which is consistent through both distributions considered so far.

Another comparison that we can make, is by considering the plot of the regression lines for the logarithm of $|Bias(\hat{H}_{N,k})|$ against the logarithm of N, for all values of k. This relationship is shown in Figure 4.15. This graph shows obviously that the larger the k that we have considered the stronger the relationship between the bias and N, since the regression line for $k = 10$ is the steepest out of them all.

4.3 1-dimensional Exponential Distribution

I will now be looking at the entropy of samples from the exponential distribution $exp(\lambda)$, where $\lambda > 0$ is the rate or inverse scale parameter. In a similar fashion to the previous distributions, the exponential also has an exact formula for the entropy, given the rate parameter λ. Using equation (1.1) and the density function for the exponential distribution $f(x) = \lambda e^{-\lambda x}$ for $x \in [0, \infty)$ and for

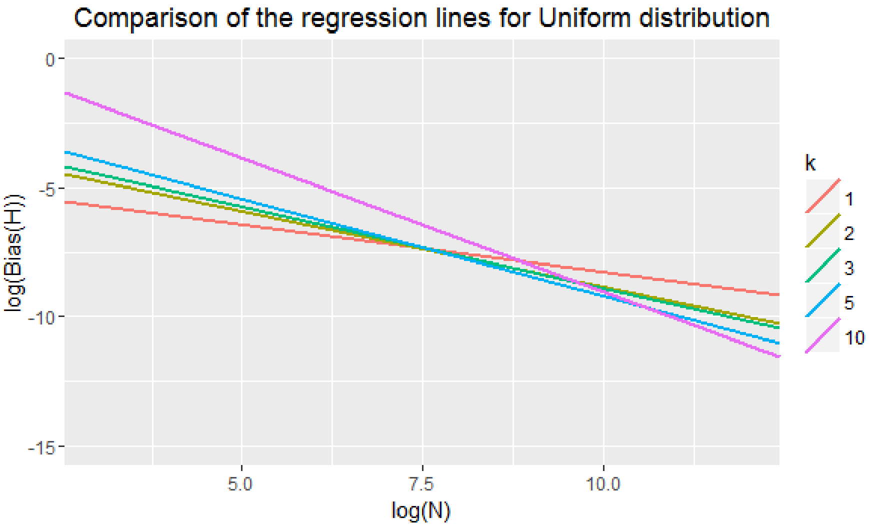

Figure 4.15: *Plot of regression lines for* $\log|Bias(\hat{H}_{N,k})|$ *against* $\log(N)$*, for* $k = 1, 2, 3, 5, 10$*, for samples from the uniform distribution*

$\lambda > 0$, we can write the exact entropy;

$$\begin{aligned}
H &= -\int_{x:f(x)>0} f(x)log(f(x))dx \\
&= -\int_0^\infty \lambda e^{-\lambda x} log[\lambda e^{-\lambda x}]dx \\
&= -\lambda \int_0^\infty \lambda e^{-\lambda x}[log(\lambda) - \lambda x]dx \\
&= \lambda \int_0^\infty \lambda e^{-\lambda x} - log(\lambda)e^{-\lambda x}dx \\
&= -\lambda \left[xe^{-\lambda x}\right]_0^\infty + \int_0^\infty \lambda e^{-\lambda x}dx + log(\lambda)\left[e^{-\lambda x}\right]_0^\infty \\
&= 0 + (log(\lambda) - 1)\left[e^{-\lambda x}\right]_0^\infty \\
&= -(log(\lambda) - 1)
\end{aligned}$$

Thus we have the the exact value of entropy for the exponential distribution, given the rate parameter $\lambda > 0$, is;

$$H = 1 - log(\lambda) \tag{4.7}$$

Moreover, I am again considering a 1-dimensional distribution; thus $V_d =$

$V_1 = 2$, and;

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

is the form of the Kozachenko-Leonenko estimator that I will be considering here, equation (3.15).

I have decided to choose to explore the exponential distribution with rate parameter $\lambda = 0.5$, this is because, we know that for the exponential distribution we must have the rate parameter $\lambda > 0$ and if I choose $\lambda > e \approx 2.7183$ we get a negative values of entropy, $H < 0$. This will introduce problems when considering the modulus of the bias; hence, for this analysis it will be more beneficial to consider a positive value of entropy, H. Also, for $\lambda \geq 1$, we have a very small value of entropy, $H \leq 1$, which would cause problems when calculating large samples and their entropy. Therefore, I have chosen a random number for the rate parameter such that $\lambda \in (0,1)$, so for this value of $\lambda = 0.5$, the exact entropy is given by;

$$H = 1 - log(0.5) \approx 1.693147 \tag{4.8}$$

4.3.1 Estimator Conditions

Samples from the exponential distribution must satisfy the conditions of Theorem **??** and **??**, to be an asymptotically unbiased and consistent estimator. For Theorem **??** to be satisfied, it must first satisfy equation (**??**), thus using that the density here is $f(x) = \lambda e^{-\lambda x}$ for $x \in [0,\infty)$, where $\lambda > 0$, and considering some $\epsilon > 0$;

$$\begin{aligned}\int_{\mathbb{R}} |log(f(x))|^{1+\epsilon} f(x)dx &= \lambda \int_0^\infty |log(\lambda) - \lambda x|^{1+\epsilon} e^{-\lambda x} dx \\ &< \int_0^\infty \frac{|-\lambda x|^{1+\epsilon}}{e^{\lambda x}} dx \\ &< \int_0^\infty \frac{|x|}{e^{\lambda x}} dx < \infty\end{aligned}$$

Also, the second condition, equation (**??**), of Theorem **??** is satisfied;

$$\begin{aligned}\int_{(\mathbb{R})^2} |log(\|x-y\|)|^{1+\epsilon} f(x)f(y)dxdy &= \lambda^2 \int_0^\infty \int_0^\infty |log(\|x-y\|)|^{1+\epsilon} e^{-\lambda(x+y)} dxdy \\ &< \int_0^\infty \int_0^\infty \frac{|log(\|x\|+\|y\|)|^{1+\epsilon}}{e^{\lambda(x+y)}} dxdy \\ &< \int_0^\infty \int_0^\infty \frac{|log(\|x\|+\|y\|)|}{e^{(x+y)}} dxdy < \infty\end{aligned}$$

