Statistical Inference for Estimation of Entropy

Karina Marks

February 27, 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" (Fraser and Swinney, 1986) [9]. Thus the "entropy of S can be seen as the uncertainty of S" [14].

1.1.1 Shannon Entropy

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

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

1.1.2 Rényi and Tsallis Entropy

These entropies are for the order $q \neq 1$ and the construction of them relies upon the generalisation of the Shannon entropy 1.1. For a random vector $X \in \mathbb{R}^d$ with density function f, we define;

Rényi entropy

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

Tsallis entropy

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

When the order of the entropy $q \to 1$, both the Rényi, (1.2), and Tsallis, (1.3), entropies tend to the Shannon entropy, (1.1), 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; and then progress to those for other types of entropy. Kapur and Kesavan's book on *Entropy Optimization Principles with Applications* [16], gives an account of some properties of the Shannon entropy H. First recall the definition of Shannon entropy (equation (1.1));

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

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

- H is permutationally symmetric
- For $f(x)$ continuous on some interval; H is also continuous everywhere in the same interval
- Entropy doesn't change by the inclusion of an impossible event
- $H > 0$ for all circumstances unless if f is any of the N degenerate distributions; where $f(x_i) = 1$ if $i = k, k \in [1, N]$ otherwise $f(x_i) = 0$, then $H = 0$
- H is a concave function
- The maximum value of H is attained by different distributions depending on how the distribution f is supported. For example, the maximum of H is attained when f is the;
  - Uniform distribution, if $supp\{f\} = [a, b]$, for $a, b, \in \mathbb{R}$
  - Exponential distribution, if $supp\{f\} = [0, \infty)$
  - Normal distribution, if $supp\{f\} = \mathbb{R} = (-\infty, \infty)$

- For two independent distributions ($f_X(x)$ and $f_Y(y)$), the entropy of their joint distribution ($f_{X,Y}(x,y)$) is just the sum of the entropies of the two distributions; $H(f_{X,Y}) = H(f_X) + H(f_Y)$

Shannon entropy, as mentioned earlier, is a special case of the Rényi and Tsallis entropies as $q \to 1$. 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;

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

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

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

Furthermore, there are some interesting relationships between the different specific types of entropy, for example Leonenko and Seleznjev [21] show the following relationship between H_2^* and H_∞^*;

$$H_\infty^* \leq H_2^* \leq 2H_\infty^* \tag{1.5}$$

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.

There are also some interesting properties of the general q-entropy; firstly, H_q is concave when $q > 0$ (convex when $q < 0$), implying for Shannon entropy, H is concave - as stated earlier. Also, the maximising distribution is the uniform distribution, for all q-entropies, as well as for Shannon, with a finite support. Lastly, for any d-dimensional sample ($d \geq 1$), given $\frac{d}{d+2} < q < 1$ and a covariance matrix, the q-entropy maximising distribution is of the multidimensional Student-t distribution [20].

1.2.2 Applications of Entropy

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

of information transmission. Thus, entropy has many applications across both these areas.

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

Wikipedia gives an appropriate overview of the applications, that the estimation of Shannon entropy is useful in "various science/engineering applications, such as independent component analysis, image analysis, genetic analysis, speech recognition, manifold learning, and time delay estimation" [27].

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

Image analysis is the investigation of an image and the extraction of useful information. Hero and Michel [13] first discuss the applications of Rényi entropy in image processing, then Neemuchwala, Hero and Carson [22] discuss how in image analysis an important task is that of image retrieval, which uses entropy estimation to compute entropic similarities that are used to match a reference image to another image. Moreover Du, Wang, Guo and Thouin, [7] considered the importance of entropy-based image thresholding; using both Shannon and relative entropy.

Genetic analysis is the study and research of genes and molecules to find information on biological systems. Statistical analysis of specific cells can help us understand how genomic entropy can help diagnose diseases and cancers. Wieringen and Vaart, [25] discuss how chromosomal disorganisation increases as cancer progresses, they mention how the K-L estimator can be used to help find this disorganisation/entropy; thus finding that "as cancer evolves, and the genomic entropy increases, the transcriptomic entropy is also expected to surge".

"Speech recognition (SR) is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers" [28]. Shen, Hung and Lee [23] discuss how an entropy based algorithm can conduct accurate SR in noisy environments. Moreover, Kuo and Gao [18] focus on a method where the probability of a state or word sequence given an observation sequence is computed directly from the maximum entropy direct model.

