MMath Project

C1231497

Statistical Inference for Entropy

Author:
Karina Marks

Supervisor:
Prof. Nikolai Leonenko

April 6, 2017

Contents

Chapter 1

Introduction

1.1 Entropy

Entropy $H(S)$, 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.

"$H(S)$ is the quantity of surprise you should feel upon reading the result of a measurement" [10]; thus, the "entropy of S can be seen as the uncertainty of S" [15].

1.1.1 Shannon entropy

The Shannon entropy of a random vector $X \in \mathbb{R}^d$, from a continuous distribution, with density function f is given by:

$$H = -\int_{x \in \mathbb{R}^d} f(x) log(f(x)) dx \tag{1.1}$$

(This is often referred to as the differential entropy [25])

If a random vector X has a discrete distribution with probability mass function $f(x)$, $x \in \mathbb{Z}^d$, then we have Shannon entropy defined as:

$$H = -\sum_{x \in \mathbb{Z}^d} f(x) log(f(x))$$

1.1.2 Rényi and Tsallis entropies

The Rényi and Tsallis 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 $X \in \mathbb{R}^d$, from a continuous distribution with density function f, we define:

Rényi entropy

$$H_q^* = \frac{1}{1-q} log \left(\int_{\mathbb{R}^d} f^q(x) dx \right) \qquad (q \neq 1) \tag{1.2}$$

Tsallis entropy

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

Moreover, if we consider a random vector X from a discrete distribution, with probability mass function $f(x)$ for $x \in \mathbb{Z}^d$, then for $q \neq 1$, the Rényi and Tsallis entropies are defined as follows:

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

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

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) as this is a special case for when $q = 1$.

1.2 Background

1.2.1 Properties of entropy

I will begin by exploring properties specific to the Shannon entropy; then progress to those for other types of entropy. Kapur and Kesavan's book on *Entropy Optimization Principles with Applications* [18], gives an account of some properties of the Shannon entropy H of a discrete distribution. Whilst Johnson [17] discusses properties for the

Shannon entropy of continuous distributions - the differential entropy, and compares them to the discrete properties in his book on *Information Theory and The Central Limit Theorem*.

The Shannon entropy of a discrete random variable is always non-negative, $H \geq 0$, and is strictly positive unless if f is any of the N degenerate distributions (X is deterministic). That is, the only case for which $H = 0$, for a discrete distribution with probability density function f, is when $f(x_i) = 1$ if $i = k$, $k \in [1, N]$ and $f(x_i) = 0$ otherwise, for a discrete distribution. However, for the differential entropy, H can be either positive or negative - can take any value between $-\infty$ and ∞ [25]. This is obvious to see if we consider a random variable Y from the continuous uniform distribution on $[0, c]$, then:

$$H = -\int_0^c \frac{1}{c} \log\left(\frac{1}{c}\right) dx = \log c$$

Thus, depending on the value of c we can obviously have either a positive or negative entropy.

Moreover, for a discrete random variable X, the Shannon entropy is both shift and scale invariant, which means for all $a \neq 0$ and $b \in \mathbb{R}$ we have:

$$H(aX + b) = H(X) \tag{1.4}$$

On the contrary, for a continuous random variable Y, only the shift invariance holds. Thus for all $a \neq 0$ and $b \in \mathbb{R}$, we have the following relationship:

$$H(aY + b) = H(Y) + \log a \tag{1.5}$$

The maximum value of the continuous entropy H is attained with different distributions, dependent on a number of factors. The maximising distribution f depends on how it is supported, and what conditions hold on the moments of X. For example, when $supp\{f\} = \mathbb{R}$, for a fixed variance σ^2, the maximum of H is attained when f is from the Normal/Gaussian distribution, in comparison to any other distribution with the same σ^2 [18].

Additionally, for the Shannon entropy, there is an interesting result on an upper bound of H, one which depends on the distribution of X, which can be used to show

if entropy is finite. Consider a random variable X, then for $s > 0$ we have:

$$H(X) \leq \frac{1}{s} \log \left(\frac{2^s e \Gamma^s(\frac{1}{s}) \mathbb{E}(|X|^s)}{s^{s-1}} \right) \tag{1.6}$$

This indicates that if the s-th moment of X is finite, $\mathbb{E}(|X|^s) < \infty$, for some $s > 0$ then the Shannon entropy of X will also be finite, $H(X) < \infty$ [27].

Shannon entropy, as mentioned earlier, is a special case of the Rényi and Tsallis entropies as $q \to 1$. Thus, for a general q-entropy, different properties hold. Namely, H_q is concave when $q > 0$ (and convex when $q < 0$), implying for the Shannon entropy, when $q = 1$, that H is concave [23].

There are also other special cases that have certain properties; for example the Rényi entropy with $q = 2$ is known as the Quadratic Rényi entropy [24]. This is defined as follows for a continuous random variable X with density function f:

$$H_2^* = -log \left(\int_{\mathbb{R}^d} f^2(x) dx \right) \tag{1.7}$$

Moreover, another special case occurs when considering the Rényi entropy as $q \to \infty$, if the limit exists, which is defined as the minimum entropy, since it is the smallest possible value of H_q^*:

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

Furthermore, there are some intriguing relationships between the different specific types of entropy, for example Leonenko and Seleznjev [24] show the following relationship between H_2^* and H_∞^*:

$$H_\infty^* \leq H_2^* \leq 2H_\infty^* \tag{1.8}$$

Additionally, they show an approximate relationship between the Shannon entropy, H, and the Quadratic Rényi entropy, H_2^* :

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

where d is the dimension of the distribution that the entropy is taken from.

1.2.2 Applications of entropy

Entropy began as a concept in thermodynamics - about the idea 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 I will consider applications accordingly. The following is a short overview of some of its applications; however, this is not an exhaustive list, since the applications of entropy are extensive.

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.

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; in most cases this must be estimated, as an exact entropy is not always known. Kraskov, Stögbauer and Grassberger [19] discussed how estimating the mutual information (MI) using entropy estimators is useful for assessing the independence of components from ICA. Learned-Miller and Fisher [21] 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 [14] first discuss the applications of Rényi entropy in image processing, then Neemuchwala, Hero and Carson [26] discuss how, in image analysis, an important task is that of image retrieval. This uses entropy estimation to compute entropic similarities that are used to match a reference image to another image. Furthermore, Du, Wang, Guo and Thouin, [8] considered the importance of entropy-based image thresholding, using both Shannon and relative entropy. Where relative entropy is a modification of the Shannon entropy, introduced to recover the

scale invariance property for the entropy of a continuous distribution [17].

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, [32] discuss how chromosomal disorganisation increases as cancer progresses; they mention how the estimation of entropy (specifically the Kozachenko-Leonenko estimator) can be used to help find this disorganisation; 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, enabling the recognition and translation of spoken language into text by computers. Shen, Hung and Lee [29] discuss how an entropy based algorithm can conduct accurate SR in noisy environments. Moreover, Kuo and Gao [20] focus on a method where the probability of a state or word sequence, given an observation sequence, is computed directly from the maximum entropy model.

It is also 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 [33]. Dudewicz and van der Meulen discussed the properties for when the uniform distribution maximises the entropy [9]. Moreover, others have explored different distributions and their entropic properties; see [11, 5].

1.2.3 Other estimators of entropy

There are several 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) [3], gives an account 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) [15] gives a more detailed observa-

tion of these different types of estimators. Below is a summary of the types of estimators, which will lead us to understand why we choose the Kozachenko-Leonenko estimator for entropy.

First, the types of consistency must be established so it can be obviously seen how it compares to the Kozachenko-Leonenko estimator, for $X_1, ..., X_N$ a i.i.d sample from the distribution with density function f, where H_N is the estimator of $H(f)$. Then we have the following definitions as $N \to \infty$:

- Asymptotic Unbias

$$\mathbb{E}(H_N) - H(f) \to 0 \tag{1.9}$$

- Weak Consistency

$$H_N \xrightarrow{p} H(f) \tag{1.10}$$

- Mean Square Consistency

$$\mathbb{E}\{(H_N - H(f))^2\} \to 0 \tag{1.11}$$

- Strong Consistency (Asymptotic Normality)

$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{1.12}$$

$$\mathbb{E}(\hat{H}_{N,k} - H)^2 = \frac{\sigma^2}{N}\left(1 + O\left(\frac{1}{N}\right)\right) \tag{1.13}$$

(This is the consistency shown with the Kozachenko-Leonenko estimator in Theorem 2)

The types of nonparametric estimators can be split into three 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 2.2.

The plug-in estimates [3, 15] are based upon a consistent density estimate f_N, of density f, which depends on the sample $X_1, ..., X_N$.Two of these will be considered;

the most obvious estimator of this type is the integral estimate of entropy, given by:

$$H_N = -\int_{A_N} f_N(x) log(f_N(x)) dx \tag{1.14}$$

where the set A_N excludes the tail values of f_N. When the sample is from a 1-dimensional distribution, Dmitriev and Tarasenko, [7] for $A_N = [-b_N, b_N]$ and f_N the kernel density estimator, proved a strong consistency. However, if f_N is not estimated in this form, due to the numeric integration, for dimensions $d \geq 2$, Joe [16] 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:

$$H_N = -\frac{1}{N} \sum_{i=1}^{N} log(f_N(X_i)) dx \tag{1.15}$$

This was first proposed in 1976, by Ahmad and Lin [2] who showed the mean-square consistency of this estimator, where f_N is a kernel density estimate. Joe [16] then proceeded to obtain the asymptotic bias and variance, and whilst satisfying certain conditions reduced the mean square error. Moreover, Hall and Morton [13] progressed to say that under more restrictive conditions, we have strong consistency for 1-dimensional distributions. However, when $d = 2$ the root-n consistent estimator has a significant bias.

There are also two more plug-in estimates explored by Beirlant et al [3]: 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 $1 \leq d \leq 3$.

Hence, so far the estimates for entropy observed are only consistent whilst under strong conditions on f and f_N, and mostly for a 1-dimensional distribution. It is therefore important to consider the next category of estimates: estimates of entropy based on sample-spacings, namely the m-spacing estimate. Sometimes, it is not practical to estimate f_N, so this estimate is found based on spacings between the sample observations [31, 12].

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 $X_{N,i+m} - X_{N,i}$ is the m-spacing.

Firstly we look at an estimator of the following form, with fixed m:

$$H_{m,N} = \frac{1}{N} \sum_{i=1}^{N-m} log\left(\frac{N}{m}(X_{N,i+m} - X_{n,i})\right) - \Psi(m) + log(m) \tag{1.16}$$

where $\Psi(x)$ is the digamma function (a more detailed explanation of this function is in section 2.2). For a sample from a uniform distribution, this estimator has been shown to be consistent - proved by Tarasenko [31]. Under some conditions on f, the weak consistency and asymptotic normality was shown by Hall [12].

To decrease the asymptotic variance of the estimator, we consider the estimator when $m = m_N \to \infty$, so it is no longer fixed. This is defined slightly differently:

$$H_{m,N} = \frac{1}{N} \sum_{i=1}^{N-m_N} log\left(\frac{N}{m_N}(X_{N,i+m_N} - X_{n,i})\right) \tag{1.17}$$

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 [3].

The last category of estimators discussed by Beirlant, Dudewicz, Gyorfi and Muelen are those based on nearest neighbour distances [3]. 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 refrain from going into detail on this estimator presently; however, I will mention that strong consistency holds for dimension $d \leq 3$, but higher dimensions can cause problems. Henceforth, it is important to note that recently, a new estimator has been proposed by Berrerrt, Samworth and Yuan [4], 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 some assumptions, strong consistency holds.

Chapter 2

Kozachenko-Leonenko Estimator

2.1 History

The Kozachenko-Leonenko (K-L) estimator was first introduced by L.Kozachenko and N.Leonenko, in 1987, in the article *Sample Estimate of the Entropy of a Random Vector*, in the journal *Problems of Information Transmission* [22]. 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. They then established conditions under which asymptotic unbiasedness and consistency hold.

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*, this time using the k-nearest neighbour method, to consider estimators for the Rényi and Tsallis entropies [23]. 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 [4]; 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 K-L estimators for different values of k". This estimator will not be explored in depth; however, the understanding of the value of k depending on N will be examined in detail. Additionally, for dimension $d \leq 3$, 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* [6], where they studied in detail the bias and variance of this estimator considering all 3 proposed values of k: $k = 1$, k fixed, and k depends on N. They also provided a development for the bias of this estimator when $k = 1$, in dimensions $d \leq 3$, 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 to show how the bias acts for large N and different k when $d = 1$.

2.1.1 Estimator with $k = 1$

Firstly, I considered the article *On Statistical Estimation of Entropy of Random Vector* (N.Leonenko and L.Kozachenko, 1987) [22], which studies the estimation of 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, as discussed in Section 1.2.3.

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

$$H_N = d\log(\bar{\rho}) + \log(c(d)) + \log(\gamma) + \log(N-1) \tag{2.1}$$

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 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 the negative Euler constant. Thus this estimator can be written in the following form:

$$\begin{aligned} H_N &= \log(\bar{\rho}^d) + \log(c(d)) - \Psi(1) + \log(N-1) \\ &= \log\left(\left[\prod_{i=1}^{N} \rho_i\right]^{\frac{d}{N}}\right) \log(c(d)(N-1)) + \log\left(\frac{1}{\exp(\Psi(1))}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \log(\rho_i^d) + \log\left(\frac{c(d)(N-1)}{\exp(\Psi(1))}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \log(\rho_i^d) + \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{c(d)(N-1)}{\exp(\Psi(1))}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\rho_i^d c(d)(N-1)}{\exp(\Psi(1))}\right) \end{aligned} \tag{2.2}$$

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.

2.1.2 Estimator with fixed k

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) [23], 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. Here, these entropies 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 value 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, $i \neq j$.

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:

$$\hat{H}_q^* = \frac{1}{1-q} log(\hat{I}_{N,k,q}) \tag{2.3}$$

$$\hat{H}_q = \frac{1}{q-1}(1 - \hat{I}_{N,k,q}) \tag{2.4}$$

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 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 (2.2); however, it is now extended from the first nearest neighbour to the kth nearest neighbour:

$$\hat{H}_{N,k,1} = \frac{1}{N}\sum_{i=1}^{N} \log(\xi_{N,i,k}) \tag{2.5}$$

where $\xi_{N,i,k} = (N-1)\exp[-\Psi(k)]V_d(\rho_{k,N-1}^{(i)})^d$, with V_d and $\rho_{k,N-1}^{(i)}$ 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 $q_1 > 1$, H 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

$\hat{H}_{N,k,1} \xrightarrow{L_2} H$, which implies that as $N \to \infty$, both $N^{\frac{1}{2}}(\hat{H}_{N,k,1} - H) \xrightarrow{d} N(0, \sigma^2)$ - it's asymptotically efficient - and $\mathbb{E}(\hat{H}_{N,k,1}) \to H$ - it is asymptotically unbiased.

2.1.3 Estimator with k dependent on sample size, N

The last main paper,being explored is *Efficient Multivariate Entropy Estimation via k-Nearest Neighbour Distances* (T.Berrett, R.Samworth, M.Yuan, 2016) [4], 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).

Considering dimensions $d \leq 3$, and a sample of size N from a continuous d-dimensional distribution with density $f(x)$, they defined the k-nearest neighbour estimator of entropy - just as in section 2.1.2 - to be:

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\rho_{(k),i}^{d} V_d (N-1)}{e^{\Psi(k)}} \right] \tag{2.6}$$

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 differentiability 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:

$$\mathbb{E}(\hat{H}_N) - H = O\left(max \left\{ \frac{k^{\frac{\alpha}{\alpha+d} - \epsilon}}{N^{\frac{\alpha}{\alpha+d} - \epsilon}}, \frac{k^{\frac{\beta}{d}}}{N^{\frac{\beta}{d}}} \right\} \right) \qquad N \to \infty \tag{2.7}$$

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 differentiability of f hold and with $k \in \{k_0, ..., k_1\}$. Then the variance of the estimator is given by:

$$Var(\hat{H}_{N,k}) = \frac{\sigma^2}{N} + o(\frac{1}{N}) \tag{2.8}$$

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 [4] 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:

$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{2.9}$$

and

$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to \sigma^2 \tag{2.10}$$

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

It is important to note that for higher dimensions ($d > 3$), these results do not necessarily hold; since just the specific dimension $d = 1$ is being considered, 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 1.2.3.

2.2 Focus of this paper

I now wish to more explicitly introduce the K-L estimator of the entropy H, in the form that I will be considering. For this consider a sample $X_1, X_2, ..., X_N$, $N \geq 1$ of 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 $\|X_{(1),i} - X_i\| \leq \cdots \leq \|X_{(N-1),i} - X_i\|$. Thus, let the metric ρ, defined as:

$$\rho_{(k),i} = \|X_{(k),i} - X_i\| \tag{2.11}$$

  denote the distance between X_i and its kth nearest neighbour.

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

$$V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(1 + \frac{d}{2})} \tag{2.12}$$

where the Γ function is defined as:

$$\Gamma(m) = \int_0^\infty x^{m-1} e^{-x} dx \tag{2.13}$$

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

$$\Psi(k) = -\gamma + \sum_{j=1}^{k-1} \frac{1}{j} \tag{2.14}$$

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 K-L estimator for the Shannon entropy, H, is given by:

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\rho_{(k),i}^d V_d (N-1)}{e^{\Psi(k)}} \right] \tag{2.15}$$

where, d is dimension, N the sample size, $\rho_{(k),i}^d$ is defined in (2.11), V_d is defined in (2.12) and $\Psi(k)$ is defined in (2.14).

This paper focuses only on distributions for which $d = 1$, 1-dimensional distributions. 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 K-L estimator takes the following form:

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{2\rho_{(k),i} (N-1)}{e^{\Psi(k)}} \right] \tag{2.16}$$

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 2.1.3, where we have the following 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 1-dimension, since that is what is considered here.)