Thus we can say that for the exponential distribution, $\hat{H}_{N,k}$ is an asymptotically unbiased estimator for entropy. Moreover, the exponential distribution also satisfies Theorem **??**, as it fulfills the first condition shown in equation (**??**);

$$\begin{aligned}\int_{\mathbb{R}} |log(f(x))|^{2+\epsilon} f(x)dx &= \lambda \int_0^\infty |log(\lambda) - \lambda x|^{2+\epsilon} e^{-\lambda x} dx \\ &< \int_0^\infty \frac{|-\lambda x|^{2+\epsilon}}{e^{\lambda x}} dx \\ &< \int_0^\infty \frac{|x|^2}{e^{\lambda x}} dx < \infty\end{aligned}$$

and the second condition, equation (**??**);

$$\begin{aligned}\int_{(\mathbb{R})^2} |log(\|x-y\|)|^{2+\epsilon} f(x)f(y)dxdy &= \lambda^2 \int_0^\infty \int_0^\infty |log(\|x-y\|)|^{2+\epsilon} e^{-\lambda(x+y)} dxdy \\ &< \int_0^\infty \int_0^\infty \frac{|log(\|x\| + \|y\|)|^{2+\epsilon}}{e^{\lambda(x+y)}} dxdy \\ &< \int_0^\infty \int_0^\infty \frac{|log(\|x\| + \|y\|)|^{2}}{e^{(x+y)}} dxdy < \infty\end{aligned}$$

Thus, the estimator $\hat{H}_{N,k}$, for a sample from the exponential distribution is a consistent estimator by Theorem **??**.

For Theorems **??** and **??** to be satisfied by the estimators generated by samples from the exponential distribution, this distribution must meet the Conditions 1, 2 and 3. Firstly, to satisfy Condition 1, the density function $f(x) = \frac{e^{\frac{-x}{2}}}{2}$ for $x \in [0, \infty)$, where we have chosen $\lambda = 0.5$, must be such that;

- f is bounded - this is true, since for any probability distribution we have $f(x) \geq 0$, also for the exponential distribution we always have for $x \in [0, \infty)$ that $f(x) \leq 1$, so f is a bounded function.

- f is m-times differentiable - using Laguerre Polynomials;

  $$L_m(x) = \frac{1}{m!} e^x \frac{d^m}{dx^m}(e^{-x}x^m)$$

  we can see that by taking $x := \frac{x}{2}$, and multiplying the equation through by $\lambda = \frac{1}{2}$ we can formulate an equation containing the m-th derivative of the exponential distribution;

  $$\begin{aligned}\frac{d^m}{dx^m}\left(\frac{e^{\frac{-x}{2}}}{2}\left(\frac{x}{2}\right)^m\right) &= L_m\left(\frac{x}{2}\right) m! \frac{e^{\frac{-x}{2}}}{2} \\ \frac{d^m}{dx^m}(f(x)x^m) &= 2^m L_m\left(\frac{x}{2}\right) m! f(x)\end{aligned}$$

  TODO .. use Leibniz rule? rearrange and write something like $\frac{d^m}{dx^m}(f(x)) = G(x)f(x)$..

- $\exists r_* > 0$ and a Borel measurable function g_*, with $\|y-x\| \leq r_*$ so that $\|f^{(t)}(x)\| \leq g_*(x)f(x)$ and $\|f^{(m)}(x) - f^{(m)}(x)\| \leq g_*(x)f(x)\|y-x\|^{\eta}$, for some g_* such that $\sup_{\{x:f(x)<\delta\}} g_*(x) = O(\delta^{-\epsilon})$ as $\delta \searrow 0$ for some $\epsilon > 0$

  TODO.. can be done once previous point has been written out correctly

Next, to satisfy Condition 2, for the density function f of the exponential distribution, must fulfill that;

- The α-moment of f must be finite, so $\int_{\mathbb{R}^d} \|x\|^{\alpha} f(x)dx < \infty$ - this is true since the moments of the exponential distribution are given by;

  $$\begin{aligned}\int_{\mathbb{R}^d} \|x\|^{\alpha} f(x)dx &= \int_0^{\infty} |x|^{\alpha} \lambda e^{-\lambda x} dx \\ &= \frac{\alpha!}{\lambda^{\alpha}} < \infty\end{aligned}$$

  for all $\alpha \in \mathbb{N}$, which is obviously finite.

Lastly, to satisfy Condition 3, we must find the values of k for which the estimator provides a uniform convergence for Theorems **??** and **??**. As previously, these values are independent of the distribution that the sample is from, and only depends on the size of the sample, the dimension of the distribution that sample is taken from and the value chosen for α, where we have chosen $\alpha = 2$. Thus, the values of k found in section 4.1.1 are $\{1, 2, ..., 11\}$.

Due to the above conditions for Theorems **??**, **??**, **??** and **??** being met, we can say that the Kozachenko-Leonenko estimator, of a sample from the exponential distribution is an asymptotically unbiased and consistent estimator for entropy.

4.3.2 k=1

In a similar fashion to how the previous distributions were explored in sections 4.1.2 to 4.1.9 and 4.2.2 and 4.2.6, I will be considering 500 samples of size N from the exponential distribution, find the estimator in each case and take the average of these estimators to find our entropy estimator, using the equation;

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

Considering the bias for this estimator against the actual value of entropy for the exponential distribution, with $\lambda = 0.5$, shown in equation (**??**), for different samples sizes N gives Table 4.19.

These results show an overall decreasing trend in the bias, and a strong decreasing variance of the bias. However, this is not a decreasing sequence of numbers, as in many places (i.e. from $N = 10,000$ to $N = 25,000$) the bias is increasing, which disagrees with Theorem **??**, that the estimator is asymptotically unbiased; since the smallest value of bias occurs at $N = 10,000$. There could be a number of reasons for the behaviour shown;

Table 4.19: *1-dimensional exponential distribution, $k = 1$*

N	$\hat{H}_{N,1}$	$\lvert Bias(\hat{H}_{N,1})\rvert$	$Var(Bias(\hat{H}_{N,1}))$
100	1.682163	0.01098459356	0.03000921515
200	1.692696	0.00045124177	0.01814679310
500	1.690508	0.00263961497	0.00634492739
1000	1.687560	0.00558717255	0.00319726793
5000	1.695081	0.00193336147	0.00058924735
10000	1.693198	0.00005051942	0.00032789400
25000	1.694373	0.00122575211	0.00012790693
50000	1.693321	0.00017399576	0.00005779114

- The estimator for the exponential distribution is not asymptotically unbiased - we know that this is not true since in section 4.3, I have shown how samples from this distribution satisfies the conditions to imply that the Theorems **??** and **??** hold. Thus the estimator for samples from the exponential distribution must be asymptotically unbiased and consistent, hence we should see the trend that $|Bias(\hat{H}_{N,1})| \to 0$ as $N \to \infty$.