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 [26]. Dudewicz and van der Meulen [8]

discussed the property mentioned in section 1.2.1, that the uniform distribution maximises the entropy. Moreover, others have explored different distributions and their entropic properties; see [10, 4]

1.2.3 Other Estimators of Entropy

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

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

- Weak Consistency
$$H_N \xrightarrow{p} H(f) \tag{1.6}$$
- Mean Square Consistency
$$\mathbb{E}\{(H_N - H(f))^2\} \to 0 \tag{1.7}$$
- Strong Consistency
$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{1.8}$$
$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to \sigma^2 \tag{1.9}$$

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

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

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

$$H_N = -\int_{A_N} f_N(x) log(f_N(x)) dx \tag{1.10}$$

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

The resubstitution estimate is of the form;

$$H_N = -\frac{1}{N}\sum_{i=1}^{N} log(f_N(X_i))dx \tag{1.11}$$

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

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

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

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

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

$$H_{m,N} = \frac{1}{N}\sum_{i=1}^{N-m} log\left(\frac{N}{m}(X_{N,i+m} - X_{n,i})\right) - \Psi(m) + log(m) \tag{1.12}$$

where $\Psi(x)$ is the digamma function - more detailed explanation in section **??**. For a sample from a uniform distribution this estimator has been shown to be consistent; proved by Tarasenko [24]. Under some conditions on f, on its boundedness, the weak consistency and asymptotic normality was shown by Hall [11].

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

$$H_{m,N} = \frac{1}{N} \sum_{i=1}^{N-m_N} log\left(\frac{N}{m_N}(X_{N,i+m_N} - X_{n,i})\right) \tag{1.13}$$

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

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

Chapter 2

Estimation of Entropy

2.1 Kozachenko-Leonenko Estimator

2.1.1 History

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

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

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

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

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

Estimator with k=1

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

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

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

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 what is used in the latter papers, so it is useful to introduce it here. $\Psi(x)$ is the Digamma function, and when $x = 1$, this is just the negative Euler constant. Thus this estimator can be written if the form;

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

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.

Estimator with k fixed

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

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

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

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

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

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

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_1 by taking the limit of the estimator for the Tsallis entropy, $\hat{H}_{N,k,q}$ as $q \to 1$, again with a fixed value of k. This estimator is similar to that proposed in 1987, equation 2.2; however, it is now extended from the nearest neighbour to the kth nearest neighbour;

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

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 $q_1 > 1$; then H_1 exists and the estimator $\hat{H}_{N,k,1}$ is a consistent estimator for the Shannon entropy. This means that for large N, we have $\hat{H}_{N,k,1} \xrightarrow{L_2} H$; 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's asymptotically unbiased.

Estimator with k dependent on N

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

Considering dimensions $d \leq 3$, and a sample size N from distribution with density $f(x)$, they defined the k-nearest neighbour estimator of entropy - just as in section 2.1.1 - 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 differentaibility of f are proposed, with $k \in \{1, ..., O(N^{1-\epsilon})\}$, for some $\epsilon > 0$, we have asymptotic unbias of the estimator; where the bias can be expressed as;

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

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

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

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 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 here is the best possible.

It is important to note that for higher dimensions ($d > 3$), these results do not necessarily hold; since I am just considering the specific dimensions $d = 1$ and $d = 2$, there is no need to detail this. However, they do then go on to discuss a more appropriate estimator for higher dimensions, given sufficient smoothness, which is efficient in arbitrary dimensions, which was previously mentioned in section 1.2.3 - Other Estimators of Entropy.

2.1.2 Focus of this Paper

I now wish to more explicitly introduce the Kozachenko-Leonenko estimator of the entropy H, in the form that I will be considering. Let $X_1, X_2, ..., X_N$, $N \geq 1$ be independent and identically distributed random vectors in $\mathbb{R}^d$, and denote $\|.\|$ the Euclidean norm on $\mathbb{R}^d$.

- For $i = 1, 2, ..., N$, let $X_{(1),i}, X_{(2),i}, .., X_{(N-1),i}$ denote an order of the X_k for $k = \{1, 2, ..., N\} \setminus \{i\}$, such that $\|X_{(1),i} - X_i\| \leq \cdots \leq \|X_{(N-1),i} - X_i\|$. Let the metric ρ, defined as;

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

  denote the kth nearest neighbour or X_i.

