Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 3, pp. 705～712

A Comparison on the Differential Entropy

Daehak Kim[1]

Abstract

Entropy is the basic concept of information theory. It is well defined for random varibles with known probability density function(pdf). For given data with unknown pdf, entropy should be estimated. Usually, estimation of entropy is based on the approximations. In this paper, we consider a kernel based approximation and compare it to the cumulant approximation method for several distributions. Monte carlo simulation for various sample size is conducted.

Keywords : cumulant, density expansion, differential entropy, Entropy, kernel density

1. Introduction

Entropy is the basic concept of information theory. Entropy of random variable can be interpreted as an amount of randomness. Hyvärinen(2001) studied the connection between entropy and randomness by considering coding length. Entropy $H(X)$ is defined for a discrete random variable X as

$$H(X) = -\sum_i P(X=a_i)\cdot \log P(X=a_i) \tag{1.1}$$

where a_i's are the possible values of X. Depending on what the base of the logarithm is, different units of entropy are obtained. Usually, the logarithm with base e is used. Now let us define the function g as

$$g(p) = -p\log p, \quad \text{for } 0 \le p \le 1. \tag{1.2}$$

[1] Professor, Catholic university of Daegu, Department of Statistical information, Kyungsan, Korea.
E-mail : dhkim@cu.ac.kr

This is a nonnegative function that is 0 for $p=0$ and for $p=1$, and positive for vlaues in between. Using this function, entropy can be written as

$$H(X) = \sum_i g(P(X=a_i)). \tag{1.3}$$

Considering the shape of g, we see that the entropy is small if the probabilities $P(X=a_i)$ are close to 0 or 1, and large if the probabilities are in between 0 and 1. The definition of entropy for a discrete random variable can be generalized for continuous random variables and vectors, in which case it is often called differential entropy.

The differential entropy of a random variable X with probability density function $f_X(\cdot)$ is defined as

$$H(X) =- \int f_X(t)\log f_X(t)\,dt = \int g(f_X(t))\,dt \tag{1.4}$$

and will be abbreviated H. For the details about entropy, see Cover(1991) and Papoulis(1991).

However, when we don't know the true pdf, estimation of entropy H is required from the data itself. For the given sample $X_1, X_2, \cdots, X_n$ with unknown population, the approximations of entropy H can be considered.

In this paper, we investigate several approximation methods of entropy H and consider the application of kernel density estimates to the approximation of H. A comparisons on the estimates of differential entropy H is made via Monte Carlo simulation study.

2. Approximation of Entropy

In the previous section, we saw that entropy is a function of pdf. Entropy can be considered as a regularization measure that help us find the least structured density compatible with the measurements. In other words, the maximum entropy density can be interpreted as the density that is compatible with the measurements and makes the minimum number of assumptions on the data. This is because entropy can be interpreted as a measure of randomness and therefore the maximum entropy density is the most random of all the pdf's that satisfy the constraints.

The problem in using the entropy estimate is, however, that it is computationally very difficult. To use differential entropy in practice, we could compute the integral in the definition in (1.4). This is, however, quite difficult since the integral involves the unknown probability density function. The density could be estimated for the evaluation of entropy. In practice, some approximations, possibly rather coarse, have to be used. In this section, we discuss about the approximations of entropy.

2.1 Approximation by kernel density

We can use the basic kernel density estimator

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h}) \tag{2.1}$$

to the approximation of differential entropy H. Silverman(1995) introduced a good guide for kernel estimation of probability density estimation. Such a simple approach would be very ideal. But it should be used cautiously because the kernel density estimator would depend on the choice of the kernel parameters called bandwidth. Bowman(1984) considered the cross-validation method. It is the most popular method for the choice of data dependent bandwidth. Cross-validatory choice select the bandwidth h by minimizing

$$M_0(h) = \int \hat{f}^2 - \frac{2}{n}\sum_i \hat{f}_{-i}(X_i) \tag{2.2}$$

where

$$\hat{f}_{-i}(x) = \frac{1}{(n-1)h}\sum_{j\neq i} K(\frac{x-X_j}{h}) \tag{2.3}$$

Optimal bandwidth choice for given data is also available for certain cases. For example, in integrated mean squared error sense with normal kernel, we can get an optimal kernel by

$$h_{opt} = 1.06\sigma n^{-1/5} \tag{2.4}$$

Then, the entropy estimates based on kernel density estimator (2.1) can be

$$\hat{H}_K = -\int \hat{f}(x)\log\hat{f}(x)\,dx \tag{2.5}$$

The behaviour of this estimates is examined in section 3.

2.2 Approximation by polynomial density expansions

The classical method of approximating entropy is using higher-order cumulants.