- The relationship isn't smooth but does overall show asymptotic unbias - this could be true, since we are not looking at every data point from $N = 1 \to 50,000$, we have only considered 8 values of N. Between each N we would expect variability and not an exact smoothness; this is because we are considering simulations so are not expecting to have every point of bias against N in exactly the correct place. Thus considering each point could confirm Theorem **??**; this will be done in the graph of the $log|Bias(\hat{H}_{N,1})|$ against $log(N)$.

- The table could just contain outliers - the amount that the values of bias jumps around would indicate that the majority of them being outliers is not feasible. Additionally, the estimator for each value N is computed 500 times, and then the average is recorded in table 4.19, so this should smooth out any outliers that could occur from the computation.

To further examine the relationship between bias and N, which was inconclusive in table 4.19, I have depicted Figure 4.16, shows the 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 $a_1 = 0.4174$ and $c_1 = 0.0198$ for the relationship shown in (4.1);

$$|Bias(\hat{H}_{N,1})| \approx \frac{0.0198}{N^{0.4174}}$$

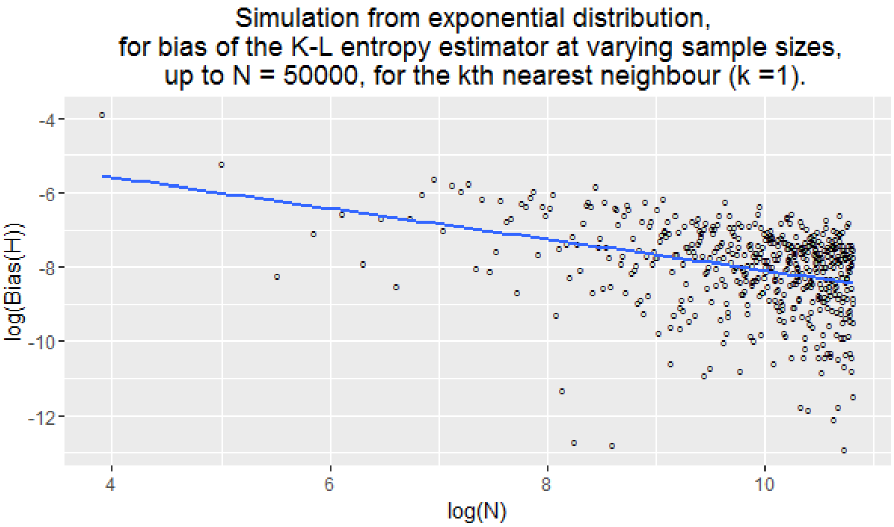

Figure 4.16: *Regression plot of* $\log|Bias(\hat{H}_{N,1})|$ *against* $\log(N)$

To say that this estimator is asymptotically unbiased, we would like to have the value of $a_k > 0.5$, which is not the case here. This is similar to the relationship shown in sections 4.1.2 and 4.2.2, where we also had $a_1 = 0.4594, 0.3698 < 0.5$. Although this is not what we would like in our estimator, it is consistent throughout the different distribution considered so far. Also, it still implies a negative relationship between the bias and N, as N increases the bias will decrease, just not a relationship as strong one would like. These three regression lines can be plotted together to further show their similarities, depicted in figure 4.17.

This graph shows that for each distribution, with $k = 1$, has a similar trend between the logarithm of the bias and the logarithm of N. It appears that the normal has the strongest negative correlation, but only by a small amount; considering the graph for larger values of $log(N)$ show the regression lines being particularly close to one and other for large N. A more detailed comparison will be further explored in section 4.3.7.

4.3.3 k=2

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

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

Using the same parameters as before; taking 500 samples of size $N = 100 \to 50,000$ from the exponential distribution, with rate parameter $\lambda = 0.5$, and con-

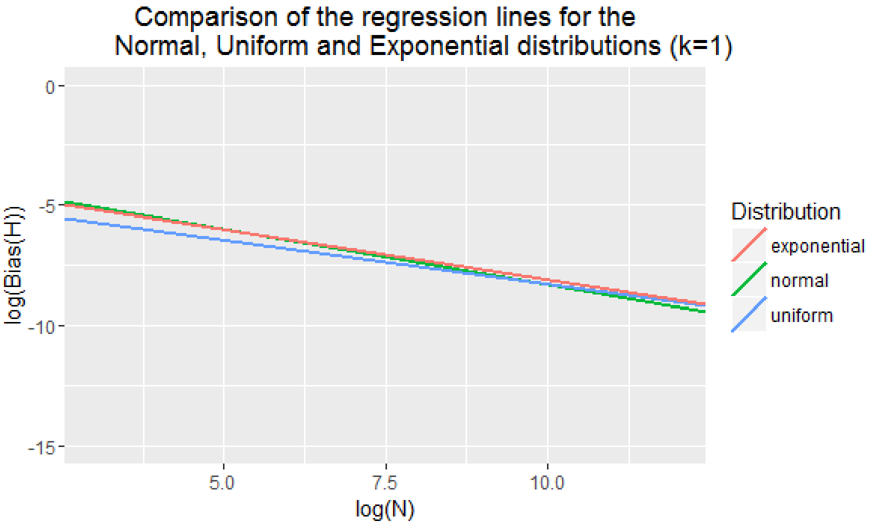

Figure 4.17: *Regression plot of* $\log|Bias(\hat{H}_{N,1})|$ *against* $\log(N)$*, for the 1-dimensional distributions; exponential, uniform and normal*

Table 4.20: *1-dimensional exponential distribution,* $k = 2$

N	$\hat{H}_{N,2}$	$\lvert Bias(\hat{H}_{N,2})\rvert$	$Var(Bias(\hat{H}_{N,2}))$
100	1.692627	0.00051996867	0.01705110058
200	1.686513	0.00663429438	0.00942333921
500	1.691328	0.00181927748	0.00371778428
1000	1.692399	0.00074835082	0.00203062072
5000	1.692660	0.00048693059	0.00039750041
10000	1.693193	0.00004543429	0.00019553001
25000	1.693504	0.00035725960	0.00007458969
50000	1.693084	0.00006338711	0.00004293387

