Chapter 2 - Background

Karina Marks

February 27, 2017

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* [**?**], gives an account of some properties of the Shannon entropy H. First recall the definition of Shannon entropy (equation (**??**));

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

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;

$$\begin{aligned} H_2^* &= -log\left(\int_{\mathbb{R}^d} f^2(x)dx\right) \qquad (1) \\ &= -log\left(\sum_{x\in\mathbb{R}^d} f^2(x)\right) \end{aligned}$$

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 [**?**] show the following relationship between H_2^* and H_∞^*;

$$H_\infty^* \leq H_2^* \leq 2H_\infty^* \qquad (2)$$

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 [**?**].

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" [?].

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 [?] discussed how estimating the mutual information (MI) using entropy estimators is useful for assessing the independence of components from ICA. Learned-Miller and Fisher [?] 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 [?] first discuss the applications of Rényi entropy in image processing, then Neemuchwala, Hero and Carson [?] 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, [?] 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, [?] 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" [?]. Shen, Hung and Lee [?] discuss how an entropy based algorithm can conduct accurate SR in noisy environments. Moreover, Kuo and Gao [?] 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 [?]. Dudewicz and van der Meulen [?] discussed the property mentioned in section 1, that the uniform distribution maximises the entropy. Moreover, others have explored different distributions and their entropic properties; see [?, ?]

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) [**?**], 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
$$H_N \xrightarrow{p} H(f) \tag{3}$$

- Mean Square Consistency
$$\mathbb{E}\{(H_N - H(f))^2\} \to 0 \tag{4}$$

- Strong Consistency
$$\sqrt{N}(\hat{H}_{N,k} - H) \xrightarrow{d} N(0, \sigma^2) \tag{5}$$
$$N\mathbb{E}(\hat{H}_{N,k} - H)^2 \to \sigma^2 \tag{6}$$

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

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 [**?**], [**?**] 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;
$$H_N = -\int_{A_N} f_N(x) log(f_N(x)) dx \tag{7}$$
where the set A_N excludes the tail values of f_N. When the sample is from a 1-dimensional distribution, Dmitriev and Tarasenko, [**?**] 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 [**?**] 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;

$$H_N = -\frac{1}{N}\sum_{i=1}^{N} log(f_N(X_i))dx \tag{8}$$

which was first proposed in 1976, by Ahmad and Lin [**?**] who showed the mean-square consistency of this estimator, where f_N is a kernel density estimate. Joe [**?**] then went on to obtain the asymptotic bias and variance, and whilst satisfying certain conditions reduced the mean square error. Moreover, Hall and Morton [**?**] 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;

$$H_{m,N} = \frac{1}{N}\sum_{i=1}^{N-m} log\left(\frac{N}{m}(X_{N,i+m} - X_{n,i})\right) - \Psi(m) + log(m) \tag{9}$$

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 [**?**]. Under some conditions on f, on its boundedness, the weak consistency and asymptotic normality was shown by Hall [**?**].

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

$$H_{m,N} = \frac{1}{N}\sum_{i=1}^{N-m_N} log\left(\frac{N}{m_N}(X_{N,i+m_N} - X_{n,i})\right) \tag{10}$$

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 [**?**], 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.