Journal of Informatics and Mathematical Sciences
Vol. 9, No. 4, pp. 1117–1123, 2017
ISSN 0975-5748 (online); 0974-875X (print)
Published by RGN Publications
DOI: 10.26713/jims.v9i4.1023

http://www.rgnpublications.com

Research Article

Analyzing the Effect of Bin-width on the Computed Entropy

Sri Purwani[1,*], Sudradjat Supian[1] and Carole Twining[2]

[1] *Department of Mathematics, Padjadjaran University, Bandung, Indonesia*
[2] *Imaging Science, The University of Manchester, Manchester, UK*
***Corresponding author:** sri.purwani@unpad.ac.id

Abstract. The Shannon entropy is a mathematical expression for quantifying the amount of randomness which can be used to measure information content. It is used in objective function. Mutual Information (MI) uses Shannon entropy in order to determine shared information content of two images. The Shannon entropy, which was originally derived by Shannon in the context of lossless encoding of messages, is also used to define an optimum message length used in the Minimum Description Length (MDL) principle for groupwise registration. We first derived the Shannon entropy from the integral of probability density function (*pdf*), and thenfound that Gaussian has maximum entropy over all possible distribution. We also show that the entropy of the flat distribution is less than the entropy of the Gaussian distribution with the same variance. We then investigated the effect of bin-width on the computed entropy. We analyzed the relationship between the computed entropy and the integral entropy when we vary bin-width, but fix variance and the number of samples. We then found that the value of the computed entropy lies within the theoretical predictions at small and large bin-widths. Wealso show two types of bias in entropy estimators.

Keywords. Entropy; Objective function; Gaussian distribution; Flat distribution

MSC. 94Axx

Received: May 8, 2016 **Accepted:** August 15, 2016

Copyright © 2017 Sri Purwani, Sudradjat Supian and Carole Twining. *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

1. Introduction

The Shannon entropy is a measure of information. It is the underlying concept of mutual information [Pluim et al., 2000] used to determine shared information content of two images. Although some early papers [Viola, 1995, Viola and Wells III, 1997] used the 'Parzen Window' method for density estimation to compute MI, a review paper by Pluim et al. [Pluim et al., 2003] noted that the majority of the papers used histogram for computing MI, and hence the entropy (see e.g., [Studholme et al., 1995, Sabuncu and Ramadge, 2008, Twining and Taylor, 2011]). We then probed the effect of bin-width on the computed entropy, and analyzed the relationship between the computed entropy and the integral entropy when we vary bin-width, but fix variance. It is shown that the value of the computed entropy lies within the theoretical predictions at small and large bin-widths. In other case,suppose our data and our events really are discrete. According to Paninski and Carlton [Paninski, (2003), Carlton, (1969)], the entropy estimated from the finite sample using the probability estimate (1) is always an under-estimate on average. We then investigated the effect of sample size to the entropy of the actual distribution (a flat distribution). It is shown that as the number of samples increases, estimates approach the true value but always under-estimate the entropy. The following section defines the computed entropy and calculates the integral entropy.

2. The Computed and Integral Entropy

The Shannon entropy [Shannon, 1948] is defined as

$$E = -\sum_i P_i \log P_i \tag{1}$$

where P_i is the probability of bin i of a histogram. This can be approximated from the continuous one, the integral of probability density function (*pdf*), as follows

$$\begin{aligned}\int -\rho(x)\log\rho(x)\,dx &\approx -\sum_i w\left(\frac{P_i}{w}\right)\log\left(\frac{P_i}{w}\right),\\ &= -\sum_i P_i \log P_i + \log w\end{aligned} \tag{2}$$

where $\rho(x)$ and w are the*pdf* and the bin-width respectively. P_i is an area under $\rho(x)$ approximated by $P_i \approx \rho(x_i)\times w$. To understand this phenomenon (see Section 4), eqn. (2) is applied to a 1D Gaussian distribution with varying bin width, a fixed σ and probability density function of the form,

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{3}$$

The left hand side of eqn. (2) gives

$$\int -\rho(x)\log\rho(x)\,dx = -\int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right)\times\left(-\frac{x^2}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)\right)dx. \tag{4}$$

We need to transform eqn. (4) to the polar coordinate system by using,

$$(I(\alpha))^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(-\alpha x^2-\alpha y^2\right)dx\,dy \tag{5}$$

where $x^2 + y^2 = r^2$. The transformation results in

$$I(\alpha) = \int_{-\infty}^{\infty} exp\left(-\alpha x^2\right) dx = \sqrt{\frac{\pi}{\alpha}} \tag{6}$$

$$\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = \left(2\pi\sigma^2\right)^{\frac{1}{2}}$$

where $\alpha = \frac{1}{2\sigma^2}$. This shows the probability of the Gaussian over the whole range is equal to 1