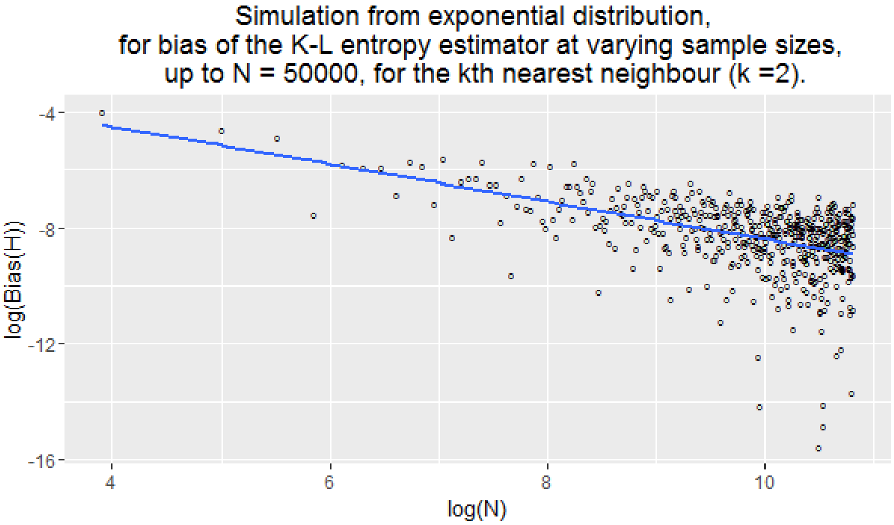

Figure 4.18: *Regression plot of* $\log|Bias(\hat{H}_{N,2})|$ *against* $\log(N)$

sider the estimator of these samples against the actual value of entropy, equation 4.8, we get the results in table 4.20. These results are inconclusive in showing the relationship desired between the $|Bias(\hat{H}_{N,2})|$ and N, from Theorem **??**, this is similar to what was found for $k = 1$ in section 4.3.2, and a more convincing conclusion was made by examining the relationship in equation (4.2). Thus, I have also plotted the regression graph again, this time for $k = 2$, the results are shown in Figure 4.16 and the regression coefficients are given by $a_2 = 0.6480$ and $c_2 = 0.1482$.

Using equation (4.1), we have the relationship between the bias and N given by;

$$|Bias(\hat{H}_{N,1})| \approx \frac{0.1482}{N^{0.6480}}$$

where here we have the $a_2 > a_1$, which is the same relationship that we have found in the previous two distributions. We can also compare th value of a_2 in these other distributions and we have; $a_2 = 0.5857$ for the normal distribution and $a_3 = 0.5998$ for the uniform distribution (from sections 4.1.3 and 4.2.3 respectively). All of these values vary by only ≈ 0.06, showing that the information found from the analysis done is consistent throughout the distributions considered so far. I have also plotted these regression lines alongside each other, in Figure 4.19, to depict the differences/similarities, between the regression line for each distribution.

This graph confirms the closeness of the regression lines for each distribution, for k=2, we can see that the uniform distribution has the lowest regression line; indicating that at the largest N considered, we would have the smallest bias for

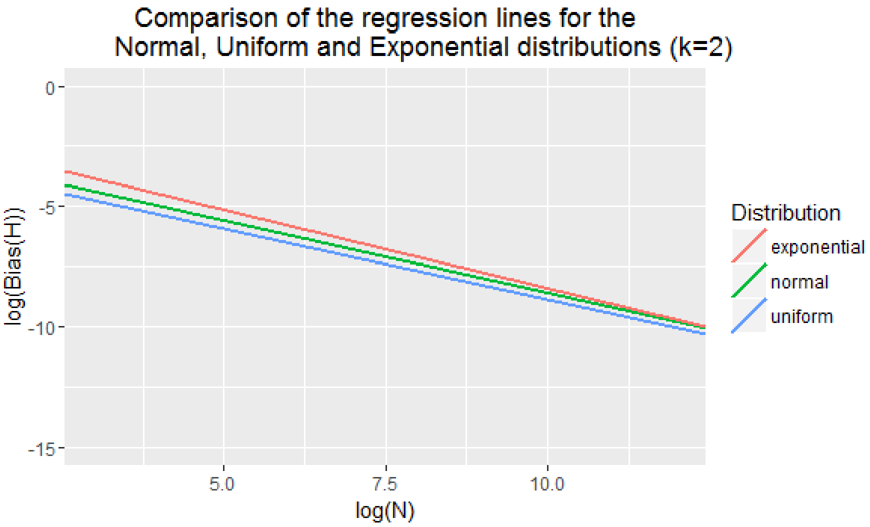

Figure 4.19: *Regression plot of* $\log|Bias(\hat{H}_{N,2})|$ *against* $\log(N)$, *for the 1-dimensional distributions; exponential, uniform and normal*

this distribution. However, the exponential distribution has the steepest slope to its line; demonstrating that for a large N (larger than what has been considered so far), we would have that the bias of samples from this distribution tends to 0 the fastest. Moreover, the regression lines are very similar between these distributions for k=2, suggesting that the results here are as to be expected.

4.3.4 k=3

Again, for $k = 3$, I will examine 500 samples of size N from the exponential 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 4.21. This table is beginning to indicate a strong trend between the bias of the estimator and the sample size, N. The previous tables from section 4.3.2 and 4.3.3, had inconclusive values in their tables; however this table does seem to suggest that the estimator agrees with Theorem **??**, that the estimator is asymptotically unbiased. To further examine the relationship between the values of bias and N, I have created the graph depicted in Figure 4.20.