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:*

$$\mathbb{E}(\hat{H}_N) - H = O\left(max\left\{\frac{k^{\frac{\alpha}{\alpha+d}-\epsilon}}{N^{\frac{\alpha}{\alpha+d}-\epsilon}}, \frac{k^{\frac{\beta}{d}}}{N^{\frac{\beta}{d}}}\right\}\right) \tag{2.17}$$

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:*

$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{2.18}$$

and

$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to \sigma^2 \tag{2.19}$$

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 the -L estimator, $\hat{H}_{N,k}$, that satisfies Conditions 1, 2, 3 is a consistent and asymptotically unbiased estimator of exact entropy H; thus is a efficient estimator. This is due to using the central limit theorem, on the estimator for entropy $\hat{H}_{N,k}$ [30], which states that:

$$\frac{\hat{H}_{N,k} - \mathbb{E}\hat{H}_{N,k}}{\sqrt{Var(\hat{H}_{N,k})}} \xrightarrow{d} N(0,1)$$

By section 2.1.3, we can assume that $Var(\hat{H}_{N,k}) = \frac{Var(\log f(x))}{N} + O(\frac{1}{N}) \approx \frac{\sigma^2}{N}$. Accordingly, the left side of the central limit theorem above can be written as:

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

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 $\mathbb{E}\hat{H}_{N,k} - H \to 0$ as $N \to \infty$. Thus as $N \to \infty$, this tends to the standard normal distribution, $N(0,1)$, and the Central Limit Theorem holds.

I will be exploring the bias in more detail later to see, analytically, which one of the two ideas show to be more true for the behaviour of the bias for a large sample size, N, in dimension $d = 1$. Where I will be considering samples from the 1-dimensional normal, uniform and exponential distributions.

- With a fixed k, by [6], for $\beta \in (0,2] \cap (0,d]$, we choose $a \in (0, \frac{\beta}{d}]$, then:

$$Bias(\hat{H}_{N,k}) = O\left(\frac{1}{N^a}\right) \tag{2.20}$$

- With k depending on N, by [4], for $\beta \in (0,2]$, we again choose $a \in (0, \frac{\beta}{d}]$, then:

$$Bias(\hat{H}_{N,k}) = O\left(\left(\frac{k}{N}\right)^a\right) \tag{2.21}$$

Moreover, I will be exploring analytically the optimal value of k, for reducing the bias, for each of these distributions according to simulations from each distribution and regression analysis conducted on them. This value of k may depend on the sample size, N: something which has been discussed theoretically in section 2.1 but will be explored analytically in this paper.

Chapter 3

Monte-Carlo Simulations

In this chapter, I will explore simulations

In this chapter, I will explore simulations on the bias of the K-L estimator (2.15) on a sample from a 1-dimensional distribution, in comparison to its sample size, with respect to different values of k. Firstly, the distributions considered will be analysed to determine if they satisfy Conditions 1, 2 and 3 stated for Theorems 1 and 2 to hold. Then, I will explore in depth the bias of this estimator of entropy for simulations of samples from certain distributions, for different values of k.

The motivation for these simulations is to analytically 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. 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 (2.17) implies $Bias(\hat{H}_{N,k}) \to 0$ as $N \to \infty$. Thus, there must be a type of inverse relationship between the bias of the estimator, $Bias(\hat{H}_{N,k})$, and N. We believe this relationship is of the form:

$$Bias(\hat{H}_{N,k}) = \frac{c}{N^a} \tag{3.1}$$

for $a, c > 0$ [6, 4]. To make this into a linear relationship, which is easier to analyse, we will be considering the logarithm of equation 3.1. For this to be defined, we must take the absolute value of the bias, since it must be positive to take the logarithm.

We obtain the following:

$$\begin{aligned} log|Bias(\hat{H}_{N,k})| &\approx log(c) - a[log(N)] + \epsilon \\ &\approx \zeta - a[log(N)] \end{aligned} \tag{3.2}$$

where $\epsilon > 0$ is some small error term. I will investigate the consistency of this estimator for a sample from a specified distribution, dependent on the value of k. This involves finding the optimum value of k for which $|Bias(\hat{H}_{N,k})| \to 0$ for $N \to \infty$. For the relationship in equation (3.1), 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 $(\frac{1}{N})^a$ (2.20) or $(\frac{k}{N})^a$ (2.21). This leads to also examining the dependence of c / ζ 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, for each k, as a_k and c_k respectively.

I will conduct a range of analyses, for each distribution, to consider how this estimator acts in reality, the process of analysis will be as follows:

1. Create a summary table of the mean absolute value of the bias of the estimator just observing the sample sizes $N = 100, 25000$ and 50000 for all values of k that satisfy Condition 3. I could also at this point consider the variance of the bias at the values of N stated above, for all applicable values of k. However, we will find that the $Var|Bias(\hat{H}_{50000,k})| \to 0$ for $k \to 10$, 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.

2. From simulations run to obtain the data, we can plot graphical representations of the linear relationship shown in equation 3.2, of $log(N)$ against $log|Bias(\hat{H}_{N,k})|$ for sample sizes $N = 100, 200, 300, ..., 50000$, for each value of k. I will take 500

samples of size N, and find the average estimator for each sample size, making use of the Mean Value Theorem to attain a more accurate representation of the estimator from that sample size.

3. Tabulate the results from the regression analysis; I will first discuss the coefficient of determination (R^2), a measure of how well the regression model describes the observed data [28]. Next I will consider the standard error/deviation of the model (σ), this is a measure of accuracy of predictions. Lastly, I will go onto consider the values of a_k and c_k from the relationship shown in equation 3.1, for each k, which is the regression line that minimizes the sum of squared deviations (σ) of prediction.

4. Lastly, graphically compare the values of a_k and c_k, for each k. From this I can examine the results for the optimal value of k, and can hopefully explore the relationship between c_k and k.

3.1 1-dimensional Gaussian/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 the definition of the Shannon entropy for a continuous distribution (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 $x \in \mathbb{R}$, when $\mu = 0$. We can write the exact entropy for

the normal distribution:

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

Thus the exact entropy for the normal distribution is given by:

$$H = log(\sqrt{(2\pi e)}\sigma) \tag{3.3}$$

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 K-L estimator. The exact entropy of this distribution is given by equation 3.3, with $\sigma^2 = 1$:

$$H = log(\sqrt{(2\pi e)}) \approx 1.418939 \tag{3.4}$$

I am considering the 1-dimensional normal distribution; whence, the estimator takes the form in equation 2.16:

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

3.1.1 Estimator conditions

The analysis set out below demonstrates how the density of the normal distribution satisfies 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 the density $f(x) = \frac{1}{\sqrt{(2\pi)}} \exp\left(\frac{-x^2}{2}\right) < 0.4$, $\forall x \in \mathbb{R}$. Hence, f is bounded above and below - so is 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:

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

Thus, we have the following for the mth derivative of f:

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

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 $\|x - y\| \leq r_*$ so that $\|f^{(t)}(x)\| \leq g_*(x) f(x)$ and $\|f^{(m)}(x) - f^{(m)}(y)\| \leq g_*(x) f(x) \|x - y\|^\eta$, 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, for Theorems 1 and 2 we have that the value of $\beta \geq 2$. Choosing $\beta = 2$ this is satisfied, so $m = \lfloor \beta \rfloor = \lfloor 2 \rfloor = 2 = \beta$ and $\eta = \beta - m = 0$. Thus to have $|f^{(t)}(x)| \leq g_*(x) f(x)$, 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 $|f^{(m)}(x) - f^{(m)}(y)| \leq g_*(x)f(x)|x - y|^0 = g_*(x)f(x)$. We currently have:

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

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

Therefore, the normal distribution satisfies Condition 1. Next, to satisfy Condition 2, the density function f of the normal distribution must fulfill the following:

- The α-moment of f must be finite; $\int_{\mathbb{R}^d} \|x\|^\alpha f(x)dx < \infty$ - 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 1 and 2. To do this we must have, let k_0^* and k_1^* denote two deterministic sequences of positive integers with $k_0^* \leq k_1^*$. We also have dimension $d = 1$ and $\alpha > d = 1$, so choosing $\alpha := 2$, we must have the following:

- $k_1^* = O(N^\tau)$, 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 $A := \frac{1}{\omega}$ for some large ω, which gives that $k_0^* = O(N^{\frac{1}{\omega}}) \approx 1$ for any $\omega > 0$.

Thus, on account of the sample size, 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 1 and 2 both hold, are $k \in \{k_0^*, ..., k_1^*\} = \{1, ..., 100^{\frac{2}{9}}\} = \{1, ..., 2.782\}$, so $k \in \{1, 2\}$. Also, for the middle sample size $N = 25,000$, we have the values of k to be in $\{k_0^*, ..., k_1^*\}$, where $k_1^* \approx 25000^{\frac{2}{9}} = 9.491 \approx 9$, thus $k \in \{1, ..., 9\}$. Moreover, for the largest sample size being considered, $N = 50,000$, we must have $k \in \{1, ..., k_1^*\} = \{1, ..., 50000^{\frac{2}{9}}\} = \{1, ..., 11.072\}$; thus, $k \in \{1, 2, ..., 11\}$.

Overall, due to Conditions 1, 2 and 3 being met, we can say that for the normal distribution, Theorems 1 and 2 hold. Henceforth, the K-L estimator, of a sample from the 1-dimensional normal distribution is an asymptotically unbiased and consistent estimator for entropy, for some values of $k \in \{1, 2, ..., 11\}$, depending on the sample size N.

3.1.2 Simulation results

I will now conduct some simulations to analyse the K-L estimator for each value of k separately. For each N, I will be considering 500 samples from this distribution, finding the estimator in each case and taking the average of these estimators to find our entropy estimator. I will then consider the relationship shown in equation 3.2 of the logarithm of the bias against the sample size, for each sample and work out the analytical values of a and c, for each $k \in \{1, 2, ..., 11\}$.

For $N = 100$, $N = 25,000$ and $N = 50,000$, using the results from Appendix B, we can create a table to compare the mean values of the bias of the estimator for the different values of k considered. These results are in table 3.1, and show that for a larger N, the modulus of the bias of the estimator is smaller, this is true for all values of k except when $N = 25,000$ in comparison to the larger value of N, where for $k = 2, 3, 7, 8$ we find the bias is smaller. There are a number of reasons why this could be; mainly due to the errors that can occur in simulations and the values of k that satisfy Theorems 1 and 2 for each sample size N. Thus, it is first important to notice that when finding the values of k that satisfy Condition 3, we found that for $N = 100$, we must have $k \in \{1, 2\}$, for $N = 25,000$ we have $k \in \{1, 2, ..., 9\}$ and for

Table 3.1: *1-dimensional normal distribution, comparison of k*

k	$\lvert Bias(\hat{H}_{100,k})\rvert$	$\lvert Bias(\hat{H}_{25000,k})\rvert$	$\lvert Bias(\hat{H}_{50000,k})\rvert$
1	0.0031912	0.0006312	0.0004428
2	0.0195347	0.0000092	0.0003632
3	0.0167902	0.0000056	0.0002278
4	0.0264708	0.0001657	0.0001196
5	0.0238265	0.0002138	0.0000003
6	0.0311576	0.0001546	0.0001471
7	0.0356302	0.0000217	0.0003024
8	0.0396299	0.0000984	0.0001021
9	0.0460706	0.0003620	0.0002070
10	0.0458648	0.0002752	0.0002611
11	0.0387339	0.0003332	0.0002458

Comparing the values of $\lvert Bias(\hat{H}_{N,k})\rvert$ *for the values of* k *with* $N = 100$, $N = 25,000$ *and* $N = 50,000$, *when the estimator is taken over* 500 *samples.*

$N = 50,000$ we have $k \in \{1, 2, ..., 11\}$.

For the smallest sample size taken; $N = 100$, we expect the best value of k to be either 1 or 2; and the table agrees with this showing that the smallest bias occurs at $k = 1$ for this sample size.

When $N = 25,000$ we have that for $k \in \{2, ..., 8\}$ that the bias is very small, especially for the values of $k = 3, 4, 7, 8$ with the smallest bias appearing when $k = 3$. This fits with the previous analysis that the best value of k will lie within 1 and 9.

Now considering the largest sample size, $N = 50,000$, the bias when $k = 5$ differs significantly since it is $\approx 10^{-3}$ smaller than the other bias values in the table. However, for all other values of k, the bias is still extremely small in comparison to the bias for $N = 100$, and even in comparison to $N = 25,000$ in some places. This extreme difference could be an outlier in my data; thus in table 3.2 I have shown the values for the modulus of the bias, when $k = 5$, for different large values of N. This table does indeed show that $|Bias(\hat{H}_{50000,5})| \approx 0.0000003$ is an anomaly in the data, and that $k = 5$ is not necessarily the best value of k for $N \approx 50,000$. Thus, we cannot yet draw any major conclusions about the best value of k for the estimator of a sample this size.

I now wish to consider the equation 3.2 and plot the simulated data, to fit a regression line for each value of k separately, as shown in Figures 3.1 and 3.2. All of these graphs agree with the relationship previously stated between the sample size and the bias of the estimator; they all show that the logarithm of this equation gives a negative linear relationship - with relatively small error bars.

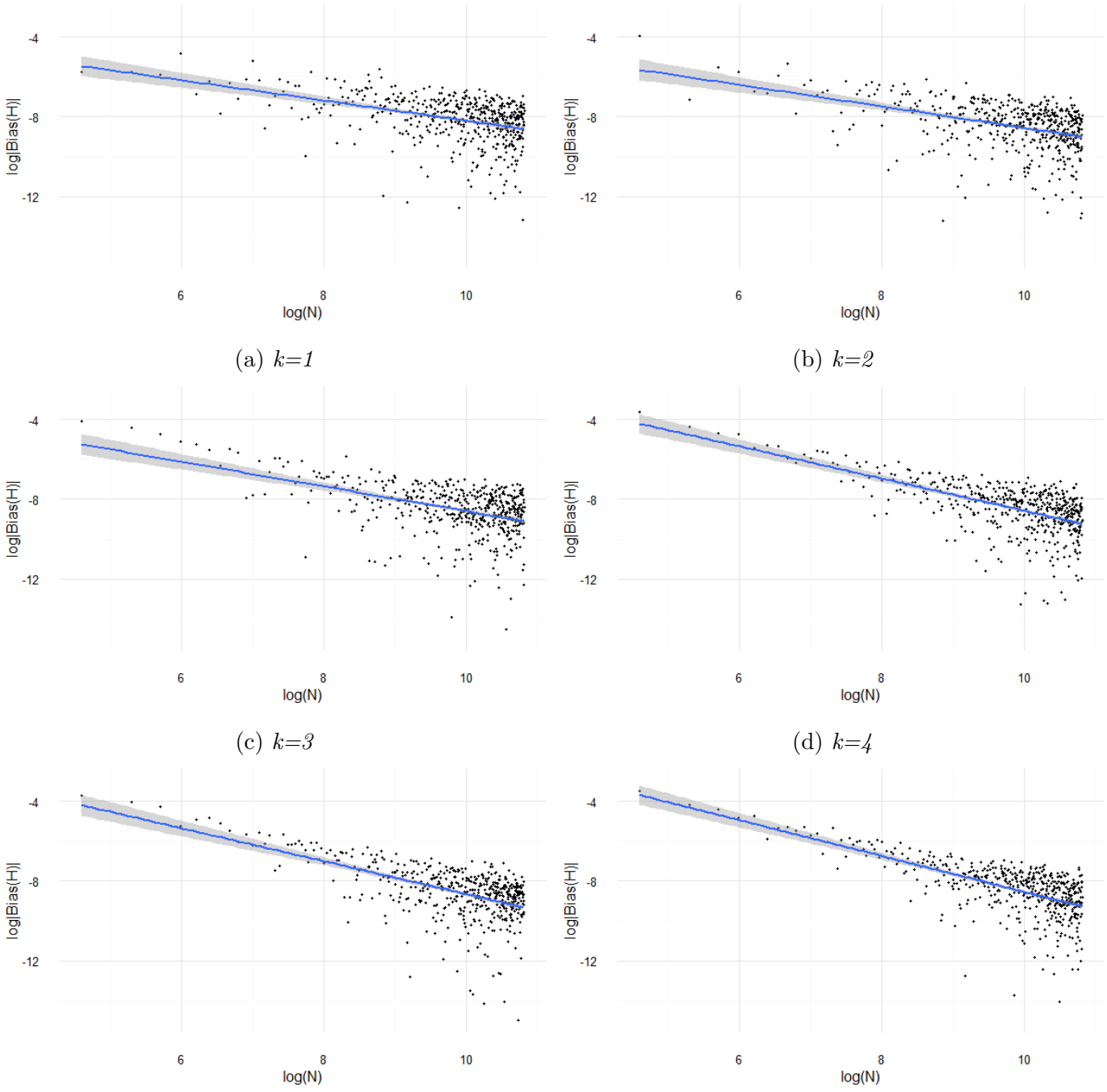
For this, I would like to consider the coefficient of determination (R^2) for each of the above regression lines; this value provides an estimate of the strength of the relationship between the model and the response variable. Also, I would like to consider the standard error/deviation (σ), for each of the different graphs, which shows a measure of the predictions' accuracy. These are depicted for each value of k in table 3.3.

Both columns of this table essentially point to the same conclusion; the larger the value of k, the more accurate the linear model is to fitting the data. This is shown by the R^2 value increasing towards 1 and the σ values decreasing positively.

Table 3.2: *1-dimensional normal distribution,* $k = 5$ *for large* N

N	$\lvert Bias(\hat{H}_{N,5})\rvert$
49100	0.0000639
49200	0.0001463
49300	0.0001700
49400	0.0001037
49500	0.0000711
49600	0.0003221
49700	0.0001047
49800	0.0000644
49900	0.0001240
50000	0.0000003

Comparing the values of $|Bias(\hat{H}_{N,5})|$ *for the large sample sizes* $N \approx 50,000$.