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = 1.$$

Then, to proceed the remaining part of eqn. (4), we differentiate eqn. (6) with respect to α, which gives

$$-\sqrt{2\pi}\sigma^3 = -\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$$

or giving

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x^2 \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = \sigma^2.$$

Hence, the Gaussian entropy (4) is given by

$$-\int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)\left(-\frac{x^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right) dx = \frac{1}{2}\left(1 + \log\left(2\pi\sigma^2\right)\right). \tag{7}$$

The following section will derive the maximum entropy.

3. The Maximum Entropy

Using a fixed variance and some constraints:

$$\int \rho(x)dx = 1, \int x\rho(x)dx = 0, \int x^2\rho(x)dx = \sigma^2,$$

we construct the lagrangian

$$\mathscr{L} = \int -\rho(x)\log\rho(x)\,dx - \alpha\left(\int \rho(x)dx - 1\right) - \beta\left(\int x\rho(x)dx\right) - \gamma\left(\int x^2\rho(x)dx - \sigma^2\right)$$

where α, β and γ are the Lagrange multipliers. Taking the functional derivative with respect to $\rho$and equating it to zero gives an extremum at

$$\frac{\delta\mathscr{L}}{\delta\rho(x)} = -\log\rho(x) - 1 - \alpha - \beta x - \gamma x^2 = 0$$

$$\Longrightarrow \quad \rho(x) \approx \exp\left(-\left[\gamma x^2 + \beta x - \alpha'\right]\right), \quad \text{where } \alpha' = -1 - \alpha.$$

This shows that for a fixed variance, a Gaussian has maximum entropy over allpossible distributions [Lisman and Zuylen (1972), Brown (1992), Brun et al. (2011)]. As a comparison, a flat distribution with a probability density function

$$\rho(x) = \begin{cases} \frac{1}{2a} & -a \le x \le a \\ 0 & \text{otherwise} \end{cases}$$

has the entropy $\frac{1}{2}\log\left(12\sigma^2\right)$, which is less than the entropy of the Gaussian distribution (7) with the same variance.Using results from Section 2, we come to the comparison in the next section.

4. The Effect of Bin-width to the Computed Entropy

We can now substitute the result of eqn. (7) into eqn. (4) to obtain the relationship to the bin-width, which equation can be written as follows:

$$\frac{1}{2}\left(1+\log\left(2\pi\sigma^2\right)\right) \approx -\sum_i P_i \log P_i + \log w. \tag{8}$$

To test the relationship in eqn. (8), we vary the bin-width and use a fixed σ and n (the number of samples). When the bin-width $w \to 0$ for finite n, we will have in the limit n occupied bins, each with only a single entry. Then $P_i = \frac{1}{n}$ and the right side of eqn. (8) results in

$$-\sum_i P_i \log P_i + \log w \approx \log n + \log w.$$

On the other hand, when the bin-width $w \to \infty$ this gives $P_i \to 1$, hence

$$-\sum_i P_i \log P_i + \log w \approx \log w.$$

We plotted a graph (see Figure 1) to compare the Gaussian entropy to these results. We used a randomly generated set of numbers from MATLAB (which was generated only once) by using Gaussian distribution, with $n = 10000$ samples and $\sigma = 3$. On the same data we computed the histograms (hence P_i) for various different bin-widths. We plot the value of the left hand side of eqn. (8) as a horizontal straight line (Red), and the values of the right hand side of eqn. (8) over varying bin-width as the black circles. The theoretical predictions at small and large bin-widths are shown as the green and purple lines respectively.