From this analysis, I have found that for this distribution we have the coefficients of regression to be $a_3 = 0.5711$ and $c_3 = 0.0585$, which implies a

Table 4.21: *1-dimensional exponential distribution,* $k = 3$

N	$\hat{H}_{N,3}$	$\lvert Bias(\hat{H}_{N,3})\rvert$	$Var(Bias(\hat{H}_{N,3}))$
100	1.687045	0.0061025449	0.01703802128
200	1.689520	0.0036271927	0.00726556895
500	1.690834	0.0023135057	0.00293535162
1000	1.696420	0.0032728456	0.00150595517
5000	1.693846	0.0006989676	0.00030230568
10000	1.692206	0.0009408184	0.00018257444
25000	1.692290	0.0008569323	0.00006568172
50000	1.693392	0.0002444838	0.00003130878

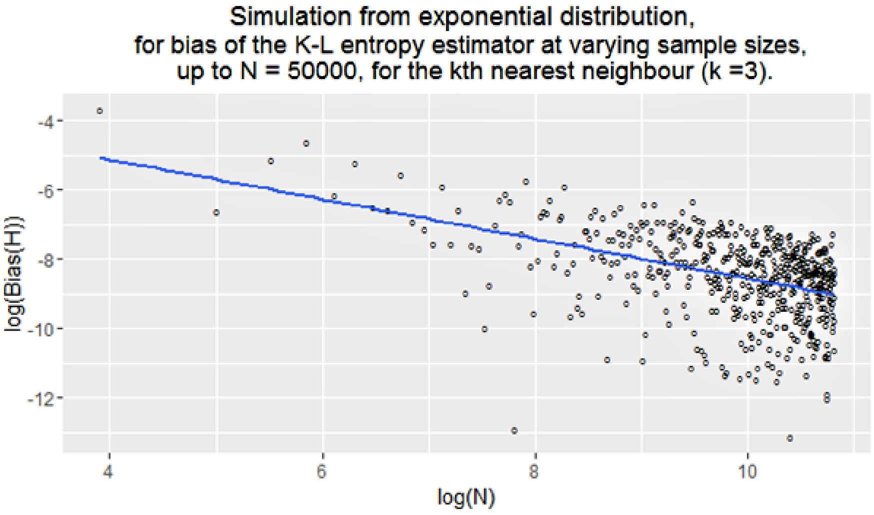

Figure 4.20: *Regression plot of* $\log|Bias(\hat{H}_{N,3})|$ *against* $\log(N)$

relationship of the form;

$$|Bias(\hat{H}_{N,1})| \approx \frac{0.0585}{N^{0.5711}}$$

In comparison to the regression coefficients found for smaller values of k, for samples also from the exponential distribution, I have found that this relationship is unlike expected. This is because, for this distribution, we have; $a_1 < a_3 < a_2$, whereas for the 1-dimensional normal and uniform distribution I have consistently found that $a_1 < a_2 < a_3$. There is also similar behaviour for the values of c_k, where here we have $c_1 < c_3 < c_2$ in contrast to $c_1 < c_2 < c_3$. This could mean a number of things;

- When $k = 2$, we have a stronger asymptotic unbias than when $k = 1, 3$ - this could be true, since we do not know what value of k will produce the best estimator, also in the previous distributions, we only had a difference of ≈ 0.045 for normal and ≈ 0.043 for uniform, between the values of a_2 and a_3. However, since the previous analysis done in sections 4.2 and 4.1 implies otherwise, I would not encourage this solution. To confirm this either way, it would be beneficial to take into account the regression relationship for higher values of k for samples fro the exponential distribution.

- There are outliers in the solution - this is a possibility, since the regression relationship for $k = 2$ and $k = 3$, have similar values of the gradient $a_2 = 0.6480$ and $a_3 = 0.5711$, they differ by ≈ 0.077. Thus, if the value of a_2 is slightly higher than it should be and if a_3 is slightly lower than it should be, both due to outliers; then we could actually have the relationship already assumed, that $a_2 < a_3$. On the other hand, one would not expect this to happen, since any outliers should have been smoothed out in the process of considering a large number of samples (300) for a lot of different sample sizes (100, 200, 300, ... , 50000).

We can also compare the found values for a_3 and c_3 between distributions; normal - $a_3 = 0.6443$ and $c_3 = 0.1156$, uniform - $a_3 = 0.6291$ and $c_3 = 0.0737$ and exponential - $a_3 = 0.5711$ and $c_3 = 0.0585$. Thus the values of the slope and intercept differ by only ≈ 0.073 and ≈ 0.057 respectively, which are not especially large values of difference. I have plotted the regression lines for each distribution together in figure 4.21 to hopefully more easily view the differences.

From this plot, we can see that the regression lines for each distribution are actually very similar to one and other, where at the largest value of N considered here we have that the regression lines implies that the uniform distribution has the smallest bias at this point. However, the normal distribution has the steepest slope, thus for a larger N than what is considered in the graph, one would have that the estimator for samples from the normal distribution has the strongest asymptotic unbias. A more detailed comparison of all values of k and all distributions will be explored in section 4.3.7.

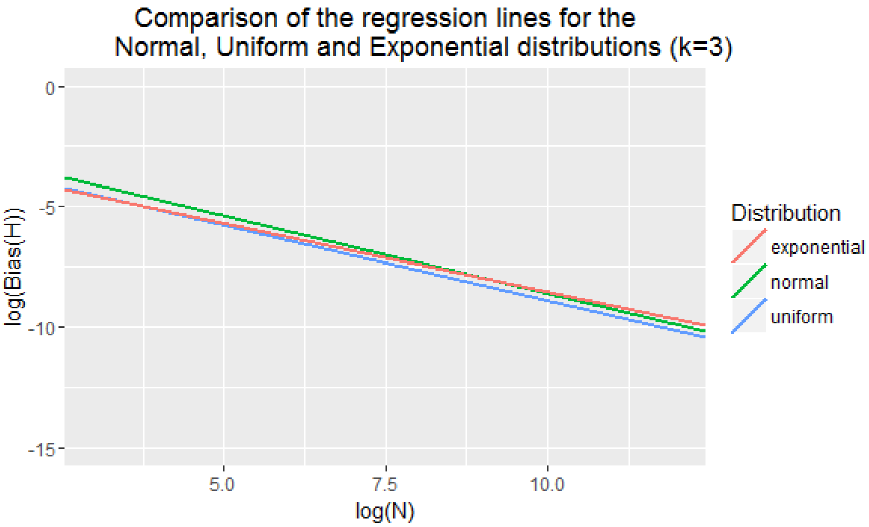

Figure 4.21: *Regression plot of* $\log|Bias(\hat{H}_{N,3})|$ *against* $\log(N)$*, for the 1-dimensional distributions; exponential, uniform and normal*

4.3.5 k=5

Now we consider the estimator, shown in equation 3.14, 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 estimator to the exact value of entropy, shown in equation (4.8), for 500 samples of size N, we get the results in table 4.22.