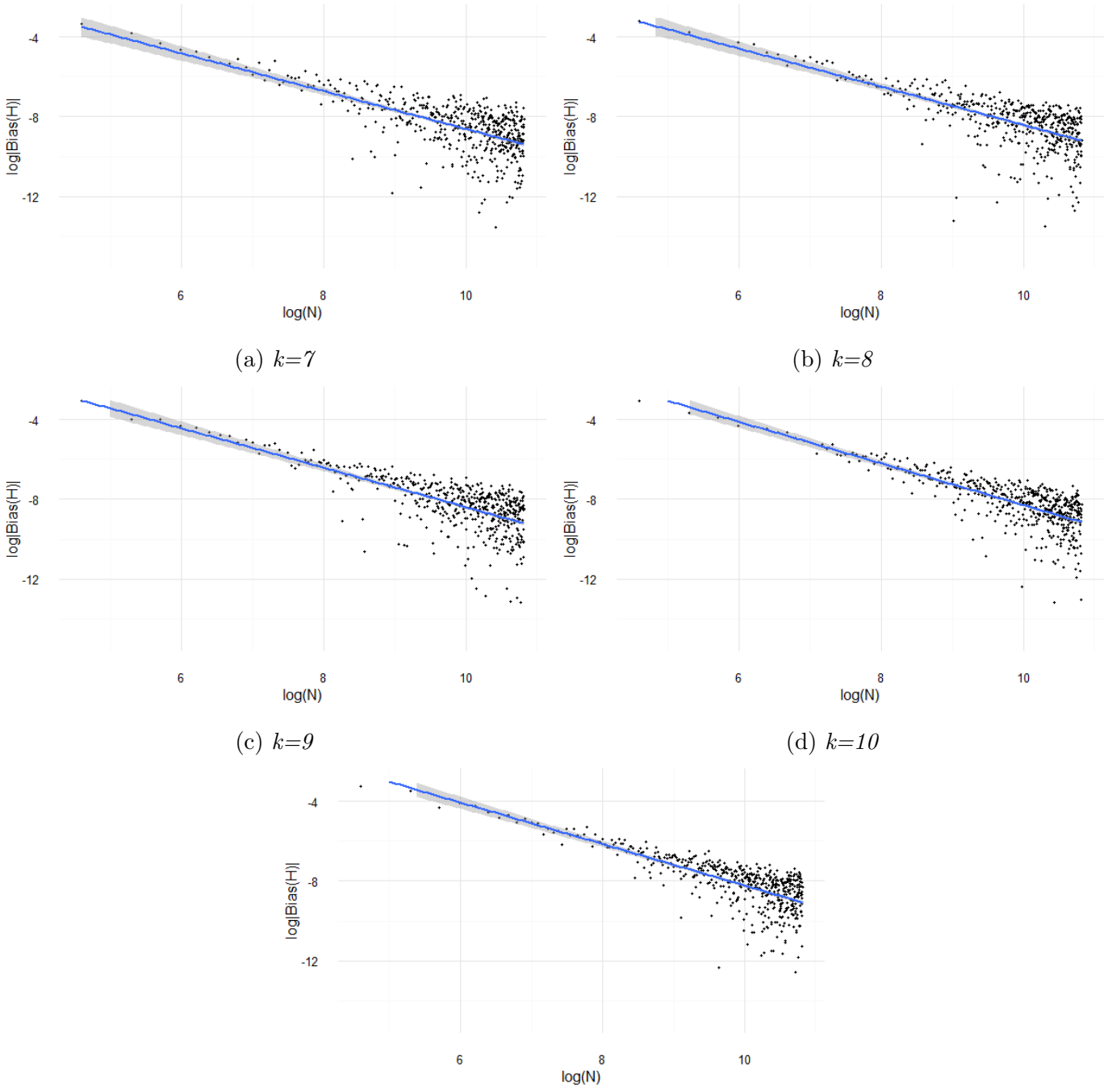


(a) *k=1* (b) *k=2*

(c) *k=3* (d) *k=4*

(e) *k=5* (f) *k=6*

Figure 3.1: *1-dimensional normal distribution with different* $k = 1, ..., 6$



(a) *k=7*

(b) *k=8*

(c) *k=9*

(d) *k=10*

(e) *k=11*

Figure 3.2: *1-dimensional normal distribution with different* $k = 7, ..., 11$

Table 3.3: *Comparison of the coefficient of determination and the standard deviations of the regression for each value of k for the 1-dimensional normal distribution*

k	R^2	σ
1	0.1766	1.0661
2	0.1793	1.1477
3	0.2292	1.1053
4	0.3556	1.0759
5	0.3322	1.1752
6	0.4260	1.0180
7	0.4532	1.0155
8	0.4623	1.0088
9	0.4962	0.9730
10	0.5227	0.9759
11	0.5839	0.8566

Table 3.4: *Comparison of coefficients of regression a_k and c_k from equation 3.1, 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

The R^2 is very small for $k \leq 3$, which points towards the line being a poor fir to the data; however, due the the standard deviation being $\sigma^2 \approx 1.1$, we cannot say that these lines are poorly fitting; since the majoring of the data is within a very small range of the line.

The most important information found from the regression analysis is shown in table 3.4; where the values of a_k and c_k are given for each value of k.

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, in the value of c_k, between $k = 3$ and 4, and $k = 9$ and 10. The higher the value of a_k, the stronger the negative relationship is between the two variables in question. So for 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 (3.1)

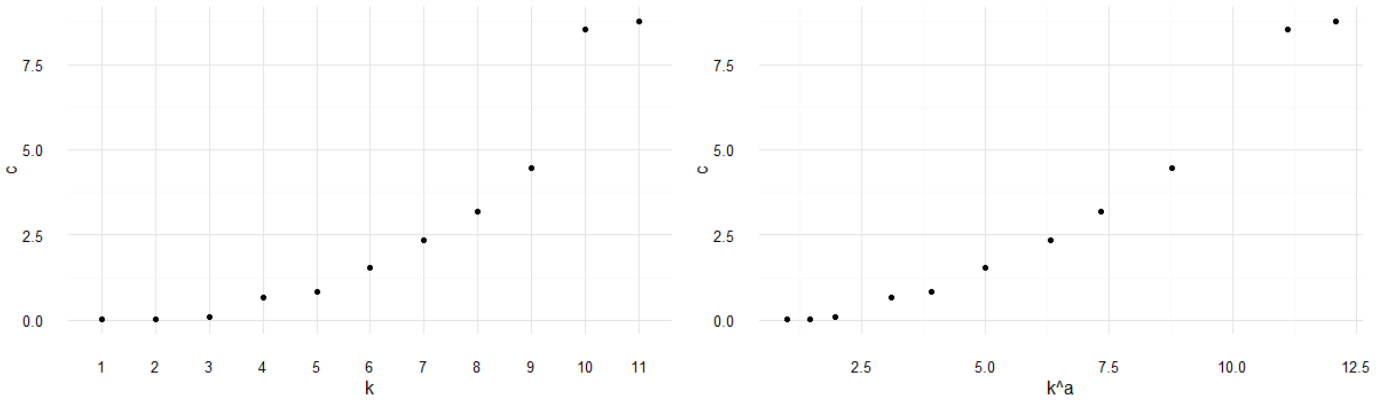
Recall, from section 2.2 we have that the bias acts in one of two ways; either of

Table 3.5: *Considering the dependence of k on c_k*

k	k^{a_k}	c_k	$\frac{k^{a_k}}{c_k}$
1	1	0.0433	23.095
2	1.4631	0.0459	31.875
3	1.9694	0.0894	22.029
4	3.1085	0.6690	4.646
5	3.9187	0.8235	4.759
6	4.9942	1.5514	3.219
7	6.3067	2.3576	2.675
8	7.3218	3.2021	2.287
9	8.7716	4.4558	1.969
10	11.1020	8.5402	1.300
11	12.0668	8.7457	1.380

$O\left(\frac{1}{N^a}\right)$ (2.20) or $O\left(\left(\frac{k}{N}\right)^a\right)$ (2.21). Thus, we have $|Bias(\hat{H}_{N,k})| \approx \frac{c_k}{N^{a_k}}$ where either c_k is constant or it depends on k and a_k - more specifically is $O(k^{a_k})$. There is evidence here to support the latter claim: the jump between $k = 3$ and 4 shown in the value of c_k, and consider the results in table 3.5. This shows that the proportional behaviour between k^{a_k} and c_k also has a large jump when k goes from $3 \to 4$. This agrees with the claim of c_k depending on k in this fashion; however, in table 3.4 we mentioned another jump between $k = 9$ and $k = 10$, and the evidence here does not show a large jump in the same area. We cannot yet make any conclusions about the dependence of c_k on k; this motivates a graphical representation of the value of c_k against k to see if there is any relation, Figure 3.3.

Interestingly, the plot from Figure 3.3 (a) shows an almost exponential or parabolic relationship between the values of c_k and the values of k. This leads me to believe that there is some kind of relationship between the two variables, and looking at the graph in Figure 3.3(b) this shows that there's a strong possibility that the relationship



(a) *The values of k against c_k*

(b) *k^{a_k} against the corresponding values of c_k*

Figure 3.3: *Graphically representing the relationship between c_k and k*

is of the form stated in equation 2.21.

To better study the linear relationship between the logarithm of the bias and the logarithm of the sample size, I have generated a comparison plot, shown in Figure 3.4.

From this, we can obviously see that for smaller values of N (i.e - smaller values of $log(N)$), the smallest bias occurs when $k = 2$, since this line is the lowest for the data up until $log(N) \approx 9$ (i.e. $N \approx 13,000$). For a larger sample size, we cannot accurately see in this graph which line is the best. This motivates us to look at a section of the graph when $9 \leq log(N) \leq 11$ - i.e. $8,000 \leq N \leq 50,000$ - Figure 3.5.

From this graph, we can obviously discount $k = 1$ for large N, since this is the most gradual descent; thus the bias will be largest for this k. Also, both the lines for $k = 2$ and $k = 3$ are more gradual in their descent at larger N, so are probably not the best to choose. Even though, for $k = 9, 10$ and 11, the slope is the steepest - a_k is largest - the intercept is larger so around the biggest sample size considered $N = 50,000$, $log(N) \approx 10.8$, the smallest bias does not occur. Actually, for large values of $N \leq 50,000$ we can see from this graph that the best lines appear to be those which are blue/green; $k = 4, 5, 6, 7, 8$. The lowest lines around the maximal sample size are those for $k = 5$ and $k = 7$; thus these values of k analytically appears to be the best nearest neighbour value to choose, when looking at a sample of size

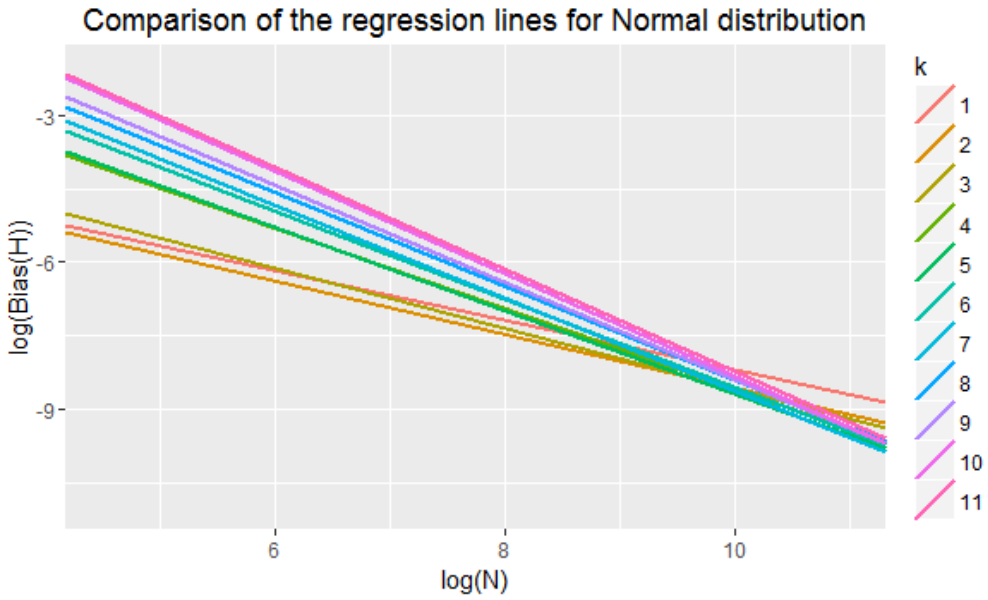


Figure 3.4: *Plot of regression lines for* $\log|Bias(\hat{H}_{N,k})|$ *against* $\log(N)$*, for* $k = 1, 2, ..., 11$*, for samples from the normal distribution*

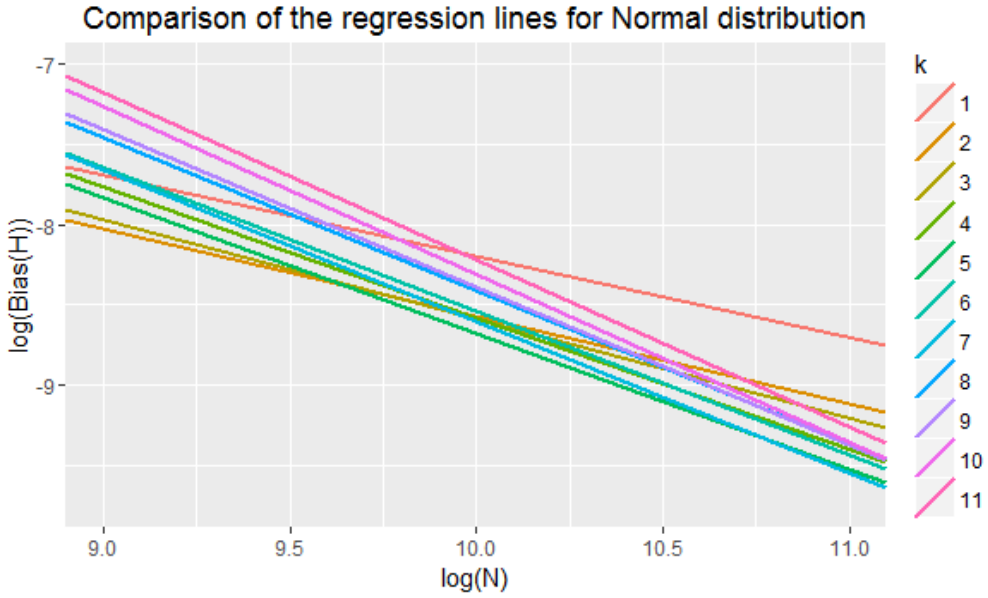


Figure 3.5: *Figure 3.4 zoomed in around large* N

$N \approx 50,000$ from the normal distribution.

3.2 1-dimensional Uniform Distribution

I will now explore the entropy of samples from the 1-dimensional continuous uniform distribution, $U[a, b]$. This distribution also has an exact formula for the entropy. We can find this formula by considering the density function, f, from the uniform distribution, given by:

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

Using the definition of Shannon entropy (1.1), we can find the exact entropy for the uniform distribution:

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

Thus, the exact value of entropy for the uniform distribution is given by:

$$H = log[b-a] \tag{3.5}$$

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 (2.16):

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

Moreover, the samples considered will not be from the standard uniform, but from the 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 density function would be $f(x) = 1,\ 0 \leq x \leq 1$. This would incur working on a very small scale (i.e taking a points with distance between them) causing problems due to the limitations that occur when using computer programs to run simulations. Thus, I will be using the density function $f(x) = 0.01,\ 0 \leq x \leq 100$, which is from the $U[0, 100]$ distribution and gives the exact entropy of:

$$H = log(100) \approx 4.605170 \tag{3.6}$$

3.2.1 Estimator conditions

For Theorems 1 and 2 to be satisfied by the estimators generated by samples from the uniform distribution, this density function 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 uniform distribution is constant for $x \in [a, b]$ and 0 otherwise; hence is bounded.
- f is m-times differentiable - as f is constant this holds.
- $\exists r_* > 0$ 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 $\epsilon > 0$. This is elementary since f is constant, thus $f^{(t)}(x) = 0$ for all $t \geq 1$, so we can always find an appropriate $g_*(x)$.

Next, to satisfy Condition 2, the density function f of the uniform distribution, must fulfill the following:

- 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.

Lastly, to satisfy Condition 3, we must find the values of k for which the estimator provides a uniform convergence for Theorems 1 and 2. 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. The values of k found in section 3.1.1, for the normal distribution, are $\{1, 2\}$ for $N = 100$, $\{1, 2, ..., 9\}$ for $N = 25,000$ and $\{1, 2, ..., 11\}$ for $N = 50,000$. These are the same for this distribution, and for all distributions considered.

Thus, due to the above conditions for Theorems 1 and 2 being met, we can say that the K-L estimator of a sample from the uniform distribution is an asymptotically unbiased and consistent estimator for entropy, for specific $k \in \{1, 2, ..., 11\}$, depending on the sample size, N.

3.2.2 Simulation results

I will now conduct simulations, in the same manner as for the normal distribution. For each value of k separately, I will consider 500 samples of size N, from the uniform distribution, finding the estimator in each case and taking the average of these estimators to find our entropy estimator. I will then consider the relationship between the bias of the estimator and the sample size (3.2) for each sample and work out the average for the values of a and c, for each $k \in \{1, 2, ..., 11\}$.

For $N = 100$, $N = 25,000$ and $N = 50,000$, using the results from Appendix B, we can create a table to compare the mean values of the bias of the estimator for the different values of k considered.

Looking at the results in table 3.6, for $N = 100$, we can see that the smallest bias, quite obviously, occurs when the estimator is taken with $k = 1$, and that for other values of k there is a significant difference in the size of the bias. This agrees with the theoretical result, that for a sample size $N = 100$, we must have $k \in \{1, 2\}$.