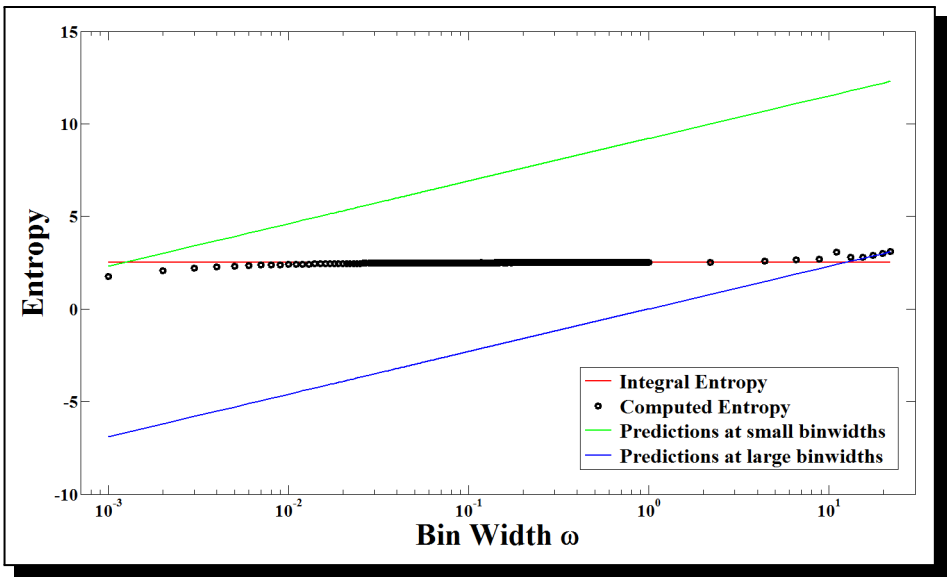

Figure 1. Comparison of integral and computed entropy

This graph shows that our analysis of the effect of bin-width on the computed entropy is correct, and that we can approximate the integral entropy by the entropy of a histogram provided we take enough care with our choice of bin-width. The following section introduces two types of bias in entropy estimators.

5. Limited Sample Size and Under-estimate the Entropy

We have seen how the entropy of a histogram can be used to estimate the integral entropy. But there is case we need to consider. Suppose our data and our events really are discrete. Then from a limited sample we are then trying to estimate the true bin probabilities. The usual maximum likelihood estimate is:

$$P_i = \frac{n_i}{n}$$

where n_i is the number of samples in bin i and n is the total number of samples. Then

$$E_{\text{estimate}} = -\sum_i P_i \log P_i.$$

Obviously, if n is small, some bins will be empty, when they have a non-zero probability. According to Paninski and Carlton [Paninski (2003), Carlton (1969)], the entropy estimated from the sample using this probability estimate is always an under-estimate on average. To demonstrate this, consider a flat distribution between 0 and 1, then divided into m equal bins $P_i = \frac{1}{m}$. Therefore, the entropy of the actual distribution is:

$$E = -\sum_{i=1}^{m} P_i \log P_i = \log m.$$

If we generate n random numbers, compute estimated probabilities $P_i \approx \frac{n_i}{n}$ and hence estimated entropy, we have the results shown in Figure 2.