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

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

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

  $$\Psi(k) = -\gamma + \sum_{j=1}^{k-1} \frac{1}{j} \tag{2.13}$$

  where $\gamma = 0.577216$ is the Euler-Mascheroni constant (where the digamma function is chosen so that $\frac{e^{\Psi(k)}}{k} \to 1$ as $k \to \infty$).

Then the Kozachenko-Leonenko estimator for entropy, H, is given by;

$$\hat{H}_{N,k} = \frac{1}{N} \sum_{i=1}^{N} log \left[\frac{\rho_{(k),i}^d V_d (N-1)}{e^{\Psi(k)}} \right] \tag{2.14}$$

where, $\rho_{(k),i}^d$ is defined in (2.11), V_d is defined in (2.12) and $\Psi(k)$ is defined in (2.13).

This paper focuses only on distributions for $d \leq 3$, more specifically, I will first be considering samples from 1-dimensional distributions, $d = 1$. Therefore, the volume of the 1-dimensional Euclidean ball is given by $V_1 = \frac{\pi^{\frac{1}{2}}}{\Gamma(\frac{3}{2})} = \frac{\sqrt{\pi}}{\frac{\sqrt{\pi}}{2}} = 2$. Hence the Kozachenko-Leonenko estimator is of the form;

$$\hat{H}_{N,k} = \frac{1}{N}\sum_{i=1}^{N} log\left[\frac{2\rho_{(k),i}(N-1)}{e^{\Psi(k)}}\right] \tag{2.15}$$

Later, I will be considering samples from 2-dimensional distributions; thus, $d = 2$ and the volume of the 2-dimensional Euclidean ball is given by $V_2 = \frac{\pi^{\frac{2}{2}}}{\Gamma(2)} = \frac{\pi}{1} = \pi$. Hence, the estimator takes the form;

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

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 2.1.1, where we have the conditions 1, 2 and 3, which imply the results stated by Theorems 1 and 2.

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

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

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

,

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

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

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

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

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

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

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

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

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

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

By section 2.1.1, 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 is more detail later to see which one of the two ideas show to be more true for the behaviors of the bias for a large value of N, in dimension $d = 1$ or $d = 2$.

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

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

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

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

Chapter 3

Monte Carlo Simulations

In this chapter I will explore simulations of the bias of estimator (2.14) in comparison to the size of the sample estimated from, with respect to different values of k; by exploring 1-dimensional distributions and then progressing onto 2-dimensional. Firstly, the distributions considered will be analysed to determine if they satisfy the conditions 1, 2 and 3 stated for Theorems 1 and 2 to hold. Then, I will explore the estimator of entropy for simulations of samples from certain distributions, for different values of k.

The motivation for these simulations is to explore the consistency of this estimator for different values of k; the relationship between the size of the bias of the estimator $\hat{H}_{N,k}$, $Bias(\hat{H}_{N,k})$, and the sample size, N. Throughout this analysis we will be considering the absolute value of this bias, since when considering its logarithm, we need a positive value. Using Theorem 1, we can write that the bias of the estimator approaches 0 as $N \to \infty$. This is because we can write $Bias(\hat{H}_{N,k}) = \mathbb{E}(\hat{H}_{N,k}) - H$, which in equation (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 modulus of the bias of the estimator, $|Bias(\hat{H}_{N,k})|$, and N. We believe this relationship is of the form;

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

for $a, c > 0$ [5, 3]. By taking the logarithm of this, we can generate a linear relationship, which is easier to analyse, and is given by;

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

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 mean finding the optimum value of k for which $|Bias(\hat{H}_{N,k})| \to 0$ for $N \to \infty$. For the relationship in equation (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 $(\frac{1}{N})^a$ (equation

(2.20)) or $(\frac{k}{N})^a$ (equation (2.21)). This leads to also examining the dependence of c on the value of k.

As I wish to consider the difference in accuracy of the estimator when using different values of k, let us denote the approximate values for a and c dependent on k as a_k and c_k.

3.1 1-dimensional Gaussian/Normal Distribution

3.2 1-dimensional Uniform Distribution

3.3 1-dimensional Exponential Distribution

3.4 2-dimensional Gaussian/Normal Distribution

Chapter 4

Conclusion

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[24] 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.

[25] 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.

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

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

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