When looking at the larger values of N, it is first important to note that the value `Inf` $= \infty$ is present in the table as for large sample sizes, N, it is difficult to work out the estimator when $k = 1$. This is because there are only extremely small distances between the closest samples of this distribution, and since any program used

Table 3.6: *1-dimensional uniform distribution, comparison of k*

k	$\lvert Bias(\hat{H}_{100,k})\rvert$	$\lvert Bias(\hat{H}_{25000,k})\rvert$	$\lvert Bias(\hat{H}_{50000,k})\rvert$
1	0.0005189	Inf	Inf
2	0.0047466	0.0001745	0.0001163
3	0.0083912	0.0001776	0.0000899
4	0.0169364	0.0001177	0.0000490
5	0.0152168	0.0000588	0.0000509
6	0.0148205	0.0000817	0.0000538
7	0.0218339	0.0002918	0.0000663
8	0.0250401	0.0001884	0.0000487
9	0.0297655	0.0001406	0.0001184
10	0.0337164	0.0001337	0.0000949
11	0.0381473	0.0001693	0.0000417

Comparing the values of $\lvert Bias(\hat{H}_{N,k})\rvert$ *for the values of* k *with sample size* $N = 100$*,* $N = 25,000$ *and* $N = 50,000$*, when the estimator is taken over* 500 *samples.*

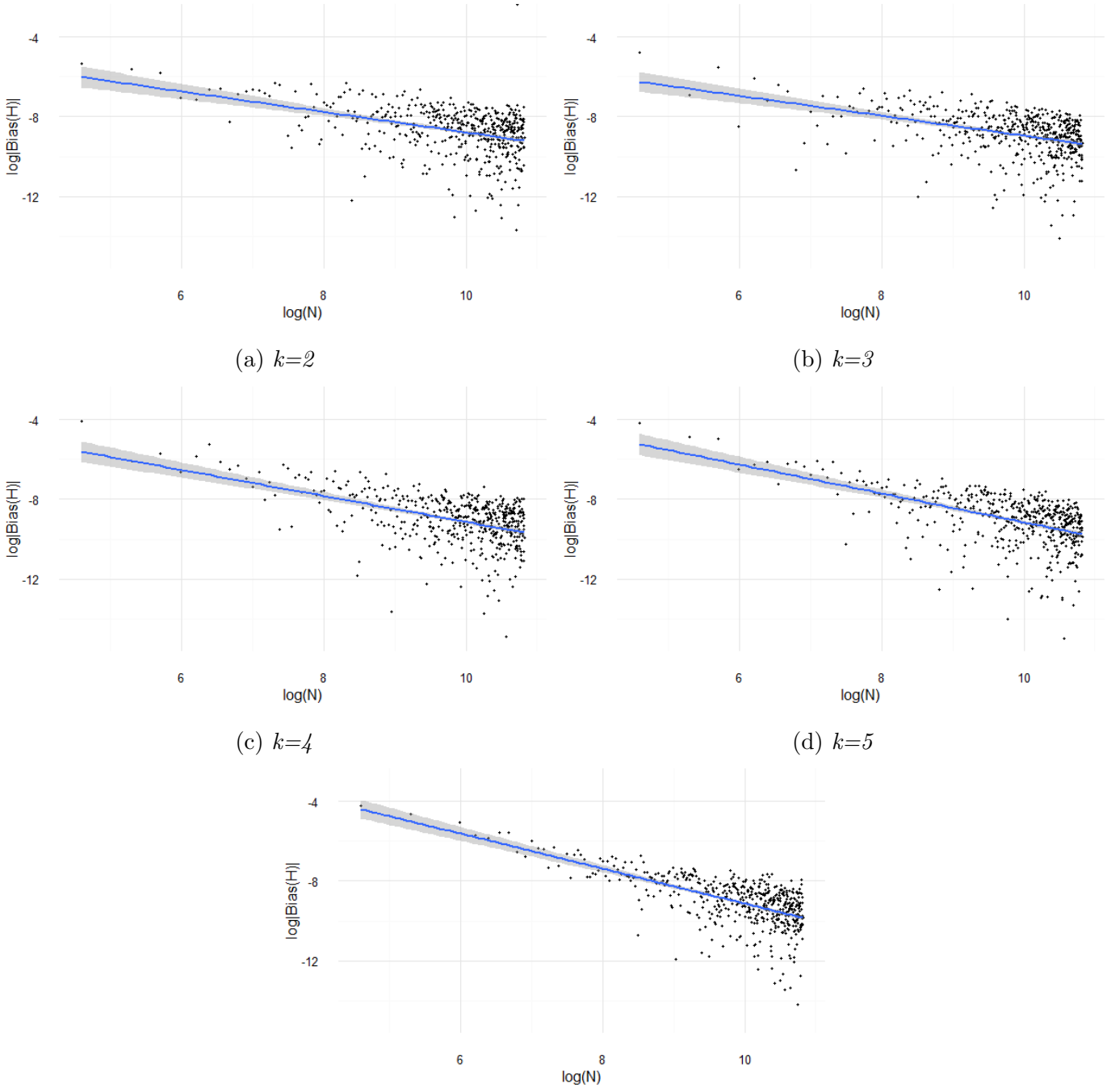
to compute the difference between these numbers will fail as it is extremely small and default sets them to `- Inf`. We only need there to be two samples with an infinitely small distance between one and other to make the whole estimator become `- Inf` (which is `Inf` when looking at the modulus). This was mentioned earlier and was the motivation for taking the uniform distribution over $[0, 100]$; however, because of this, from now onwards I will not be considering $k = 1$ in the analysis on the estimator for samples from the uniform distribution.

Thus, looking at the results shown in table 3.6, ignoring $k = 1$ and considering $N = 25,000$, it appears to show smallest bias for values of $k \in \{4, 5, 6\}$; more specifically $k = 5$ appears to have the smallest bias. This fits with the previous analysis done on the value of k for different sample sizes, which stated that we must have $k \in \{1, 2, ..., 9\}$. Next, examining the table for $N = 50,000$, we can see that the smallest bias looks to be when $k \in \{4, 5, 6, 7, 8, 11\}$, however, for all other values of k we still have a small bias < 0.000012. Thus, we cannot yet draw any conclusions about the optimal value of k for a large sample size, N, from this distribution.

Next I wish to examine the graphs showing the relationship between the logarithm of the bias of the estimator and the logarithm of the sample size (3.2). This is shown by plotting the simulated data and fitting a regression line for each value of k separately, in Figures 3.6 and 3.7. These show a negative linear relationship for all values of k in a similar fashion to the normal distribution, Figures 3.1 and **??**. Also, looking more closely, the regression lines fitted to the data appear to become more steep for higher values of k, whilst the standard error bars appear to become smaller.

Another important thing to consider, before looking at the equations of the regression lines, is to see how well these lines actually fit the data. To do this I have examined the coefficient of determination and the standard deviation of the lines, Table 3.7, where $k = 1$ is removed due to the reasons set out above.

This table is very similar to that for the normal distribution in that both columns point towards the same conclusion; the larger that k is, the more accurate the linear model is to fitting the data. This is shown by the R^2 value generally increasing towards 1 and the σ values decreasing positively - the deviation about the line is decreasing for higher k. There is slight fluctuation in the middle values of k, where



(a) *k=2*

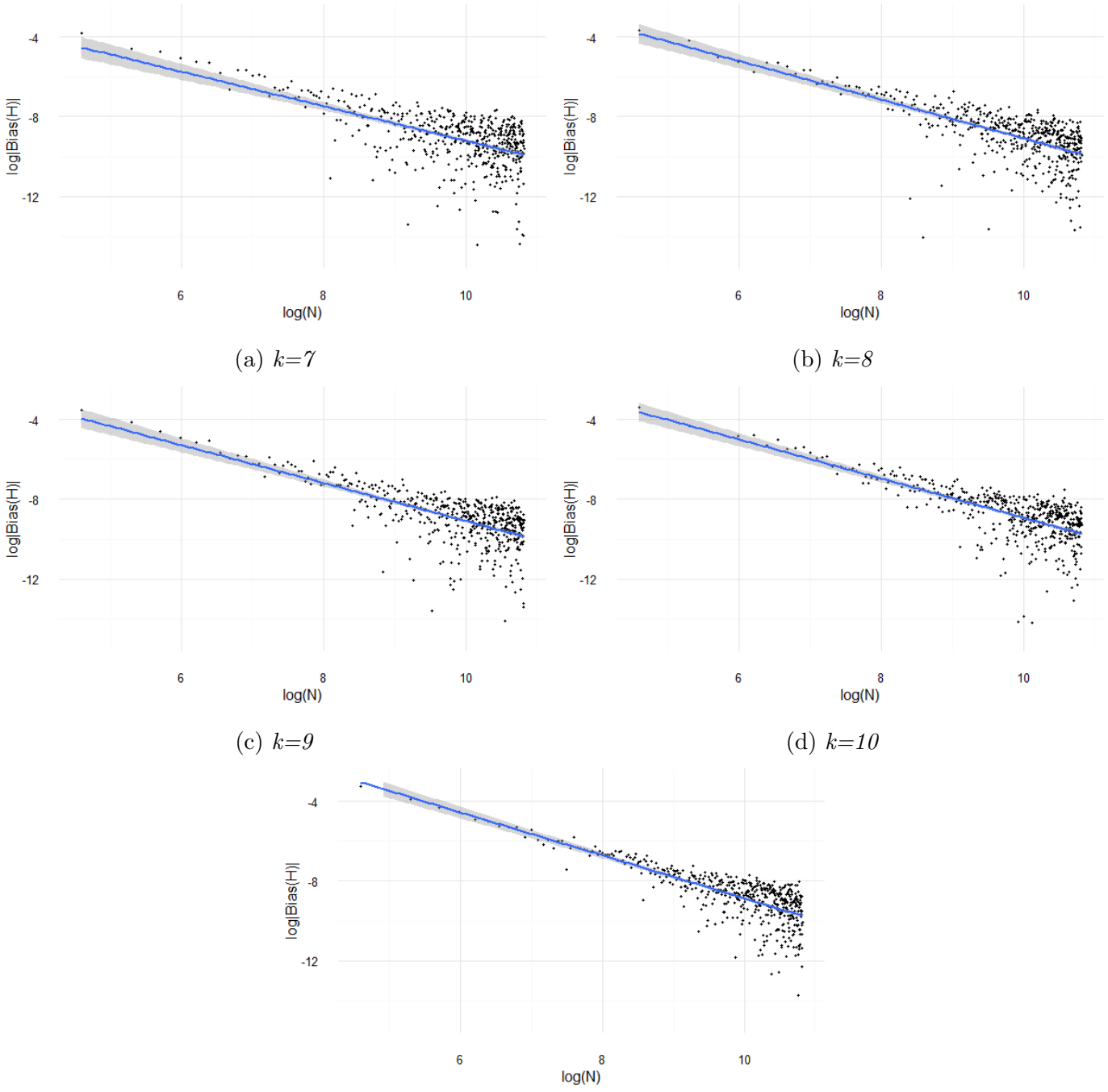
(b) *k=3*

(c) *k=4*

(d) *k=5*

(e) *k=6*

Figure 3.6: *1-dimensional uniform distribution with different* $k = 2, ..., 6$



(a) *k=7*

(b) *k=8*

(c) *k=9*

(d) *k=10*

(e) *k=11*

Figure 3.7: *1-dimensional uniform distribution with different* $k = 7, ..., 11$

Table 3.7: *Comparison of the coefficient of determination and the standard deviations of the regression for each value of k for the 1-dimensional uniform distribution*

k	R^2	σ
2	0.1721	1.0982
3	0.1576	1.1402
4	0.2485	1.1200
5	0.2784	1.1459
6	0.4027	1.0286
7	0.3605	1.1252
8	0.4689	1.0073
9	0.4653	0.9940
10	0.5088	0.9408
11	0.6215	0.8208

Table 3.8: *Comparison of coefficients of regression a_k and c_k from equation 3.1, for 1-dimensional uniform distribution*

k	a_k	c_k
2	0.5125	0.0258
3	0.5048	0.0199
4	0.6593	0.0779
5	0.7286	0.1531
6	0.8645	0.6090
7	0.8648	0.5758
8	0.9688	1.8377
9	0.9492	1.5095
10	0.9801	2.4041
11	1.0765	6.6932

for $k = \{6, 7, 8\}$ there is not an exact direction that the relationship is going. However, there is nothing to say that this is not normal behaviour, since we do not yet know the relationship between the bias and k.

Moreover, R^2 is very small (≤ 0.25) for $k \leq 4$, which points towards the line being a poor fit to the data; however, due the the standard deviation being $\sigma \approx 1.1$, we cannot discard the importance of these lines since most of the data is in a very small range. Additionally, for $k = 11$ we have the strongest relationship seen so far - the largest R^2 and smallest σ values appear. This implies that for this distribution, when $k = 11$, the linear relationship between the logarithm of the bias and the logarithm of the sample size occurs in a stronger manner than for the other k and distributions considered thus far.

Possibly the most important information found from the regression analysis, is shown in Table 3.8; where the values of a_k and c_k are given for each value of $k = \{2, 3, ..., 11\}$.

Table 3.9: *Considering the dependence of k on c_k*

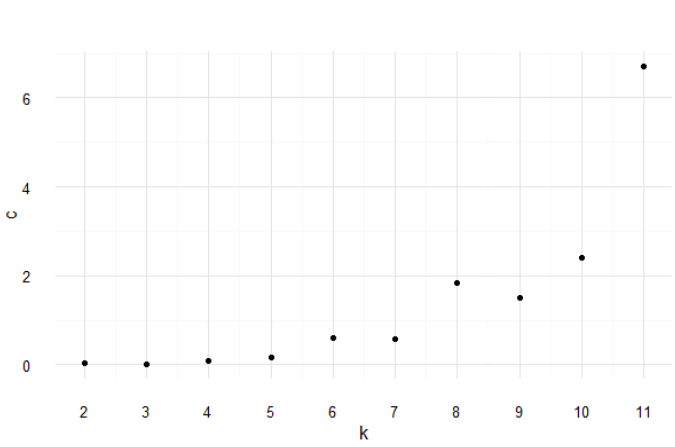
k	k^{a_k}	c_k	$\frac{k^{a_k}}{c_k}$
2	1.4265	0.0258	55.291
3	1.7412	0.0199	87.498
4	2.4942	0.0779	32.019
5	3.2305	0.1531	21.101
6	4.7066	0.6090	7.729
7	5.3807	0.5758	9.345
8	7.4975	1.8377	4.080
9	8.0495	1.5095	5.332
10	9.5521	2.4041	3.973
11	13.2148	6.6932	1.974

We wish to have a value of $a_k \geq 0.5$ to show the relationship desired, and this is true for all values of k in this table. As k increases, we have that both a_k and c_k increase, expect for $k = 8$, which has larger values that $k = 9$. However, this slight change around $k = 8$ does not necessarily imply anything dramatic since the overall trend seems to fit with that found for the normal distribution.

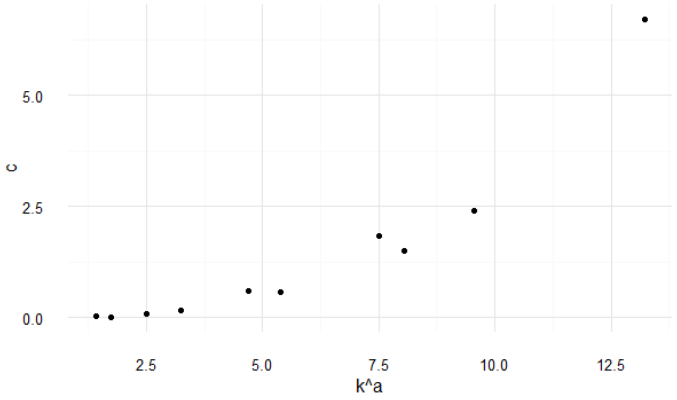
To better see if the relationship of the bias of the estimator and the sample size is of the form of $O\left(\frac{1}{N^a}\right)$ (2.20) or $O\left(\left(\frac{k}{N}\right)^a\right)$ (2.21), I will consider the value of c_k for each k, to see if it is dependent on k^{a_k} or not, shown in Table 3.9.

This shows that the proportional behaviour between k^{a_k} and c_k does not seem to imply that $c_k = O(k^{a_k})$, since there is no exact trend in the numbers, other than a slight decrease as k increases. However, this increase is not uniform so we can make little assumptions about the behaviour of c_k, perhaps a better way to show this relationship is through a graphical representation of c_k against k to see what form it takes.

Figure 3.8, for the uniform distribution tells a slightly different story to that of



(a) *The values of k against the values of c_k*



(b) *The values of k^{a_k} against the corresponding values of c_k*

Figure 3.8: *Graphically representing the relationship between c_k and k for the uniform distribution*

the normal. It does seem to show an almost exponential or parabolic relationship, as before; however, there are a large number of outlier values that do not fit within a smooth line in these graphs. This is different in comparison to the smoother relationship shown in those for the normal distribution, Figure 3.3. However, it does imply some correspondence between the two variables which leads me to believe that there is something important between them, which could possibly be of the $O\left(\left(\frac{k}{N}\right)^a\right)$ (2.21).

To better study the linear relationship between the logarithm of the bias and the logarithm of the sample size, I have generated a comparison plot, shown in Figure 3.4 (where $k = 1$ is removed due to reasons stated earlier).

For the smaller values of N, when $N \approx 100$, we can see obviously from this plot that the lowest line occurs from $k = 3$, and this value of k stays the best for the estimator until $log(N) \approx 8.5$, so the sample size $N \approx 5,000$. Above this value, it appears to be that the smallest bias occurs from when $k \in \{6, 7, 8, 9, 11\}$, to see this more accurately consider an enlarged version of this graph for $5,000 < N < 50,000$, Figure 3.10.