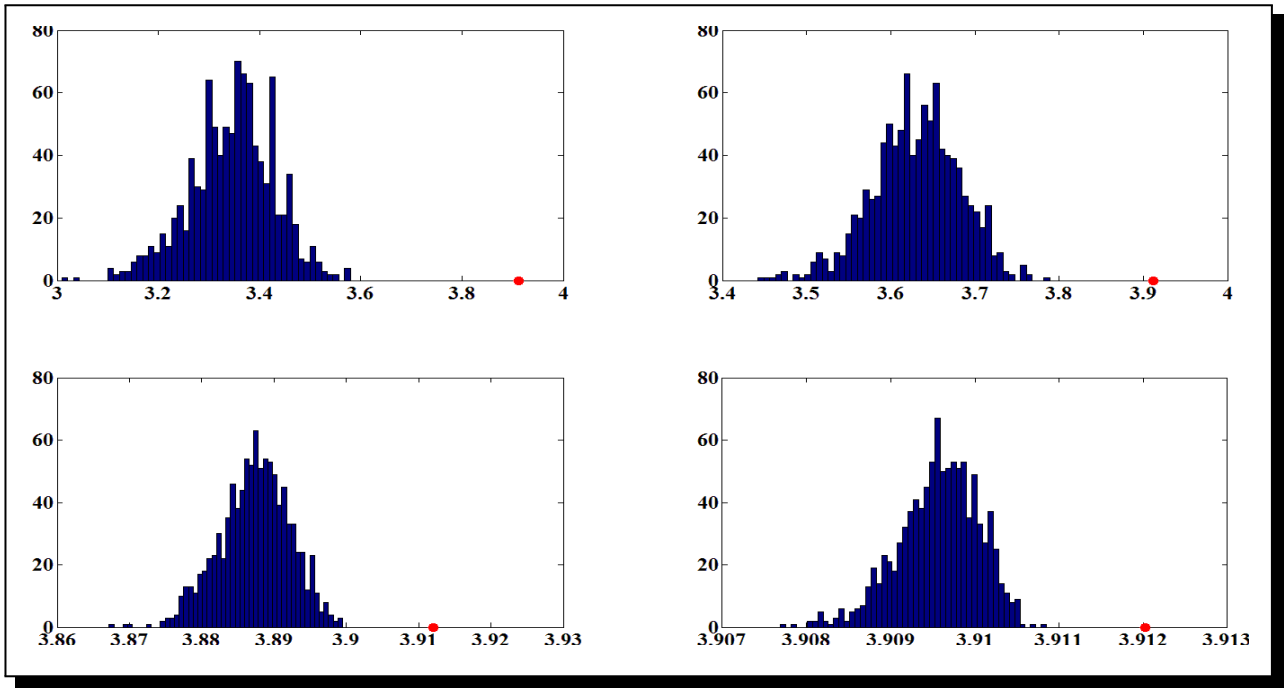

Figure 2. The number of samples on the top from left to right is 50 and 100 and on the bottom 1000 and 10000 respectively. The number of bins between 0 and 1 is kept fixed at $m = 50$. Estimated entropy is calculated repeatedly 1000 times and plotted in 50 bins. As n increases, estimates approach the true value (small red circle on the x-axis) but always under-estimate the entropy. Note varying scale on the x-axis

6. Conclusions

The effect of bin-width on the computed entropy suggests that we can approximate the integral entropy by the entropy of a histogram provided we take enough care with our choice of bin-width.

In other case, the effect of limited sample size to the estimated entropy of the actual distribution (see Figure 2) shows, as n increases, estimates approach the true value, but always under-estimate the entropy. This then shows the two types of bias in entropy estimators, as given by Moddemeijer [Moddemeijer (1989)],

- Approximating a continuous distribution by a histogram, and
- Finite sample size means the histogram of the sample different from the actual histogram.

Acknowledgement

Ms Purwani would like to thank the Directorate General of Higher Education of Indonesia (DIKTI) for providing funding for her PhD. She would also like to thank the Academic Leadership Grant 1-1-6 (ALG) led by Prof. Dr. Sudradjat Supian for all support given.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] L.G. Brown, A survey of image registration techniques, *ACM Computing Surveys* **24** (1992), 325 – 376.

[2] C.C. Brun, N. Leporé, X. Pennec, Y. Chou, A.D. Lee, G.D. Zubicaray, M.J. Wright, J.C. Gee, P.M. Thompson and et al., A non-conservative lagrangian framework for statistical fluid registration - SAFIRA, *IEEE Transactions on Medical Imaging* **30**(2) (2011), 184 – 202.

[3] A.G. Carlton, On the bias of information estimates, *Psychological Bulletin* **71**(2) (1969), 108 – 109.

[4] J.H.C. Lisman and M.C.A. van Zuylen, Note on the generation of most probable frequency distributions, *Statistica Neerlandica* **26**(1) (1972), 19 – 23.

[5] R. Moddemeijer, On estimation of entropy and mutual information of continuous distributions, *Signal Processing* **16**(3) (1989), 233 – 248.

[6] L. Paninski, Estimation of entropy and mutual information, *Neural Computation* **15**(6) (2003), 1191 – 1253.

[7] J.P.W. Pluim, J.B.A. Maintz and M.A. Viergever, Interpolation artefacts in mutual information-based image registration, *Computer Vision and Image Understanding* **77** (2000), 211 – 232.

[8] J.P.W. Pluim, J.B.A. Maintz and M.A. Viergever, Mutual-information based registration of medical images: a survey, *IEEE Transactions on Medical Imaging* **22**(8) (2003), 986 – 1004.

[9] M.R. Sabuncu and P. Ramadge, Using spanning graphs for effcient image registration, *IEEE Transactions on Image Processing* **17**(5) (2008), 788 – 797.

[10] C.E. Shannon, A mathematical theory of communication, *The Bell System Technical Journal* **27**(1948), 379 – 423.

[11] C. Studholme, D.L.G. Hill and D.J. Hawkes, Multiresolution voxel similarity measures for MR-PET registration, in *Information Processing in Medical Imaging*, Vol. 3 of Computational Imaging and Vision (1995), 287 – 298, Kluwer.

[12] C.J. Twining and C.J. Taylor, Specificity: a graph-based estimator of divergence, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(12) (2011), 2492 – 2505.

[13] P.A. Viola, *Alignment by Maximization of Mutual Information*, PhD thesis, Massachusetts Institute of Technology (1995).

[14] P. Viola and W.M. Wells III, Alignment by maximization of mutual information, *International Journal of Computer Vision* **24**(2) (1997), 137 – 154.