These are based on the idea of using an expansion. This expansion is taken for the pdf of a random variable, say X, in the vicinity of the gaussian density. For simplicity, let us first make X zero-mean and of unit variance. Then, we can make the technical assumption that the density $p_X(\xi)$ of X is near the standardized gaussian density

$$\varphi(\xi) = exp(-\xi^2/2)/\sqrt{2\pi} \tag{2.6}$$

Two expansions are usually used in this context: the Gram-Charlier expansion and the Edgeworth expansion. They lead to very similar approximations. Gram-Charlier expansion use the so-called Chebyshev-Hermite polynomials, denoted by H_i where the index i is a nonnegative integer. These polynomials are defined by the derivatives of the standardized gaussian pdf $\varphi(\xi)$ by the equation

$$\frac{\partial^i \varphi(\xi)}{\partial \xi^i} = (-1)^i H_i(\xi)\varphi(\xi) \tag{2.7}$$

Thus, H_i is a polynomial of order i. The Gram-Charlier expansion of the pdf of X, truncated to include the two first non constant terms, is then given by

$$p_X(\xi) \approx \hat{p}_X(\xi) = \varphi(\xi)(1 + \kappa_3(X)\frac{H_3(\xi)}{3!} + \kappa_4(X)\frac{H_4(\xi)}{4!}) \tag{2.8}$$

where $\kappa_3(X) = E\{X^3\}$ is the third order cumulant and $\kappa_4(X) = E\{X^4\} - 3$. Note that the expansion starts directly from higher-order cumulants due to the standardization of X to zero mean and unit variance. Now we could plug the density in (2.8) into the definition of entropy, to obtain

$$H(X) \approx - \int \hat{p}_x(\xi) \log \hat{p}_x(\xi) d\xi \tag{2.9}$$

This integral is not very simple to evaluate, though. But again using the idea that the pdf is very close to a gaussian one, we see that the cumulants in (2.8) are very small, and thus we can use the simple approximation

$$\log(1+\epsilon) \approx \epsilon - \epsilon^2/2 \tag{2.10}$$

which gives

$$H(X) \approx - \int \varphi(\xi) \log \varphi(\xi) d\xi - \frac{\kappa_3^2(X)}{2 \times 3!} - \frac{\kappa_4^2(X)}{2 \times 4!} \tag{2.11}$$

Thus we finally obtain an approximation of the entropy of a standardized random variable as

$$H(X) \approx - \int \varphi(\xi)\log\varphi(\xi)\,d\xi - \frac{1}{12}E\{X^3\}^2 + \frac{1}{48}\mathrm{kurt}^2(X) \tag{2.12}$$

where $\mathrm{kurt}(X) = E(X^4) - 3$. For the given sample, we can get the entropy estimates

$$\hat{H}_C \approx - \int \varphi(\xi)\log\varphi(\xi)\,d\xi - \frac{1}{12}\hat{E}\{X^3\}^2 + \frac{1}{48}\widehat{\mathrm{kurt}}^2(X) \tag{2.13}$$

by the cumulant approximation.

3. Simulation Study

In this section, we consider the comparison of kernel density entropy estimates $\hat{H}_K$ and cumulant approximation entropy $\hat{H}_C$ which are explained in section 2. For the comparison, we consider the 4 distributions, standard normal, lognormal with mean 0 and variance 1, uniform and double exponential distribution, respectively as a population distribution with mean 0 and variance 1. Various sample sizes are considered for each case. Due to the heavy dependence of kernel density estimate (2.1) to the bandwidth h, we allow the two bandwidth selections, cross-validatory choice and optimal choice simultaneously to the entropy estimation of $\hat{H}_K$.

For the better understanding of the approximation of entropy, we calculated true entropy H with known the 4 probability density functions. In this case, numerical integration of (1.4) and cumulant approximation of (2.9) with exact cumulants are also considered. The results are shown in table 1. That means, the values in table 1 is calculated with known distributional properties without sample variation.

Table 1. Entropy of several distributions.

Distributions		True entropy	Numerical integral approximation	Cumulant approximation
Standard normal $N(0,1)$		1.41893	1.41831	1.23144
Lognormal $(0, 0.5^2)$		0.72579	0.72525	0.61608
Uniform$(-\sqrt{3}, \sqrt{3})$		1.24245	1.24242	1.23143
Exponential	$(\lambda = 1)$	1.00000	0.98264	-4.58106
	$(\lambda = \sqrt{2})$	0.65343	0.65203	0.85643

From the table 1, we can find that numerical integration approximation estimates very close to the true entropy for all the considered cases. It looks that cumulant approximation method underestimates a little bit but works well generally except exponential distribution in the presence of no sampling variation. The shape of exponential distribution with $\lambda = 1$ near 0 may be the reason of the negative entropy estimate.

For the comparison of entropy estimates from the given data, we generated random samples from the 4 distributions, respectively with sample size n. Two symmetric distributions, standard normal and uniform distribution $(-\sqrt{3}, \sqrt{3})$, and two asymmetric distributions, lognormal and exponential distribution are considered. We allowed $n = 5, 10, 20, 30, 50$. For the accurate evaluation of comparison, we have $1,000$ replication for each case. The grid for the cross-validatory choice was set to 100 for appropriate integral range. For optimal choice of h, we used (2.4). Simpson method is used for numerical integration.