This graph shows that for large $N \leq 50,000$, the lowest line on the graph is when

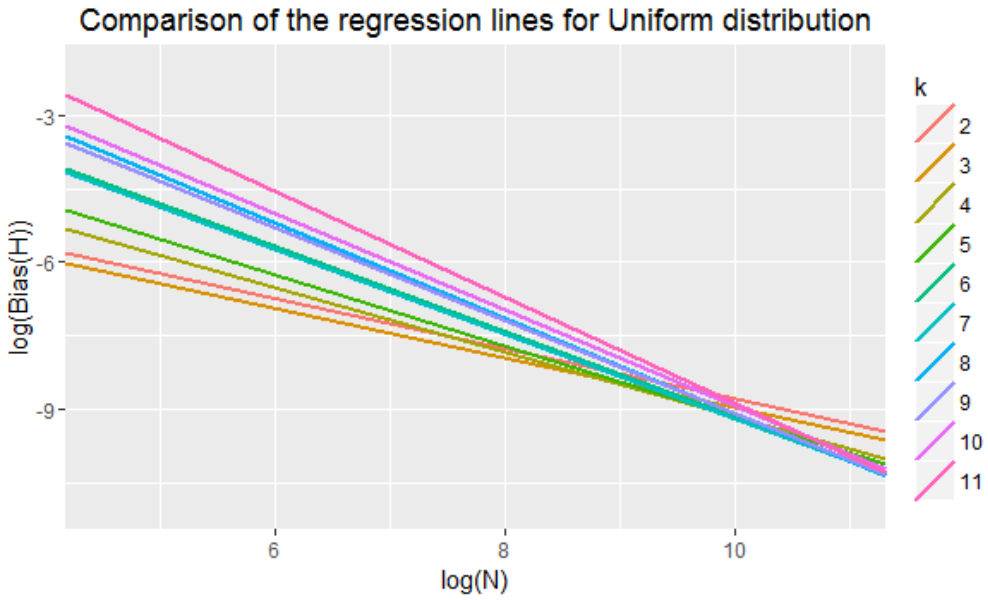


Figure 3.9: *Plot of regression lines for* $\log|Bias(\hat{H}_{N,k})|$ *against* $\log(N)$*, for* $k = 2, 3, ..., 11$*, for samples from the uniform distribution*

Figure 3.10: *Figure 3.9 zoomed in around large* N*.*

the estimator is found using $k = 7$; thus there is a possibility that this value of k could be the best nearest neighbour value to choose, when considering sample size $N \approx 50,000$ from the uniform distribution.

3.3 1-dimensional Exponential Distribution

Lastly, I consider 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 the definition of the Shannon entropy (1.1) and the density function for the exponential distribution, $f(x) = \lambda e^{-\lambda x}$ for $x \in [0, \infty)$ and $\lambda > 0$, we can write the exact entropy:

$$\begin{aligned}
H &= -\int_{x\in\mathbb{R}^d} f(x)log(f(x))dx \\
&= -\int_0^\infty \lambda e^{-\lambda x} log[\lambda e^{-\lambda x}]dx \\
&= -\lambda \int_0^\infty e^{-\lambda x}[log(\lambda) - \lambda x]dx \\
&= \lambda^2 \int_0^\infty x e^{-\lambda x} dx - \lambda \int_0^\infty log(\lambda) e^{-\lambda x} dx \\
&= \lambda \left[-xe^{-\lambda x}\right]_0^\infty + \lambda \int_0^\infty e^{-\lambda x} dx - \lambda log(\lambda) \int_0^\infty e^{-\lambda x} dx \\
&= 0 + (1 - log(\lambda)) \left[e^{-\lambda x}\right]_0^\infty \\
&= 1 - log(\lambda)
\end{aligned}$$

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

$$H = 1 - log(\lambda) \tag{3.7}$$

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 K-L estimator that I will be considering here (2.16).

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

$$H = 1 - log(0.5) \approx 1.693147 \tag{3.8}$$

3.3.1 Estimator conditions

Samples from the exponential distribution must satisfy the conditions of Theorem 1 and 2, to be an asymptotically unbiased and consistent estimator. For these theorems to hold, this distribution must satisfy the Conditions 1, 2 and 3.

Firstly, to satisfy Condition 1, the density function $f(x) = \frac{1}{2}e^{\frac{-x}{2}}$ 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 - this is obvious, since if we consider the mth derivative of the density function f we get:

$$\begin{aligned}\frac{d^m}{dx^m}(f(x)) &= \frac{d^m}{dx^m}(\lambda e^{-\lambda x}) \\ &= (-1)^m \lambda^{m+1} e^{\frac{-x}{\lambda}} \\ &= (-1)^m \lambda^m f(x)\end{aligned}$$

where $(-1)^m\lambda^m$ is a constant, thus a polynomial of order $O(1)$, which exists for all $x \in [0, \infty)$; 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, the norm $\|\cdot\|$ can be written as $|.|$. Moreover, considering that for Theorems 1 and 2, we have the value of $\beta \geq 2$; thus choosing $\beta = 2$, and since $m = \lfloor\beta\rfloor = \lfloor 2\rfloor = 2 = \beta$ and $\eta = \beta - m$, we have that $\eta = 0$, just as previously. 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 $g_*(x) = |(-1)^m\lambda^m| = |\lambda^m| = \lambda^m$, 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 $|f^{(m)}(x) - f^{(m)}(y)| \leq g_*(x)f(x)|y - x|^0 = g_*(x)f(x)$. We currently have:

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

  since we know that $f(x) > 0$ for all $x \in \mathbb{R}$, and $g_*(x) = \lambda^m > 0$, which is the g_* before; thus satisfying Condition 2.

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 1 and 2. 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. Thus, the values of k found in section 3.1.1 are $\{1, 2, ..., 11\}$.

Due to the above conditions for Theorems 1 and 2 being met, we can say that the K-L estimator, of a sample from the exponential distribution, is an asymptotically unbiased and consistent estimator for entropy.

3.3.2 Simulation results

I will be exploring the simulation results using the same process as for the previous two distributions, which begins with examining the values of the bias of the estimator at certain values of N for all different $k \in \{1, 2, ..., 11\}$.

Using $N = 100$, $N = 25,000$ and $N = 50,000$, from Appendix B, we can create a table to compare the mean values of the bias of the estimator for the different values of k considered. Recall that the information is taken over 500 sample of size N and the mean of these estimators is considered, then the bias is found by taking the modulus of this value minus the exact entropy (3.8).

For the sample size $N = 100$, Table 3.10 shows that the optimal value of k to choose for the estimator is obviously given by $k = 1$, since the value of the bias is $\approx 10^{-3}$ smaller than that for all the other values of k. This fits with the analysis previously conducted, that for $N = 100$, we must have $k \in \{1, 2\}$.

Considering $k = 1$ for larger sample sizes, we have `Inf` $= \infty$ - an error, similar to that for uniform distribution. This happens due to limitations in computer programs; since, we cannot have continuous data, it must be discretised, so for infinitely small distances between two points, we get a $-\infty$ value for the distance to the values closest neighbour, which only has to happen between two points for the whole estimator to become infinite. Because of this, I will not be examining the value of the estimator for $k = 1$, when considering large sample sizes.

Table 3.10: *1-dimensional exponential distribution, comparison of k*

k	$\lvert Bias(\hat{H}_{100,k})\rvert$	$\lvert Bias(\hat{H}_{25000,k})\rvert$	$\lvert Bias(\hat{H}_{50000,k})\rvert$
1	0.0008253	Inf	Inf
2	0.0210589	0.0008092	0.0000643
3	0.0173133	0.0001295	0.0004322
4	0.0146824	0.0000769	0.0000023
5	0.0155819	0.0000654	0.0000749
6	0.0194462	0.0000117	0.0001435
7	0.0127644	0.0005499	0.0000490
8	0.0174169	0.0000440	0.0004291
9	0.0216625	0.0003314	0.0000560
10	0.0177220	0.0004974	0.0000544
11	0.0162163	0.0000472	0.0006244

Comparing the values of $\lvert Bias(\hat{H}_{N,k})\rvert$ *for the values of* k *with* $N = 100$, $N = 25,000$ *and* $N = 50,000$, *when the estimator is taken over* 500 *samples*

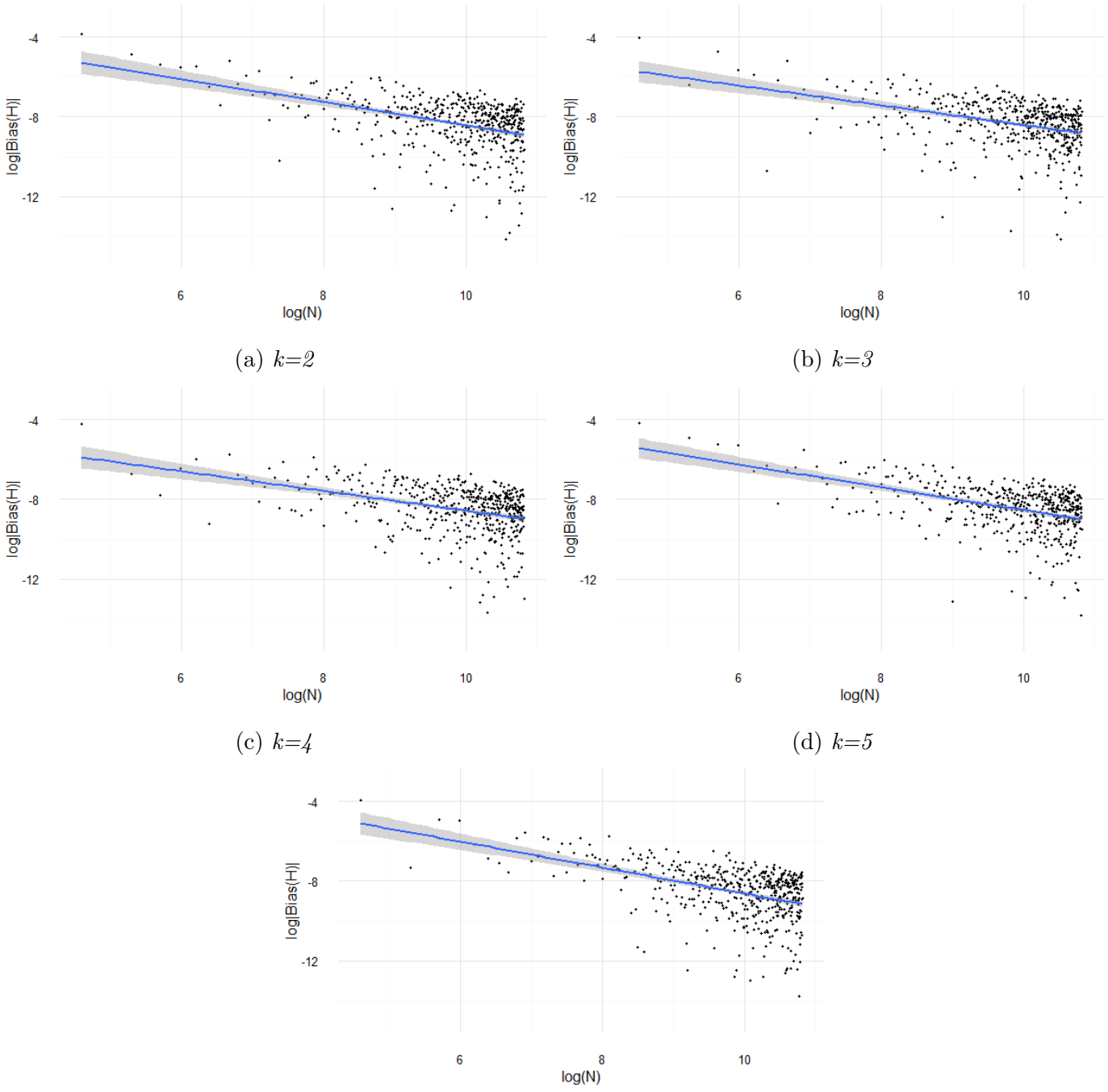
Looking at the magnitude of the bias for samples of size $N = 25,000$, when $k \in \{2, 3, ..., 11\}$, we can see that the smallest bias occurs at $k = 6$; but it is also quite small for $k = 4, 5$ and 11. Because of the closeness of the size of the bias for all the values of k just mentioned, we cannot currently draw up any conclusions about which value of k is better to use - in terms of reducing the bias.

Considering the information found from samples of size $N = 50,000$, it appears to show that the estimator for $k = 4$ has a significantly smaller bias than that for the other values of k. The next smallest bias is $\approx 10^1$ larger than that for $k = 4$, and occurs when k is either $2, 5, 7, 9$ or 10. Since we are only looking at the bias at this specific sample size, and not looking at the information about the value of the bias found for samples of a similar size, we cannot yet draw any conclusive decisions from this table.

The relationship between the logarithm of the bias and N (3.2), can be depicted in a plot of the simulated data to see if analytically the estimator does, in fact, show the linear relationship proposed. I have plotted the data separately for each value of k, and have fitted a linear regression line to each plot to indicate the general trend. I have not considered $k = 1$ since a large proportion of the simulated values are infinite; thus, the graph does not make much sense. These plots are shown in Figures 3.11 and 3.12.

These graphs all show what is expected, and what has been confirmed with the previous two distributions; the logarithm of the sample size against the logarithm of the bias of the estimator does indeed show a negative linear relationship. In these graphs there seems to be little difference between the values of k in how close the data actually fits the line. This motivates us to consider both the coefficient of determination and the standard deviations of the regression for each value of k, which is given in Table 3.11.

The R^2 value (coefficient of determination) is ≈ 0.207, with very little variation, for all values of k. This is an indication that the regression line could be a poor fit to the data, since a value of 1 means a perfect fit. However, the standard error of the regression line is small, ≈ 1.15 for all k, again with little variation. The standard error for the lines here is similar to that for both the uniform and normal distributions;



(a) *k=2*

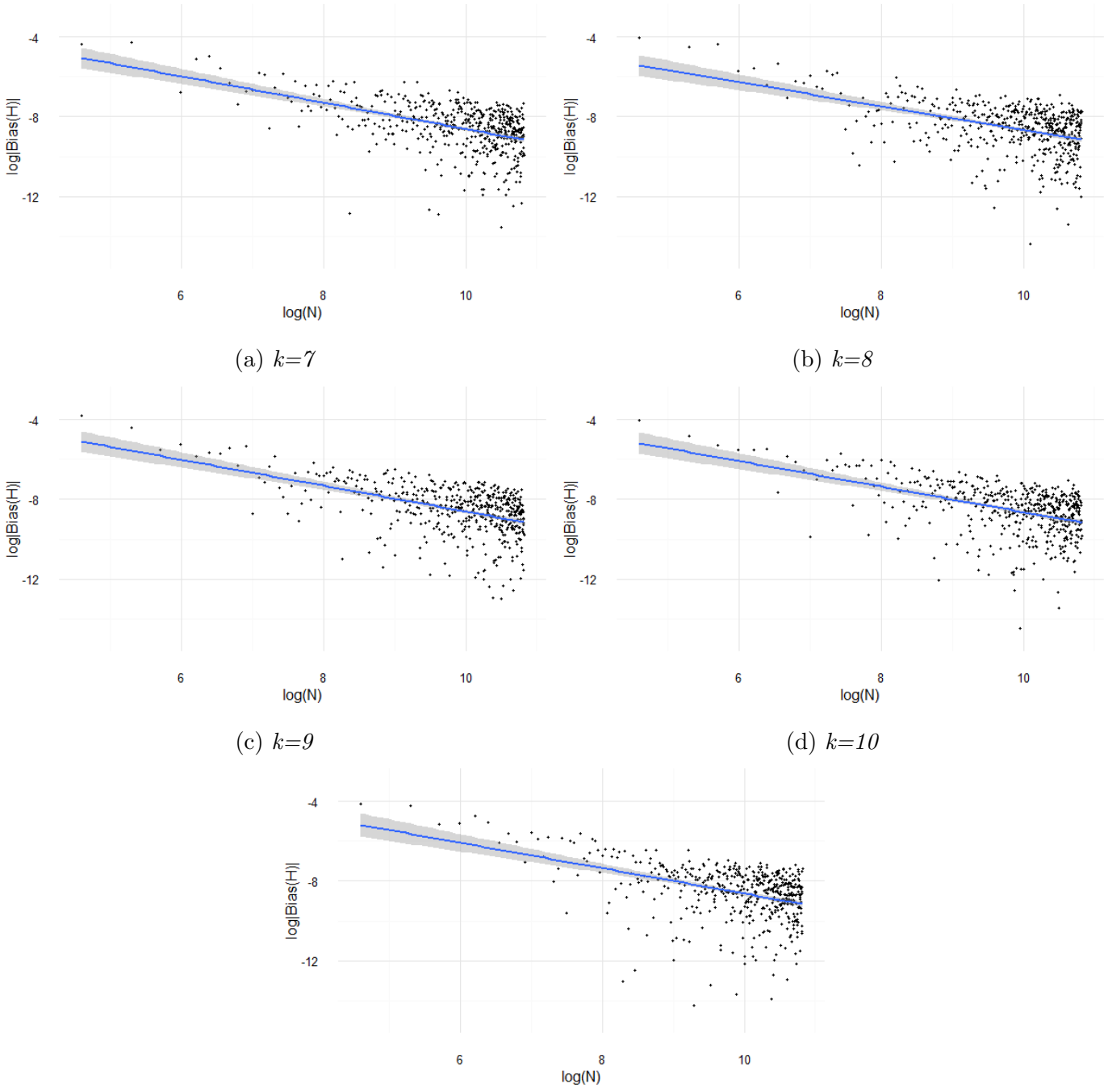
(b) *k=3*

(c) *k=4*

(d) *k=5*

(e) *k=6*

Figure 3.11: *1-dimensional Exponential distribution with different* $k = 2, ..., 6$



(a) *k=7*

(b) *k=8*

(c) *k=9*

(d) *k=10*

(e) *k=11*

Figure 3.12: *1-dimensional Exponential distribution with different* $k = 7, ..., 11$

Table 3.11: *Comparison of the coefficient of determination and the standard deviations of the regression for each value of k for the 1-dimensional exponential distribution*

k	R^2	σ
2	0.1843	1.1972
3	0.1596	1.1073
4	0.1538	1.1321
5	0.2151	1.0687
6	0.2277	1.1694
7	0.2474	1.1253
8	0.2107	1.1473
9	0.2509	1.0939
10	0.2200	1.2150
11	0.2017	1.2369

Table 3.12: *Comparison of coefficients of regression a_k and c_k from equation 3.1, for 1-dimensional exponential distribution*

k	a_k	c_k
2	0.5824	0.0739
3	0.4941	0.0310
4	0.4940	0.0266
5	0.5727	0.0602
6	0.6500	0.1199
7	0.6605	0.1324
8	0.6067	0.0739
9	0.6480	0.1189
10	0.6606	0.1267
11	0.6365	0.1042

thus, this doesn't necessarily indicate an extremely poorly fitting line. Conversely to the relationship shown for the other two distributions, here there is no obvious trend as k runs from $2 \to 11$, when previously there was an clear increase in R^2 and decrease in σ.