From this table we can see an overall decrease in bias, from $\approx 0.02 \rightarrow 0.0002$ as N goes from $100 \rightarrow 50,000$. This is fitting with Theorem **??**, that the estimator is asymptotically unbiased for this distribution. However, on closer inspection we can see, as in the previous values of k for this distribution, that the size of the bias does not decrease consistently as the sample size, N, increases in the table. For example from $N = 10,000$ to $N = 25,000$, the size of the bias actually increases by ≈ 0.00027, which one would not expect. However, we should allow for some variability in the results since they are found from simulations. Furthermore, I have considered a plot to represent equation (4.2), shown in Figure 4.24, where I have also found the coefficients of regression to be; $a_5 = 0.4793$ and $c_5 = 0.0241$.

This implies the relationship;

$$|Bias(\hat{H}_{N,5})| \approx \frac{0.0241}{N^{0.4793}}$$

Table 4.22: *1-dimensional exponential distribution,* $k = 5$

N	$\hat{H}_{N,5}$	$\lvert Bias(\hat{H}_{N,5})\rvert$	$Var(Bias(\hat{H}_{N,5}))$
100	1.671551	0.02159648310	0.01379216507
200	1.679168	0.01397964263	0.00670466906
500	1.692568	0.00057877793	0.00231523099
1000	1.694130	0.00098238978	0.00134446983
5000	1.692592	0.00055538663	0.00029049727
10000	1.693190	0.00004331284	0.00014609571
25000	1.692837	0.00031015885	0.00005897936
50000	1.692901	0.00024589568	0.00002504615

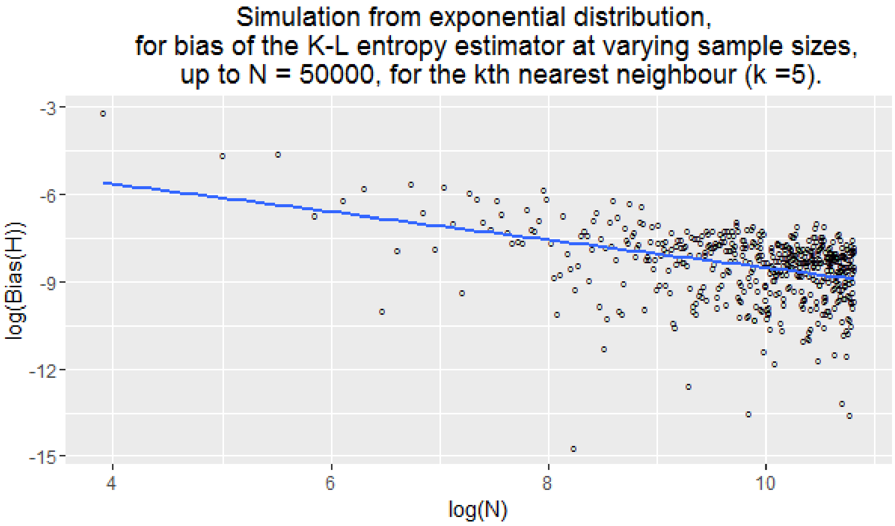

Figure 4.22: *Regression plot of* $\log|Bias(\hat{H}_{N,5})|$ *against* $\log(N)$

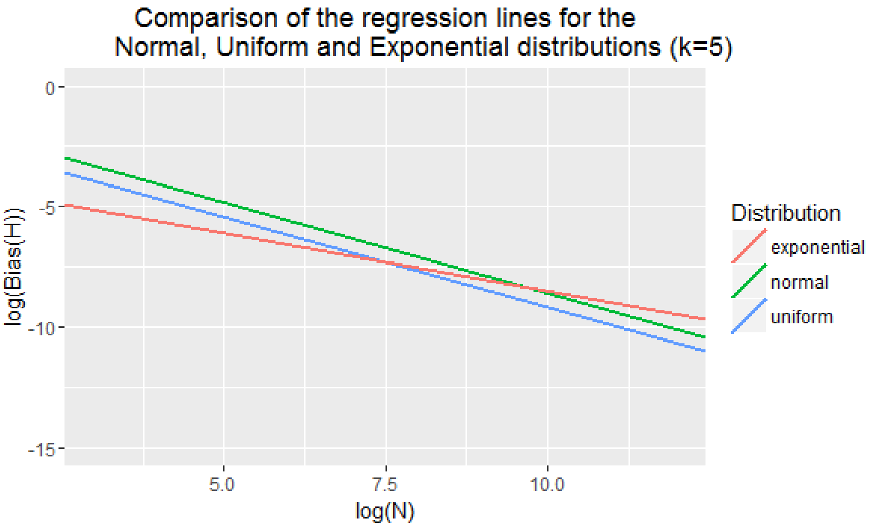

Figure 4.23: *Regression plot of* $\log|Bias(\hat{H}_{N,5})|$ *against* $\log(N)$, *for the 1-dimensional distributions; exponential, uniform and normal*

In comparison to the coefficients from the previous value of k, we find that these coefficients also do not fit with the trend that as $k \to 10$ that the values of a_k and c_k also increase. For the exponential distribution we have found, so far, that $a_1 < a_5 < a_3 < a_2$; this is atypical to the normal and uniform distributions where the results are consistently $a_1 < a_2 < a_3 < a_5 < a_{10}$. To compare the values of a_5 for these distributions, I have plotted their regression lines together, depicted in Figure 4.23.

This graph shows that, the normal and uniform have almost parallel to one and other, with uniform lowest at the largest N considered. On the other hand, we have the regression line for the bias of samples from the exponential distribution, which is a more gradual decline that those for the other two distributions. Indicating that the asymptotic unbais of this distribution is not as strong as that for the other two. The reasoning behind this is currently unclear, but a detailed analysis into this will take place in section 4.3.7.

4.3.6 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;

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

Table 4.23: *1-dimensional exponential distribution,* $k = 10$

N	$\hat{H}_{N,10}$	$\|Bias(\hat{H}_{N,10})\|$	$Var(Bias(\hat{H}_{N,10}))$
100	1.681136	0.01201069876	0.01103729658
200	1.679907	0.01323996238	0.00586663681
500	1.687144	0.00600360695	0.00209594189
1000	1.692988	0.00015893618	0.00118121417
5000	1.693597	0.00044935080	0.00025313033
10000	1.692565	0.00058186611	0.00012048276
25000	1.693052	0.00009528961	0.00004392069
50000	1.693243	0.00009553753	0.00002392945

The results for the comparison between this estimator and the exact value of entropy, equation (4.8), are displayed in table 4.23.