In order to see the average performance of each entropy estimates, mean and standard error of $1,000$ estimates of entropy for each method are calculated. The results are shown in table 2 and table 3 for the considered 4 distributions, respectively. Symmetric distributions considered have mean 0 and variance 1, so there are no need to standardization.

Table 2. Entropy estimates for symmetric distributions.

n	$N(0,1)$			Uniform$(-\sqrt{3}, \sqrt{3})$		
	$\hat{H}_K$		$\hat{H}_C$	$\hat{H}_K$		$\hat{H}_C$
	CV	OPT	CU	CV	OPT	CU
5	1.4174(0.591)	1.4338(0.397)	1.7623(0.971)	1.0179(0.279)	1.1652(0017)	1.4471(0.016)
10	1.5131(0.381)	1.4945(0.249)	1.5860(0.395)	1.0657(0.172)	1.1708(0.018)	1.4507(0.021)
20	1.5155(0.242)	1.5028(0.164)	1.5060(0.157)	1.1229(0.092)	1.1783(0.015)	1.4486(0.011)
30	1.4987(0.209)	1.5037(0.133)	1.4791(0.101)	1.1441(0.066)	1.1836(0.013)	1.4487(0.009)
50	1.4920(0.135)	1.4910(0.103)	1.4526(0.068)	1.1673(0.038)	1.1896(0.010)	1.4485(0.007)

Figure 1. Convergence plot for the three estimates.
(left : $N(0,1)$, right : Uniform$(-\sqrt{3}, \sqrt{3})$)

For normal distribution, it looks that the three estimates works well generally. Kernel based approximation method looks better than cumulant approximation method for uniform distribution. It reveals that cumulant approximation can't be a good method anymore when losing normality. Note that for uniform distribution, the pdf is continuous but not a smooth function.

In order to see the trend of three estimates for the change of sample sizes, we depict the convergence plot in figure 1. From the figure 1, we can see that entropy based on kernel estimates converges to the true entropy as sample size goes to infinity provide the same results regardless of the choice of bandwidth. Cumulant approximation method estimates entropy very accurately for normal distribution but for uniform distribution, there may be no improvement in entropy estimats as sample size increases. Figure 2 represents box and whisker plot for the three estimates with 1,000 replications.

Figure 2. Box and whisker plot for the three estimates when $n=50$. (left : normal, center : uniform, right : exponential)

Table 3. Entropy estimates for asymmetric distributions.

n	Lognormal $(0, 0.5^2)$			Exponential $(\lambda=1)$		
	$\hat{H}_K$		$\hat{H}_C$	$\hat{H}_K$		$\hat{H}_C$
	CV	OPT	CU	CV	OPT	CU
5	0.5266(0.337)	0.7931(0.016)	1.4519(0.020)	0.4909(0.330)	0.7962(0.013)	1.4554(0.022)
10	0.5934(0.201)	0.7886(0.014)	1.5008(0.104)	0.5072(0.200)	0.7896(0.015)	1.5338(0.132)
20	0.6161(0.128)	0.7779(0.022)	1.6136(0.364)	0.5454(0.121)	0.7772(0.024)	1.6890(0.389)
30	0.6310(0.103)	0.7706(0.028)	1.7611(0.774)	0.5716(0.093)	0.7677(0.034)	1.8637(0.786)
50	0.6538(0.068)	0.7650(0.025)	1.8806(0.944)	0.6006(0.063)	0.7592(0.033)	2.0103(0.848)

For lognormal distribution and exponential distribution, the results are shown in table 3. In this case, kernel based approximation method works better than the cumulant based approximation. Cumulant approximation method provides a large standard errors than kernel approximation method. By considering the asymmetry of distributions and standardization, we can grasp the superiority of kernel based approximation than cumulant based approximation for considered distributions..

4. Concluding Remarks

We considered the approximation of differential entropy. Kernel density based approximation method for the entropy estimates is proposed and compared to cumulant based approximation method. From the simulation result, we can find that differential entropy estimate for given sample depends heavily on the distribution of underlying population. It is rarely hard to find good entropy estimates always powerful. Cumulant approximation methods works well generally for normal family. Proposed kernel based approximation method for differential entropy can be a good candidate in case of non-normal family of distributions.

References

1. Bowman, A. (1984), An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, vol. 71, 353-360.
2. Cover, T. M. and Thomas, J. A. (1991), *Elements of Information theory*, John Wiley & Sons.
3. Hyvärinen, A., Karhunen, J. and Oja, E. (2001), I*ndependent Component Analysis*, Wiley-Interscience

4 Papoulis, A. (1991), Probability, *Random variables and Stochastic Process*, McGraw-Hill, 3rd edition.

5. Silverman, B. W. (1985), *Density estimation for statistics and data analysis*, Chapman and Hall, London.

[received date : Jul. 2005, accepted date : Aug. 2005]