The most important information found from these graphs are the equation of the regression lines. I have collated all the data about the coefficients a_k and c_k, of these lines, and displayed them in Table 3.12.

This distribution shows interesting results, that are somewhat different to those previously shown in the other two distributions. In a similar fashion to previously there is a general increase in a_k with k; however, in comparison to before where c_k acted similarly, here c_k does not increase uniformly with k. All the values of a_k and c_k are unvaried, with a_k ranging from 0.4940 to 0.6606 and c_k ranging from 0.0266 to 0.1324. Interestingly, the smallest values of a_k and c_k, occur at $k = 4$ then at $k = 3$. Additionally, the largest two values occur at $k = 7$ and $k = 10$. The optimal value of

Table 3.13: *Considering the dependence of k on c_k*

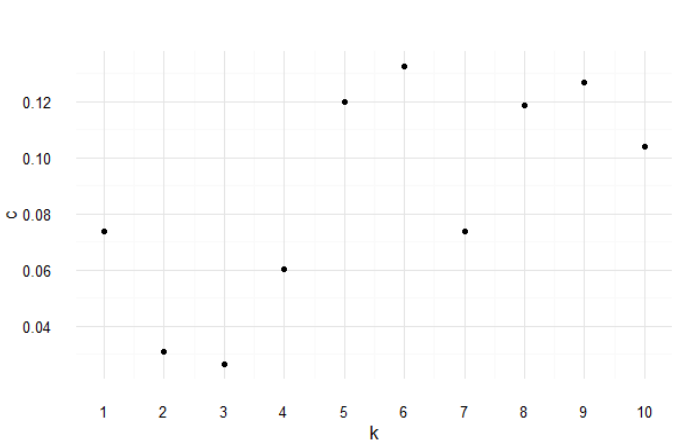
k	k^{a_k}	c_k	$\frac{k^{a_k}}{c_k}$
2	1.0000	0.0739	13.532
3	1.4084	0.0310	45.434
4	1.7207	0.0266	64.687
5	2.2121	0.0602	36.745
6	2.8466	0.1199	23.742
7	3.2656	0.1324	24.665
8	3.2563	0.0739	44.063
9	3.8477	0.1189	32.361
10	4.2695	0.1267	33.697
11	4.3301	0.1042	41.556

k will hopefully become more apparent when plotting all the regression lines against one and other.

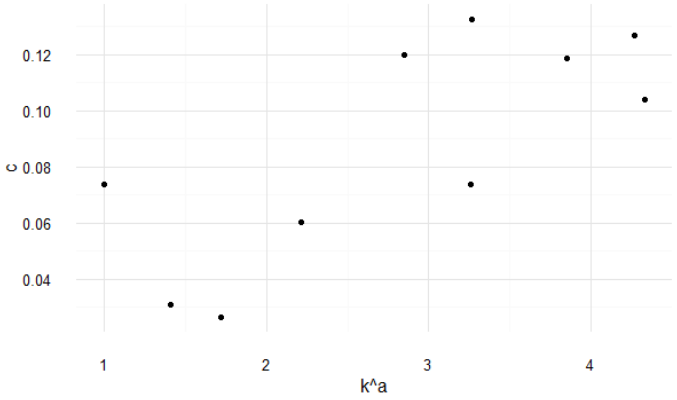
Firstly, I wish to look at the relationship between k and c_k in the exponential distribution. To do this, I will look at the values of c_k for each k and see if it depends on k^{a_k} or not. This should help us to decipher if the bias is of $O\left(\frac{1}{N^a}\right)$ or $O\left(\left(\frac{k}{N}\right)^a\right)$. This information is shown in table 3.13.

For the previous distributions there has been an obvious increase when looking at the values of $\frac{k^{a_k}}{c_k}$ for increasing k. However, for the exponential distribution this is not the case, the values seem to be all over the place, and plotting these in Figure 3.13 confirms this.

These graphs do not show the almost exponential or parabolic relationship seen before, they don't appear to show any strong relationship between k and c_k. This could imply that for the exponential distribution we have $|Bias(\hat{H}_{N,k})| = O\left(\frac{1}{N^a}\right)$, in comparison the normal and uniform which showed results in favour of the opposing idea; $|Bias(\hat{H}_{N,k})| = O\left(\left(\frac{k}{N}\right)^a\right)$.



(a) *The values of k against the values of c_k*



(b) *The values of k^{a_k} against the corresponding values of c_k*

Figure 3.13: *Graphically representing the relationship between c_k and k for the exponential distribution*

To find the optimal value of k for this distribution, I have generated a comparison plot of all of the regression lines, shown in Figure 3.14. Hopefully this can shed some light on which value of k appears to be the best for use in the estimator, depending on the sample size, N.

At smaller sample size $N \leq 5,000$, the graph shows that the estimator with $k = 4$ has the lowest line; hence, the smallest bias and the best one to use for the estimation. However, for larger N this is not true, to better visualise what is happening for large N, I have created an enlarged version in Figure 3.15.

This graph shows that for large sample sizes, $5,000 \leq N \leq 50,000$ that the estimator with $k = 10$ has the smallest bias; thus appears to be the best one to use for estimating entropy from this distribution.

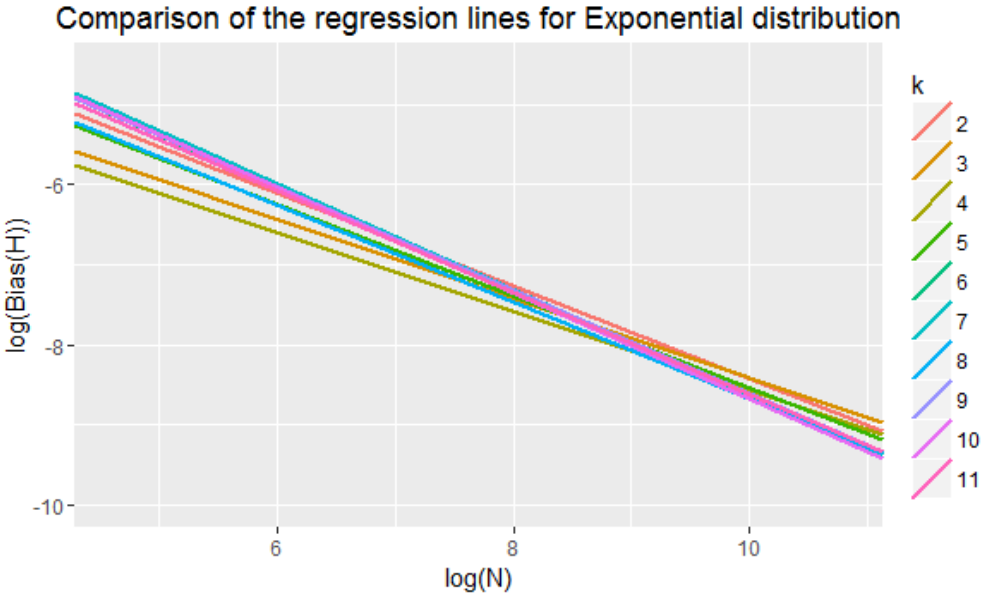


Figure 3.14: *Plot of regression lines for* $\log|Bias(\hat{H}_{N,k})|$ *against* $\log(N)$*, for* $k = 2, 3, ..., 11$*, for samples from the exponential distribution*

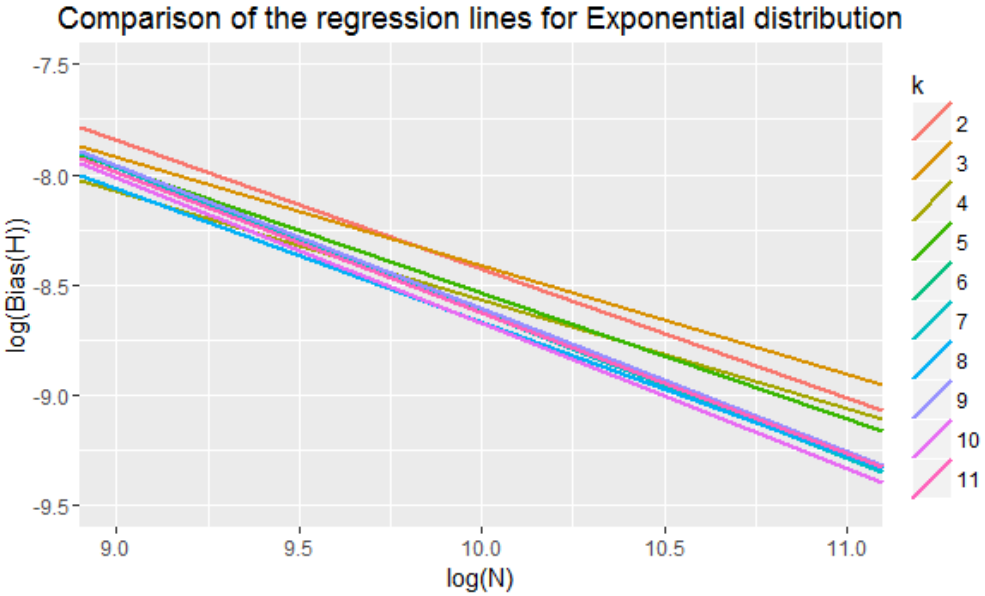


Figure 3.15: *Figure 3.14 zoomed in around large* N*.*

Chapter 4

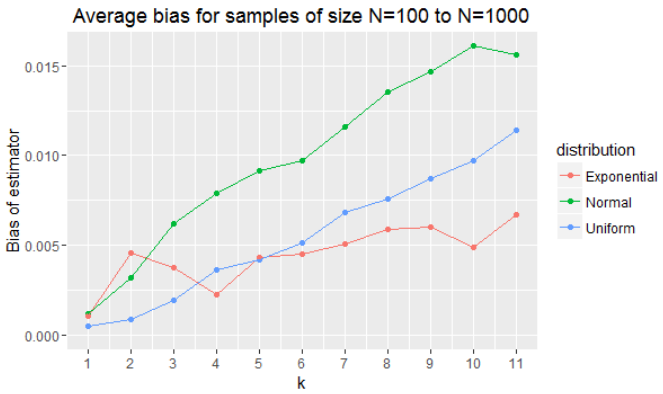
Conclusion

As seen from the analysis undertaken in Chapter 3, taking an estimation of entropy from samples of the normal, uniform and exponential distributions, appear to act differently. There is no conclusive answer, that agrees across all three distributions, as for the best value of k to use for the estimator, or if the bias of the estimator is of $O\left(\frac{1}{N^a}\right)$ or $O\left(\left(\frac{k}{N}\right)^a\right)$.

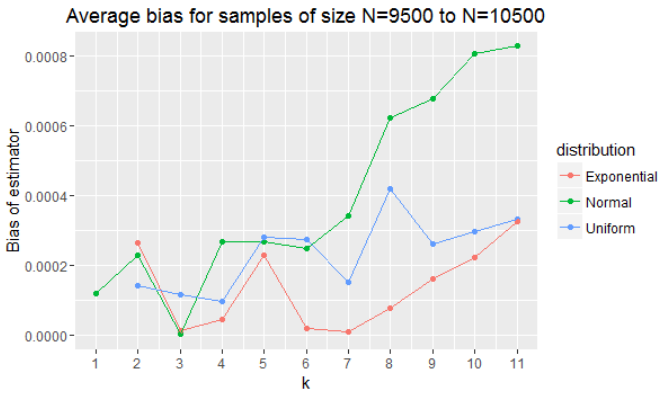
I previously discussed the supposed best value of k as shown by the graphs with regression lines of the bias for each distribution. Here, optimal is meant in terms of reducing the bias, so the estimated value is as close to the exact value as possible. To view this in more detail, comparing distributions and sample sizes, I have taken 1000 samples around the sample sizes of 500, 10000, 20000, 30000, 40000 and 49500, found the average bias of the estimator for each k and each distribution. I have graphically represented this in Figure 4.1.

If we ignore the distribution, there is not an always obvious best value of k shown by the graphs. Moreover, considering the analysis completed in section 3.1.1, we found that the values of k that satisfy Theorems 1 and 2, depend on N so that $k \in \{k_0^*, ..., k_1^*\}$, where $k_0^* \approx 1$ and $k_1^* = O(N^{\frac{2}{9}})$. Thus for the samples in the graphs above, we can create a table of the values of k that satisfy the conditions to be able to assume consistency and unbias of the estimator, Table 4.1.

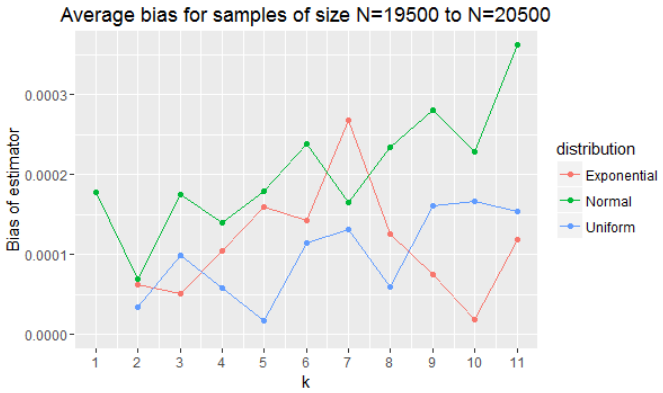
Although, for samples of size $N \leq 1,000$, we can see that, quite obviously, the smallest bias occurs with an estimator found using $k = 1$, and this is true for all the



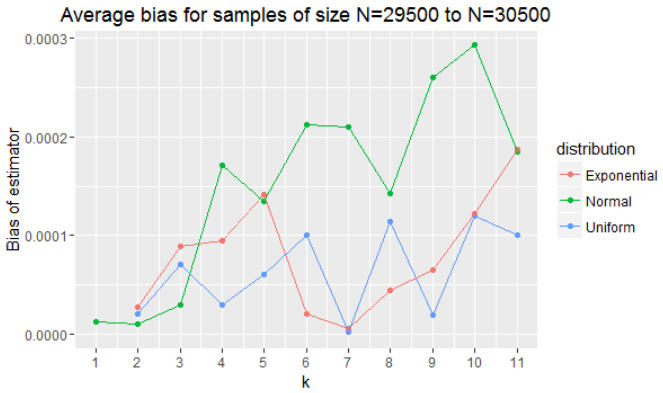
(a) $N \approx 500$



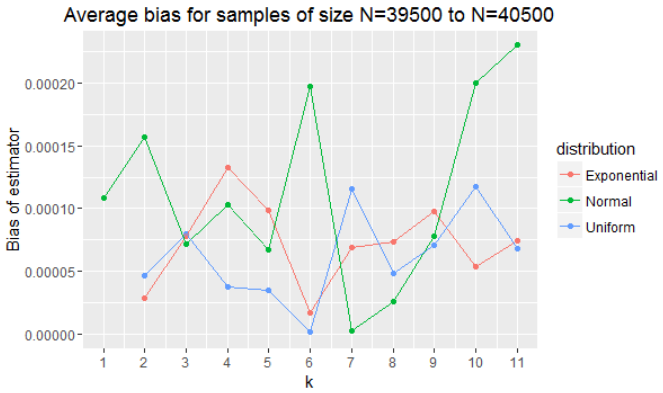
(b) $N \approx 10,000$



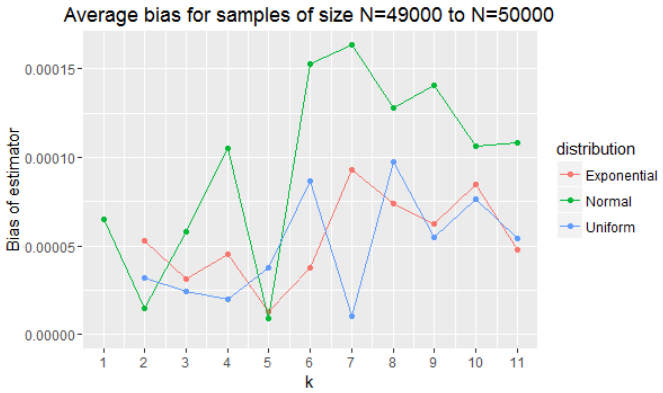
(c) $N \approx 20,000$



(d) $N \approx 30,000$



(e) $N \approx 40,000$



(f) $N \approx 49,500$

Figure 4.1: *Comparing the average bias about sample size N, for each k and each distribution.*

Table 4.1: *Values of k satisfying Theorems 2 and 1 for sample sizes N*

N	k
500	$\{1, 2, 3\}$
$1,000$	$\{1, 2, 3, 4\}$
$10,000$	$\{1, 2, ..., 7\}$
$20,000$	$\{1, 2, ..., 9\}$
$30,000$	$\{1, 2, ..., 9\}$
$40,000$	$\{1, 2, ..., 10\}$
$50,000$	$\{1, 2, ..., 11\}$

distributions, graph (a). If we consider the comparison graphs in Chapter 3, when $N \leq 1,000$ we approximately have $log(N) \leq 7$, and in Figures 3.4, 3.9 and 3.14 we have the smallest bias occurring at $k = 2$, $k = 3$ and $k = 4$ for the normal, uniform and exponential samples respectively. This difference in results could be due to a number of reasons. Firstly, from the current analysis $k = 1$ appears to be the optimal value for the smallest bias; however, when looking at the uniform and exponential samples, the regression line for $k = 1$ was omitted, due to the large number of `Inf` values (reasoning for these values were discussed previously). Secondly, the graphs shown in Chapter 3 are of regression lines, which we have assumed to be linear as $N \to \infty$. Thus, when looking at the information shown by the lines in a particular region where N is small, we cannot say that these are accurate. Therefore, the difference in the results is not a problem; and the graph in Figure 4.1 (a), is more accurate when looking at this sample size, N. Moreover, from this graph, we can see a uniform increase in the bias, against the size of k for both the uniform and normal distributions; thus for $N \approx 500$ and $k \in \{1, 2, 3\}$, analytically the optimal estimator appears to be found using $k = 1$.