This values in this table, now begin to show more what we are used to; that as N increases the bias of the estimator decreases, which is fitting with Theorem **??**. Additionally, the variance of the bias is decreasing for a larger N, again something to be expected by Theorem **??**. To view this relationship in a more desirable way, we can consider the equation (4.2), as previously done, and generate 500 samples, for each size $N = 100, 200, ..., 50000$, and find out the estimator $\hat{H}_{N,k}$ each time, to then work out the average and plot the logarithm of the modulus of the Bias for each N, $log|Bias(\hat{H}_{N,k})|$, against the logarithm of N, $log(N)$. This is shown in figure 4.24.

From plotting these points and fitting a regression line, we get the relationship, shown in equation (4.1), here with the coefficients $a_{10} = 0.7170$ and $c_{10} = 0.2292$, thus;

$$|Bias(\hat{H}_{N,5})| \approx \frac{0.2292}{N^{0.7170}}$$

We wish to have $a_k > 0.5$, which is true for this particular realisation. In comparison to the other coefficients for different k in the exponential distribution,we have that; $a_1 < a_5 < a_3 < a_2 < a_{10}$. This does show that now there is a general trend, since $a_1 < a_5 < a_{10}$, but this is still not the same as what was shown for the other distributions. Considering $k = 5$ and the values of a_5 and c_5, for the different distributions, I have plot the regression lines together in figure 4.17.

This graph shows ... TODO

4.3.7 Comparison of k

In sections 4.3.2 to 4.3.6, I have explored the Koazchenko-Leonenko estimator for samples from the 1-dimensional exponential distribution.

smth about tables and bias with N

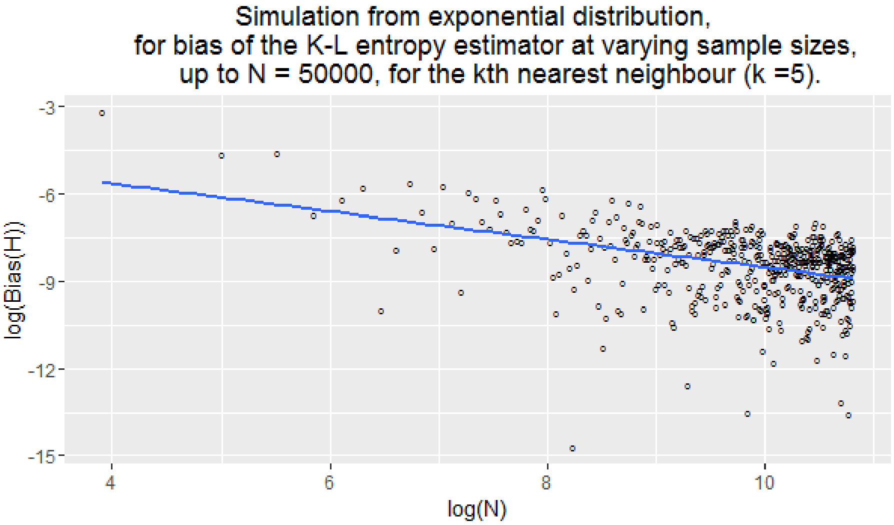

Figure 4.24: *Regression plot of* $\log|Bias(\hat{H}_{N,5})|$ *against* $\log(N)$

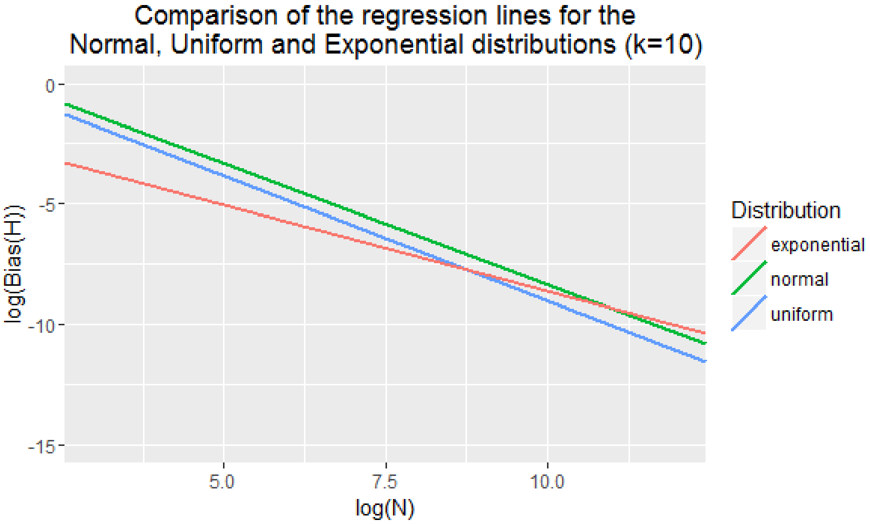

Figure 4.25: *Regression plot of* $\log|Bias(\hat{H}_{N,10})|$ *against* $\log(N)$*, for the 1-dimensional distributions; exponential, uniform and normal*

Table 4.24: *1-dimensional exponential distribution, a comparison of k*

k	a_k	c_k
1	0.4147	0.0198
2	0.6480	0.1482
3	0.5711	0.0585
5	0.4793	0.0241
10	0.7170	0.2292

Additionally, it is important to note that in all the tables for the exponential distribution, we consistently have $Var(Bias(\hat{H}_{N,k}))$ decreases as $N \to \infty$. By Theorem **??**, we know that $Var(\hat{H}_{N,k}) \approx \frac{Var(\log f(x))}{N}$, and for $N = 50,000$ we have for this distribution that the variance of $Bias(\hat{H}_{N,k})$, for $k = 1, 2, 3, 5, 10$ are given by $\approx 0.000058, 0.000043, 0.000031, 0.000025, 0.000024$ respectively. By the non-linearity of the variance we have that $Var(Bias(\hat{H}_{N,k})) = Var(\hat{H}_{N,k} - H) = Var(\hat{H}_{N,k})$. For this distribution we have;

$$Var(\log f(x)) = TODO... \tag{4.9}$$

I will be now focusing on the results of the coefficients of equation (4.1), found from the simulations above;

$$|Bias(\hat{H}_{N,k})| = \frac{c_k}{N^{a_k}}$$

The results from the investigation above have been condensed into table 4.24, showing the change in values of a_k and c_k for different k.

This table doesn't show the same as the other distributions; previously, in sections 4.2.7 and 4.1.10, we have found that $a_1 < a_2 < a_3 < a_5 < a_{10}$, similarly with the intercept coefficient we have found that $c_1 < c_2 < c_3 < c_5 < c_{10}$. However, here the coefficients at $k = 3$ and $k = 5$ are not as expected, they are smaller than those for $k = 2$. To examine this more, I will find the coefficients of regression for estimators found when $k = 4, 6, 7, 8, 9, 11$, these are displayed in table 4.25.