Considering graph (b) for sample size $N \approx 10,000$, we have that the smallest values of bias occur when $k = 3$ or 7 for the exponential, $k = 3$ for the normal, and

$k = 4$ for the uniform distributions. This agrees with the information found for k to satisfy Condition 3: for this sample size $k \in \{1, 2, ..., 7\}$. This also implies that the most accurate estimator for a sample of this size could be found when using $k = 3$ (or possibly 4), since for all three distributions, at these values of k, we do have some of the smallest sizes of bias. When $N \approx 10,000$ we have $log(N) \approx 9.2$, which is where on Figures 3.4, 3.9 and 3.14, the regression lines begin to cross. Thus, looking at these figures makes it difficult to decipher which value of k appears to have the smallest bias. Whence, we can look at Figures 3.5, 3.10 and 3.15, which zoom in around $log(N) \approx 9$, we can see that for the normal distribution, the lowest lines occur at $k = 2$ then 3, for the uniform $k = 4$ then 3 and for the exponential $k = 8$ then 3. The results in these graphs, as well as the graphs in this chapter, all give the impression that $k = 3$ is a strong choice for the optimal value of k across all distributions, when finding the estimator for sample size $N \approx 10,000$.

For graph (c), with $N \approx 20,000$, we have a range of values for k with the smallest bias, depending on the distribution; for the exponential $k = 10$, the normal $k = 2$, and the uniform $k = 5$. However, considering that for this sample size we must have $k \in \{1, 2, ..., 9\}$, the value of k for the exponential distribution cannot be $k = 10$ for Theorems 1 and 2 to be satisfied. So for the exponential distribution, the next smallest bias occurs when $k = 3$. Due to this, nothing conclusive can be said about the optimal value of k, although, we could say that for this sample size, the best k is from the range $k = \{2, 3, 5\}$. However, when choosing the values of k that satisfy Condition 3, approximations and assumptions were made, which depended on the choice of $\alpha > d = 1$, and a value τ which is smaller than a bound $(\frac{1}{4})$. Due to this, the value of τ could be chosen to be slightly larger; for example, $\tau = \frac{19}{80} < \frac{1}{4}$, which would give $k_1^* = N^{\frac{19}{80}} = 20000^{\frac{19}{80}} = 10.507$. So $k = 10$ could be included in the values that satisfy the Theorems and cannot be discounted. In fact, if we did use this new value of τ, then the values for the optimum k would be $k = \{2, 5, 10\}$, which makes even less sense.

Moreover, we can also examine the graphs in Chapter 3, for this sample size $N \approx 20,000$ which gives $log(N) \approx 9.9$. Thus, Figures 3.5, 3.10 and 3.15, which zoom around this sample size, show that the optimal choice of k for the normal distribution

is $k = 5$, for the uniform is $k = 7$ then 5 and for the exponential $k = 8$ or 10. The exponential agrees on $k = 10$ and the uniform agrees on $k = 5$, but these are both the second lowest lines on the graph, and other than this none of the above results agree across distributions with those from earlier. This could be due to a number of reasons, but the most obvious one being that there is not a fixed value of k for this sample size that coincides for all distributions as the optimal value of k. Thus, for this sample size we cannot draw any conclusions about the optimal value of k.

When $N \approx 30,000$, graph (d), we have the optimal values of k - when the smallest bias occurs - being 2 for the normal distribution and 7 for both the uniform and exponential distributions. Interestingly, for all three distributions $k = 2$ has either the first, second or third smallest bias, out of all k for its own distribution. However, while the normal distribution has a general increase for $k > 2$, the other two distributions later dip to their lowest value of bias at $k = 7$. Thus, while one may say that $k = 7$ is the optimal, since it agrees with two of the three distributions, for the third distribution, the normal distribution, the bias is significantly higher at this value of k. Henceforth, I would actually assume that a safer option, for the best value of k across any distribution would be when $k = 2$, where the bias for all three distributions is significantly small. Considering the regression lines, found in Chapter 3, for $log(N) \approx 10.3$, we can see that the optimal choice of k appears to be $k = 5$ for the normal distribution, $k = 8$ for the uniform and $k = 10$ for the exponential. None of these agree with the possibility the $k = 2$ is the optimal value, in fact when considering the graphs of the comparison of the regression lines for each distribution, we can see that the line for $k = 2$ is not nearly the lowest of the group. More specifically for both the exponential and uniform distributions $k = 2$ appears to be the worst possible value of k in decreasing the bias, and since we are now considering a relatively large sample size, one would expect these graphs to have more accurate results. Although all these values of k discussed fit within those found in Table 4.1; no value of k stands out as an optimal value, thus no conclusion can be drawn for this sample size.

Now, for the graph (e), when $N \approx 40,000$ we have the smallest bias occurring at $k = 6$ for the uniform and exponential and $k = 7$ for the normal. If we consider the

value for the bias at $k = 6$ for the normal, we can see a sudden significant increase in the size of the bias and a massive drop between this value, and that for $k = 7$. Moreover, considering $k = 7$ for the uniform and exponential distributions, we can see a substantial increase from the bias at $k = 6$. Indicating, that the best value of k may be 6 or 7; but this strongly depends on the distribution that the sample is taken from, not just the sample size. Now looking at the regression comparison graphs, around $N \approx 40,000$ we have $log(N) \approx 10.6$, which shows for the normal distribution the optimal value of $k = 5$ and 7, for the uniform $k = 7$ then $6, 8, 9$ (very close regression lines), and for the exponential $k = 10$ then $6, 7, 8$ (again very close regression lines). This slightly agrees with above, that $k = 6$ or 7 could be the best value of k to choose for the estimator; moreover, it also shows less of a dependence on distribution. Thus, I would suggest that for a sample size $N \approx 40,000$, that $k = \{6, 7\}$ would be the best candidates for the values of k, across all three distributions.

Lastly, for the final graph (f) with sample size $N \approx 50,000$, we can see that instead of the uniform and exponential agreeing as before, we now have the normal and exponential agreeing on its minimum value of bias at $k = 5$, and the uniform minimum occurring at $k = 7$. For the uniform distribution, the estimator with $k = 5$ still holds a small bias, but there are four other values of k with smaller bias for this distribution. For the normal distribution, there is a large jump at $k = 7$, where the bias is much larger than before, and the exponential also shows a substantial jump. For this sample size, we have $log(N) \approx 10.8$, so considering the enlarged regression graphs; Figures 3.5, 3.10 and 3.15, we can see that for each distribution the best value of k in reducing the bias is given by $k = 7$ then 5 for the normal distribution, $k = 7$ or 8 for the uniform and $k = 10$ for the exponential. Only the normal distribution aligns with above in that $k = 7$ is the optimum value, but the rest does no coincide. However, both $k = 5$ and $k = 7$ appear in both analysis, a number of times so would be inclined to possibly agree that $k = \{5, 7\}$ are the optimal choices for this sample size $N \approx 50,000$, but this result is not certain for all distributions.

By considering all the information found from the graphs in Figure 4.1, for small sample sizes $N \leq 1,000$ we can say that $k = 1$ would possibly be the best choice for k. However, I don't think anything conclusive can be drawn about a general distribution

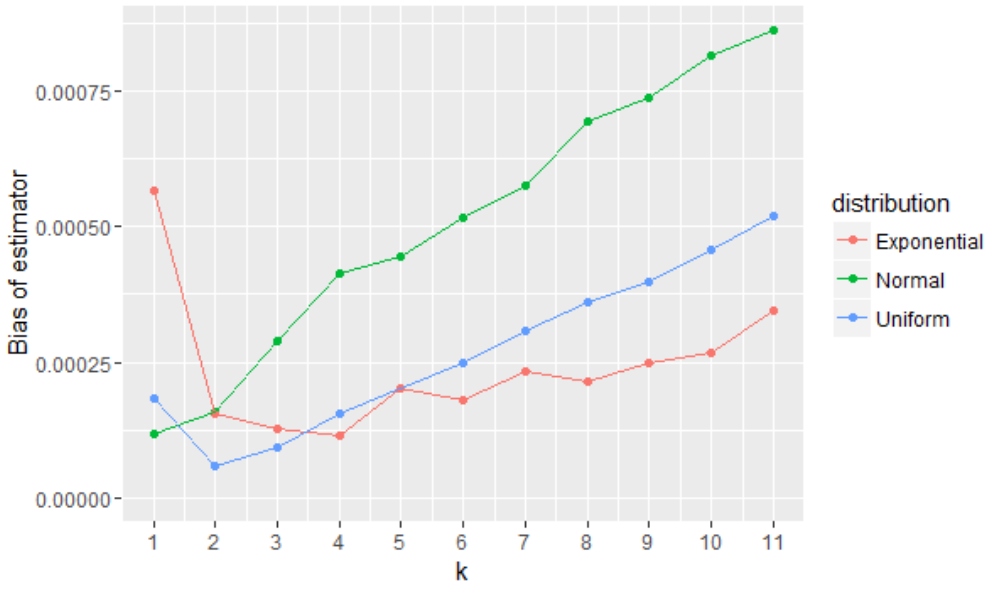


Figure 4.2: *The average bias over all sample sizes N for each k and each distribution.*

from a larger sample size N, despite some of the rough deductions above, there is no definitive choice for k being obviously shown throughout each distribution considered.

Perhaps an interesting graph to look at is that depicted in Figure 4.2, where I have taken all the data for every sample size $100 \leq N \leq 50,000$ and averaged the bias per distribution. I have done this to see if there is a safe value of k to choose when wanting the estimator of entropy from a random sample to have the smallest bias.

From this, we can see that for a sample distributed like the normal, the safest value of k to choose for any sample size is $k = 1$, since on average, this gives the smallest bias of ≈ 0.000125, regardless of its sample size. Then for all other values of k, the bias of the estimator increases with k; hence, if in doubt, I would recommend choosing $k = 1$ if you suspect your 1-dimensional data to be normally distributed.

Considering the uniform distribution, we can see that the smallest average bias occurs at $k = 2$, with the same behaviour as the normal for $k \geq 3$, whereby the bias increases with k. The average bias at $k = 2$ is ≈ 0.0000625, which is about half the size

Table 4.2: *Suspected behaviour of Bias*$|\hat{H}_{N,k}|$ *for each distribution*

Normal, $N(0,1)$	Uniform, $U[0,100]$	Exponential, $exp(0.5)$
$O\left(\left(\frac{k}{N}\right)^a\right)$	$O\left(\left(\frac{k}{N}\right)^a\right)$	$O\left(\frac{1}{N^a}\right)$

of that for the normal distribution. However, even though the relative size of the bias within a distribution is important to look at, the size between distributions is not as informative, due to the variability of the samples. Where for the normal distribution we have variance $\sigma^2 = 1$ and for the uniform we have variance $\frac{12}{100^2} = 0.0012$, which is significantly smaller - so we would expect a more accurate estimator due to this. Thus, when uncertain of which k to choose, if the 1-dimensional data appears to be uniformly distributed, then it seems to be that $k = 2$ would be the safest choice.

Lastly, looking at the most interesting result of the exponential distribution, which shows that, on average, the smallest bias occurs when $k = 4$, with the bias generally increasing on both sides as k increases/decreases away from 4. The bias here is $\approx$ 0.000125 - similar to that of the normal distribution, which is surprising owing to the fact that the samples from the exponential distribution were taken with a variability of $\frac{1}{0.5^2} = 4$, which is four times larger than the variance of samples from the normal distribution. This is something very interesting about the exponential distribution, and I have noticed throughout my analysis that samples from this distribution do indeed act quite differently to those from the other two.

Another important aim of this paper was to consider whether the distributions imply the bias is of equations 2.20 or 2.21; i.e, whether bias is of $O\left(\frac{1}{N^a}\right)$ or $O\left(\left(\frac{k}{N}\right)^a\right)$. From the previous section, namely Figures 3.3, 3.8 and 3.13, implied the behaviours for the distributions considered shown in Table 4.2.

This table shows that for samples from the normal and uniform distributions, both seemed to have the bias of the estimator showing the relation where c_k depends on k^a, since the graphs of this correspondence, appeared to show an almost exponential or parabolic relationship between the two variables. However, for the exponential

distribution a very different picture occurred; thus, implying for this distribution that there was no obvious dependence on k^a, and in fact c_k is just constant. This was confirmed by the values shown in Table 3.12, where there was very little variability for each k.

Overall, this paper has shown that analytically, the estimator is asymptotically unbiased and that we cannot know how to choose k for a sample by only considering its size N. We can only can find a range of k that satisfies the conditions from the theoretical works, but the analytical results show that there are many discrepancies as to which value of k is optimal, depending on how the sample is distributed. We have also not been able to draw any solid conclusions about the form of the bias of the estimator; however, the data here only showed implications of the connection between c_k and k^a, it has not proved or disproved it either way.

There are large amounts of further research to be done into the estimation of entropy, but time permits that this paper can look at nothing more. With more time, it would be interesting to see how the estimator works analytically on samples from different distributions; for example, Gamma, Chi-Squared or Beta distributions. Also, it would be intriguing to look analytically at the estimator for samples from higher dimensions, with known entropy.

Appendix A

Code

For the simulations in this project, I used R: "a language and environment for statistical computing and graphics" [1]. I created a package called `EntropyEst`, with the functions needed to created the K-L entropy estimator (KLEE). The functions exported form this package are; `KLEE`, `GammaFun`, `Rho`, `VolD`, `NormalEnt`, `UniformEnt` and `ExpoEnt`. The functions in this package I then used to run simulations on samples from different statistical distributions to create the results in this paper.

To create my package `Entropy-Estimators`, I used two of Hadley Wickham's [34] packages; `devtools` and `roxygen2`. `Entropy-Estimators` also has a dependent package (alongside the base R packages); `FNN`, which was used for the kth nearest neighbour function.

When running simulations and creating the graphical information in this paper, I also used 5 more packages: `ggplot2` for the graphical representation of the data, `Rcpp` to create a C++ for loop for faster computation, `dplyr` and `tidyr` for the manipulation of data, and `readr` to read and write the CSV files of data.

I will outline the important code used for the simulations in this appendix; however, the full package and a complete account of the code used can be found on my GitHub page *https://github.com/KarinaMarks/Entropy-Estimators*.

A.1 The K-L estimator

Using the definition of the K-L estimator, found in Section 2.2, I created the function `KLEE`. To do this I first had to make the following functions; `GammaFun`, `Rho`, `VolD`. For the purpose of this paper, I have only created the function to work out the estimator for a 1-dimensional sample.

Firstly, I created the function `Rho`, defined in equation 2.11, which takes three arguments:

- `X`: a random vector
- `k`: the order or nearest neighbour to be used
- `d`: the dimension of the sample, default value 1 - since currently can only work for 1-dimensional samples

It returns a vector of distances, where its first entry is the distance from the first value of `X` to its kth nearest neighbour. The function is given by the following code:

```
Rho <- function(X, k, d=1) {
  if (d == 1){
    # find the length of the sample
    n <- length(X)
    # check that k is not larger than the length of the
       vector
    stopifnot(n > k)

    # creating the matrix of kth nn distances for X
    NNdist <- FNN::knn.dist(data=X, k=k)

    # return the kth column of the matrix
    NNdist[,k]
  } else {
    return("Dimension is too high for this estimator")
```

```
  }
}
```

Next, I created the function `GammaFun`, for Γ defined in equation 2.13, which takes one argument `m` and returns a numeric value, using the following code:

```
GammaFun <- function(m) {
  # check that m > 0
  stopifnot(m > 0)
  # writing the function for the integrand
  integrand <- function(x) {x^(m-1)*exp(-x)}
  # integrating the integrand from 0 to infinity
  res <- integrate(integrand, 0, Inf)
  # selecting the result of the integral from the integrate
     class, which is a list structure
  as.numeric(res$value)
}
```

Then, using the function above for Γ, we can define the function `VolD`, equation 2.12, which is the volume of the d-dimensional unit euclidean ball. This function takes one argument `d` which is the dimension and returns a numeric value using the following code:

```
VolD <- function(d) {
  # the formula to find the d-dimensional euclidean unit ball
  (pi^(d/2))/GammaFun(1 + (d/2))
}
```

The only other function needed for the estimator is the digamma function which is defined in base R; thus, we can now define the estimator `KLEE`. This takes three arguments:

- `X`: a vector of a sample to estimate the entropy of
- `k`: the order or nearest neighbour to be used

- `d`: the dimension of the sample, default value 1 as before for `Rho`

This returns a numeric value which is the estimator of entropy for this sample. The code written to define this function is as follows:

```
KLEE <- function(X, k, d=1) {
  if (d==1){
    # length of the sample
    n <- length(X)
    # check that k is smaller than the length of the sample
    stopifnot(k < n)
    # define the vector Roe of nearest neighbour distances
    NN <- Rho(X, k)
    # find the volume of the unit ball
    V1 <- VolD(1)
    # return the estimator
    (1/n)*sum(log((NN*V1*(n-1))/exp(digamma(k))))
  } else {
    # this would be changed to include higher dimensions
    return("Dimension must be 1")
  }
}
```

A.2 Exact entropies

To consider the bias of this estimator, I had to find the exact value of entropy from a 1-dimensional normal, uniform and exponential distribution. The function written to return this for the normal distribution is `NormalEnt` with parameter `sd`, the standard deviation of the sample. The mean is not needed as a parameter since its value is not used for finding the entropy of the normal distribution. The function is defined, using equation 3.3, and the code is as follows:

```
NormalEnt <- function(sd){
  ((log(sqrt(2*pi*exp(1))*sd))
}
```

With `sd = 1`, as is true in the samples considered here, we find the exact entropy (3.4) to be:

```
> NormalEnt(sd=1)
[1] 1.418939
```

The function for the uniform distribution is `UniformEnt`, with parameters `min` and `max`, is defined using equation 3.5, and the code is written as:

```
UniformEnt <- function(min, max){
  log(max - min)
}
```

Here we use `min=0` and `max=100` in the samples considered; thus, we find the exact entropy (3.6) to be given by:

```
> UniformEnt(min = 0, max = 100)
[1] 4.60517
```

Lastly, for the exponential distribution we have the function `ExpoEnt`, with only one parameter `rate`, using equation 3.7, defined in code below:

```
ExpoEnt <- function(rate){
  1 - log(rate)
}
```

In this paper we are using the exponential distribution with parameter `rate=0.5`, thus the exact entropy here (3.8) is given by:

```
> ExpoEnt(rate = 0.5)
[1] 1.693147
```

A.3 Simulations

In this section I used the packages `readr` *to save the data,* `dplyr` *for the manipulation of data and* `Rcpp` *for creating a fast loop over hundreds of iterations.*

I created functions `normalloop`, `uniformloop` and `expoloop`, in C++ which, for each sample size N creates M samples of that size, finds the estimator for sample and puts the result in a vector of length M. I have used C++ here owing to the fact that R is slow in the computation of loops, and using `Rcpp` to compile a C++ loop, within an R script, dramatically increases computation time. These functions are as follows:

```
cppFunction('
            NumericVector normalloop(int M, int N, int k){
              NumericVector est(M);
              NumericVector x(N);
              for (int i = 0; i < M; i++) {
              int sd=1;
              Function KLEE("KLEE");
              Function rnorm("rnorm");
              x=rnorm(N, sd=sd);
              est[i]=as<double>(KLEE(x ,k=k));
              }
              return Rcpp::wrap(est);
              }
              ')
```

for the normal distribution, and for the uniform distribution:

```
cppFunction('
            NumericVector uniformloop(int M, int N, int k, int
              min, int max){
              NumericVector est(M);
              NumericVector x(N);
```

```
for (int i = 0; i < M; i++) {
Function KLEE("KLEE");
Function runif("runif");
x=runif(N, min=min, max=max);
est[i]=as<double>(KLEE(x ,k=k));
}
return Rcpp::wrap(est);
}
')
```

Lastly for the exponential distribution:

```
cppFunction('
        NumericVector expoloop(int M, int N, int k, float
          rate){
          NumericVector est(M);
          NumericVector x(N);
          for (int i = 0; i < M; i++) {
          Function KLEE("KLEE");
          Function rexp("rexp");
          x=rexp(N, rate=rate);
          est[i]=as<double>(KLEE(x ,k=k));
          }
          return Rcpp::wrap(est);
          }
          ')
```

Using these functions I created each column of the tables, where each table is a different distribution, each column is a different value of $k \in \{1, 2, ..., 11\}$ and each row is a different sample size $N \in \{100, 200, 300, ..., 50000\}$. Below is how the column with $k = 1$ for the normal distribution was created for all sample sizes N - all other columns were done similarly:

```
# initalise the data frame with all sample sizes n
```

```
data.frame(n = seq(100, 50000, 100)) %>%
  # group by n to use summarise on each n
  dplyr::group_by(n) %>%
  # for each n the mean of the normalloop function is found,
      taken over 500 samples of size n
  summarise(Ent = mean(normalloop(M=500, N=n, k=1, rate=0.5),
      na.rm=TRUE))
```

This data created is too large to fit in the appendices of this paper, details of how to view it are in Appendix B.

A.4 Analysis

In this section I use the packages `ggplot2` *for the graphs,* `dplyr` *for the data manipulation and* `readr` *to read in my csv data files.*

Once I obtained all the simulated data, I found the modulus of the bias for each sample size N, each k and each distribution. I then selected the information for all k with $N = 100, 25000$ and 50000, to display in Tables 3.1, 3.6 and 3.10, this involved taking `Data` and subtracting either `NormalEnt(sd=1)`, `UniformEnt(min=0, max=100)` or `ExpoEnt(rate=0.5)` from the estimators, depending on the distribution.

Next, I plotted graphs for each k of the logarithm of the bias of the estimator $\hat{H}_{N,k}$ against the logarithm of the sample size N, shown in Figures 3.1, 3.2, 3.6, 3.7, 3.11 and 3.12. I used the following code to do this, changing the `y` value to either `k1, k2, ..., k11` depending on which value of k I was plotting. Also the `data` would be read in from a different file for each distribution, the code below shows plotting the simulations from the normal distribution with $k = 1$.

```
# read in the data as a data frame
data <- as.data.frame(read_csv("./Data/data_normal.csv"))

# find the modulus of the bias for all n and k
data[-1] <- abs(data[-1] - NormalEnt(1))
```

```
# take the logarithm of everything
logdata <- log(data)

# the max and min x values
xmin <- min(logdata$n)
xmax <- max(logdata$n)

# the min and max y values
ymin <- -15 # this is because there are only 5 values smaller
    than -15
ymax <- ceiling(max(logdata[-1]))

# plot the graph for each k - here k=1
# defining the data
ggplot(data=logdata, aes(x=n, y=k1)) +
  # plotting the points
  geom_point(size=0.8) +
  # adding a linear regression line
  geom_smooth(method="lm") +
  # labelling the axis
  xlab("log(N)") +
  ylab("log|Bias(H)|") +
  # setting the axis limits
  xlim(c(xmin, xmax)) +
  ylim(c(ymin, ymax)) +
  # choosing the graph theme
  theme_minimal()
```

Additionally, I created a summary table of the useful information needed, containing the coefficients of the intercept ζ and the gradient $-a_k$ from the regression

analysis, the coefficient of determination R^2 and the standard error σ also from the regression analysis. I also modified ζ and$-a_k$ to find both a_k and c_k. The code below shows how I did this for the normal distribution, and it is similar for the other two distributions, just changing the input data and the exact value of entropy used to find the bias.

```
# read in the data as a data frame
data <- as.data.frame(read_csv("./Data/data_normal.csv"))

# find the modulus of the bias for all n and k - removing the
    1st column, n
data[-1] <- abs(data[-1] - NormalEnt(1))

# take the logarithm of everything
logdata <- log(data)

# initalise and empty df with everything in
Info <- data.frame(k = 1:11, ak = rep(0, 11),
                   zeta = rep(0, 11), powera = rep(0, 11),
                   ck = rep(0, 11), rsquared = rep(0, 11),
                   sigma = rep(0, 11))

# fill in data frame
for (k in 1:11){
  # find linear relationship of logarithm of bias against
     logrithm of n
  reg <- lm(logdata[[k+1]] ~ logdata$n)

  # the coeffs of log(bias)
  zeta <- round(reg$coefficients[["(Intercept)"]], 4)
  ak <- round(reg$coefficients["logdata$n"], 4)
```

```
  # the coeffs of normal bias
  ck <- round(exp(reg$coefficients[["(Intercept)"]]), 4)
  powera <- -ak

  # find the R squared value
  rsquared <- summary(reg)$r.squared

  # find the standard error
  sigma <- summary(reg)$sigma

  # fill in the each row for k=k
  Info[k,] <- c(k, ak, zeta, powera, ck, rsquared, sigma)
}

# save the Info data to a csv file
write_csv(Info, "../Data/normal_info.csv")
```

These tables are shown in Appendix B, and from these I found the information in Tables 3.3, 3.4, 3.7, 3.8, 3.11 and 3.12. Then to create Tables 3.5, 3.9 and 3.13, I just had to modify the summary tables found above to include two extra columns with the k^{a_k} and $\frac{k^{a_k}}{c_k}$, which was done by the following:

```
# read in the summary data as a data frame
Info <- as.data.frame(read_csv("./Data/normal_info.csv"))

# make sure k is an integer not a factor for the following
    computation
Info$k <- as.integer(Info$k)

# create a new data frame, Info2 with c, k^a and (k^a)/c
Info2 <- Info %>%
```

```
  mutate("k^a" = k^-ak, "(k^a)/c" = ((k^-ak)/ck)) %>%
  select(`k^a`, ck, `(k^a)/ck`)
```

From this table I than created the graphs shown in Figures 3.3, 3.8 and 3.13, using the below code:

```
# Graph (a) k against c
ggplot(data=Info2, aes(x=k, y=ck)) +
  # plotting the points
  geom_point() +
  # x axis labels
  scale_x_continuous(breaks = c(2:11), labels = c(2:11)) +
  theme_minimal()

# Graph (b) k^a against c
ggplot(data=Info2, aes(x=`k^a`, y=ck)) +
  # plotting the points
  geom_point() +
  theme_minimal()
```

The last part of analysis conducted, was plotting all the regression lines of the logarithm of N against the logarithm of the bias, for each k on the same graph. To do this I used the summary data, read in as `Info`, and the `xmin`, `xmax`, `ymin` and `ymax` found when plotting the graphs for each k separately. The following code was then used to create the graphs in Figures 3.4, 3.9 and 3.14:

```
# make k a factor
Info$k <- as.factor(Info$k)

# plot graph of comparison for each k
ggplot()+
  # add the lines for each k
  geom_abline(aes(intercept=zeta, slope=a, colour=k), data=
      Info, size=1) +
```

```
# set the axis limits
ylim(c(ymin, ymax)) +
xlim(c(xmin, xmax))+
# set the axis labels
xlab("log(N)") +
ylab("log(Bias(H))") +
# set the graph title
ggtitle("Comparison of the regression lines for Normal
    distribution")
```

For Figures 3.5, 3.10 and 3.15, the enlarged versions of Figures 3.4, 3.9 and 3.14, I changed the axis limits in the above code to be: `ymin = -9.5`, `ymax = -7.5`, `xmin = 9`, and `xmax = 11`.

A.5 Conclusion

The following code is how I created the graphs in Chapter 4. I changed `N` depending on what range of numbers I wished to have in my graph. I used the packages `readr` to read in my simulated data, `ggplot2` to plot the graphs and `dplyr` and `tidyr` to change the shape of my data. I have to work with `Inf` values in the uniform data for $k = 1$ and 2 and in the exponential data for $k = 1$, so I set these to `NA` values to make use of the `na.rm` argument, which is used in many functions. The following code plots an example graph for the range of data from $N = 100, ..., 1000$:

```
# choose a sample size
N <- seq(100, 1000, 100) # chosen this for example

# read in the data as a data frame for each distribtuion
# select the rows with N specified above
# remove the column n
Ndata <- as.data.frame(read_csv("./Data/data_normal.csv"))
  %>%
```

```
  filter(n %in% N) %>%
  select(-n)
Udata <- as.data.frame(read_csv("./Data/data_uniform.csv"))
  %>%
  filter(n %in% N) %>%
  select(-n)
Edata <- as.data.frame(read_csv("./Data/data_expo.csv")) %>%
  filter(n %in% N) %>%
  select(-n)

# setting all Inf values to NA, so that we can use na.rm in
   mean function
Udata$k1[is.infinite(Udata$k1)] <- NA
Udata$k2[is.infinite(Udata$k2)] <- NA
Edata$k1[is.infinite(Edata$k1)] <- NA

# work out the mean of the estimated entropy
# removing any NA values
Ndata <- Ndata %>%
  summarise_each(funs(mean))
Udata <- Udata %>%
  summarise_each(funs(mean(., na.rm = TRUE)))
Edata <- Edata %>%
  summarise_each(funs(mean(., na.rm = TRUE)))

# turning data around rows become columns and finding the
   bias
Ndata <- as.data.frame(t(Ndata)) %>%
  mutate(normal = abs(V1 - NormalEnt(1)))
Udata <- as.data.frame(t(Udata)) %>%
```

```
  mutate(uniform = abs(V1 - UniformEnt(min=0, max=100)))
Edata <- as.data.frame(t(Edata)) %>%
  mutate(expo = abs(V1 - ExpoEnt(0.5)))

# join all the df's together and adding kth column
df <- bind_cols(list(as.data.frame(Ndata$normal),
                     as.data.frame(Udata$uniform),
                     as.data.frame(Edata$expo))) %>%
  mutate(k=1:11)

# Change the column names
colnames(df) <- c("Normal", "Uniform", "Exponential", "k")

# change to different format
df <- gather(df, key= distribution, value=bias, Normal,
    Uniform, Exponential)

# Remove any rows which = NaN - this occurs when all values
    are NA
df <- df[!is.na(df$bias),]

# find the maximum value of bias
ymax <- max(df$bias)

# plot the graph
ggplot(aes(x=k, y=bias, col = distribution), data=df) +
  # add the points
  geom_point() +
  # add a line between points
  geom_line() +
```

```
# labels on the x axis and location of ticks
scale_x_continuous(labels = (1:11), breaks = (1:11)) +
# setting axis limits and labels
ylim(0, ymax) +
ylab("Bias of estimator") +
# adding a title to the graph
ggtitle(paste0("Average bias for samples of size N=",
                min(N), " to N=", max(N)))
```

Appendix B

Data

For all the data from the simulations visit `https://github.com/KarinaMarks/Entropy-Estimators/blob/master/Data/data_normal.csv` for the normal distribution, `https://github.com/KarinaMarks/Entropy-Estimators/blob/master/Data/data_uniform.csv` for the uniform distribution and `https://github.com/KarinaMarks/Entropy-Estimators/blob/master/Data/data_expo.csv` for the exponential distribution.

The data included here is the summary data created from the above simulated data, by the code set out in Appendix A. Note that for the uniform and exponential tables, there is no data for $k = 1$, the reasons for this are set out previously in the appropriate chapters.

k	a_k	ζ	c_k	R^2	σ
1	-0.5054	-3.1396	0.0433	0.17661828686171607	1.0660838272774293
2	-0.549	-3.0807	0.0459	0.17925206277881248	1.1477000302094302
3	-0.6169	-2.4146	0.0894	0.2291675329227359	1.105346052553444
4	-0.8181	-0.402	0.669	0.3555947364645426	1.0759440094571222
5	-0.8486	-0.1942	0.8235	0.33224404864966217	1.175234753954676
6	-0.8976	0.4392	1.5514	0.42598413789031414	1.0179591933311047
7	-0.9464	0.8576	2.3576	0.4532271066441854	1.015543671615524
8	-0.9574	1.1638	3.2021	0.4622618103677057	1.0087975591274145
9	-0.9883	1.4942	4.4558	0.49615166230448043	0.9729852501959322
10	-1.0454	2.1448	8.5402	0.522710431578095	0.975910357146507
11	-1.0386	2.1686	8.7457	0.5838759856844608	0.8565835831228029

Table B.1: *Simulations of Samples from the Normal Distribution - Summary Data*

k	a_k		c_k	R^2	σ
1	0.0	0.0	0.0	0.0	0.0
2	-0.5125	-3.657	0.0258	0.17211090724847583	1.0981929352102586
3	-0.5048	-3.9189	0.0199	0.15757092983648552	1.140209400927365
4	-0.6593	-2.5522	0.0779	0.24854564236305457	1.1199964906509245
5	-0.7286	-1.8768	0.1531	0.27842980395606237	1.1459347878922286
6	-0.8645	-0.4959	0.609	0.4026851191777433	1.0286337970721233
7	-0.8648	-0.552	0.5758	0.36052121422676875	1.1252173370666898
8	-0.9688	0.6085	1.8377	0.4688930592953607	1.0072590047360632
9	-0.9492	0.4118	1.5095	0.46531938639431797	0.9940387256809089
10	-0.9801	0.8772	2.4041	0.5087587117521769	0.9408499419440619
11	-1.0765	1.9011	6.6932	0.621458425518452	0.8208055149732554

Table B.2: *Simulations of Samples from the Uniform Distribution - Summary Data*

k	a_k		c_k	R^2	σ
1	0.0	0.0	0.0	0.0	0.0
2	-0.5824	-2.605	0.0739	0.18426832540375937	1.197183154117289
3	-0.4941	-3.4729	0.031	0.15964776874615852	1.1073499326838805
4	-0.494	-3.6277	0.0266	0.1537967126177289	1.1320666489578564
5	-0.5727	-2.8098	0.0602	0.21511352626001026	1.0687100051851457
6	-0.65	-2.1208	0.1199	0.22770117233710857	1.1694342748147517
7	-0.6605	-2.0218	0.1324	0.24742681384495552	1.1252927366709649
8	-0.6067	-2.6048	0.0739	0.21065071093727172	1.1472557143016018
9	-0.648	-2.1296	0.1189	0.2508890257166511	1.0938570355565644
10	-0.6606	-2.0658	0.1267	0.22004533438334092	1.2149786893084469
11	-0.6365	-2.2611	0.1042	0.2017082704261763	1.2369403770481109

Table B.3: *Simulations of Samples from the Exponential Distribution - Summary Data*

Bibliography

[1] R Project. `https://www.r-project.org/about.html`. Accessed: 18th March 2017.

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

[3] 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.

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

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

[6] S. Delattre and N. Fournier. On the Kozachenko-Leonenko entropy estimator. *arXiv preprint arXiv:1602.07440*, 1, 2016.

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

[8] 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.

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

[10] A. Fraser and H. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33:1134–1140, 1986.

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

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

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

[14] 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.

[15] 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.

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

[17] Oliver Thomas Johnson. *Information theory and the central limit theorem*. World Scientific, 2004.

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

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

[20] H. Kuo and Y. Gao. Maximum entropy direct models for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:873–881, 2006.

[21] E. Learned-Miller and J. Fisher. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

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

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

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

[25] Joseph Victor Michalowicz, Jonathan M Nichols, and Frank Bucholtz. *Handbook of differential entropy*. CRC Press, 2013.

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

[27] Murray Rosenblatt. *Gaussian and non-Gaussian linear time series and random fields*. Springer Science & Business Media, 2012.

[28] A. Schneider, G. Hommel, and M. Blettner. Linear regression analysis. *Dtsch Ä Rztebl Int*, 107:776–782, 2010.

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

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

[31] 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.

[32] 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.

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

[34] H. Wickham. Hadley Wickham. `http://hadley.nz/`. Accessed: 18th March 2017.