This tables appears to show that the values of a_k and c_k for $k = 2, 5, 8$ seem to be different to the trend shown with the other values of k for this distribution and also for the other distributions; normal and exponential. To try and smooth out any of these errors, I will now consider $3,000$ (instead of the previous 500) samples of size N, to then find the estimator and compute the relationship in equation (4.2). This will be done too find the coefficients shown in table 4.25 when taken as an average over a larger number of samples, for $k = 2, 5, 8$, the results are as follows; TODO...

I can also compare the values found for a_k and c_k for the exponential distribution, to those found from the other 1-dimensional distributions previously considered, I have condensed this information into table 4.26.

Table 4.25: *1-dimensional exponential distribution, a comparison of more k*

k	a_k	c_k
1	0.4147	0.0198
2	0.6480	0.1482
3	0.5711	0.0585
4	0.5914	0.0731
5	0.4793	0.0241
6	0.6515	0.1333
7	0.6023	0.0772
8	0.5384	0.0392
9	0.7139	0.2202
10	0.7170	0.2292

Table 4.26: *Comparison between 1-dimensional Exponential, Uniform and Normal distribution*

—	Normal		Uniform		Exponential	
k	a_k	c_k	a_k	c_k	a_k	c_k
1	0.4594	0.0249	0.3698	0.0103	0.4147	0.0198
2	0.5998	0.0746	0.5857	0.0503	0.6480	0.1482
3	0.6443	0.1156	0.6291	0.0737	0.5711	0.0585
5	0.7568	0.3557	0.7501	0.1889	0.4793	0.0241
10	1.0055	5.5942	1.0357	3.8217	0.7170	0.2292

Chapter 5

Conclusion

Bibliography

[1] I. Ahmad and P. Lin. A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Transactions on Information Theory*, 22:372–375, 1976.

[2] J. Beirlant, E. Dudewicz, L. Gyorfi, and E. van der Muelen. Nonparametric entropy estimation : an overview. *International Journal of Mathematical and Statistical Sciences*, 6:17–39, 2001.

[3] T. Berrett, R. Samworth, and M. Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *arXiv preprint arXiv:1606.00304*, 2, 2016.

[4] P. Crzcgorzewski and R. Wirczorkowski. Entropy-based goodness-of-fit test for exponentiality. *Communications in Statistics-Theory and Methods*, 28:1183–1202, 1999.

[5] S. Delattre and N. Fournier. On the kozachenko-leonenko entropy estimator. *arXiv preprint arXiv:1602.07440*, 1, 2016.

[6] Y. Dmitriev and F. Tarasenko. On the estimation functions of the probability density and its derivatives. *Theory Probability Applications*, 18:628–633, 1973.

[7] Y. Du, J. Wang, S-M. Guo, PD. Thouin, et al. Survey and comparative analysis of entropy and relative entropy thresholding techniques. *IEEE Proceedings-Vision, Image and Signal Processing*, 153:837–850, 2006.

[8] E. Dudewicz and E. Van Der Meulen. Entropy-based tests of uniformity. *Journal of the American Statistical Association*, 76:967–974, 1981.

[9] W. Gao, S. Oh, and P. Viswanath. Demystifying fixed k-nearest neighbour information estimators. *arXiv preprint arXiv:1604.03006*, 2, 2016.

[10] D. Gokhale. On entropy-based goodness-of-fit tests. *Computational Statistics & Data Analysis*, 1:157–165, 1983.

[11] P. Hall. Limit theorems for sums of general functions of m-spacings. *Mathematical Proceedings of the Cambridge Philosophical Society*, 96:517–532, 1984.

[12] P. Hall and S. Morton. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45:69–88, 1993.

[13] A. Hero and O. Michel. Asymptotic theory of greedy approximations to minimal k-point random graphs. *IEEE Transactions on Information Theory*, 45:1921–1938, 1999.

[14] K. Hlaváčková, M. Paluš, M. Vejmelka, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441:1–46, 2007.

[15] H. Joe. On the estimation of entropy and other functionals pf a multivariate density. *Annals of the Institute of Statistical Mathematics*, 41:171–178, 1989.

[16] J. Kapur and H. Kesavan. Entropy optimization principles with applications. 1992.

[17] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical review E*, 69, 2004.

[18] H. Kuo and Y. Gao. Maximum entropy direct models for speech recognition. *IEEE Transactions on audio, speech, and language processing*, 14:873–881, 2006.

[19] E. Learned-Miller and J. Fisher. Ica using spacings estimates of entropy. *Journal of machine learning research*, 4:1271–1295, 2003.

[20] N. Leonenko and L. Kozachenko. On statistical estimation of entropy of random vector. *Problems of information transmission*, 23:9–16, 1987.

[21] N. Leonenko, L. Pronzato, and V. Savani. A class of rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36:2153–2182, 2008.

[22] N. Leonenko and O. Seleznjev. Statistical inference for the ϵ - entropy and the quadratic rényi entropy. *Journal of Multivariate Analysis*, pages 1981–1994, 2010.

[23] H. Neemuchwala, A. Hero, and P. Carson. Image matching using alpha-entropy measures and entropic graphs. *Signal processing*, 85:277–296, 2005.

[24] J. Shen, J. Hung, and L. Lee. Robust entropy-based endpoint detection for speech recognition in noisy environments. *ICSLP*, 98:232–235, 1998.

[25] K-S. Song. Limit theorems for nonparametric sample entropy estimators. *Statistics & probability letters*, 49:9–18, 1998.

[26] F. Tarasenko. On the evaluation of an unknown probability density function, the direct estimation of the entropy from independent observations of a continuous random variable, and the distribution-free entropy test of goodness-of-fit. *IEEE Transactions on Information Theory*, 56:2052–2053, 1968.

[27] W. van Wieringen and A. van der Vaart. Statistical analysis of the cancer cell's molecular entropy using high-throughput data. *Bioinformatics*, 27:556–563, 2011.

[28] O. Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 54–59, 1976.

[29] Wikipedia. *Entropy Estimation.* https://en.wikipedia.org/wiki/Entropy estimation.

[30] Wikipedia. *Speech Recognition.* https://en.wikipedia.org/wiki/Speech recognition.