Iranian Int. J. Sci. ***6****(1), 2005, p.119-129*

A Version of the Entropy Estimator via Spacing

E. Pasha[1], M. Kokabi Nezhad[2], G. R. Mohtashami[3]
[1]*Department of Mathematics, Teacher Training University, Tehran, Iran.*
[2]*Corresponding author: Department of Statistics, Science and Research Branch, Islamic Azad University, Tehran, Iran.*
[3]*Department of Mathematics, Birjand University, Birjand, Iran.*
(received: 6/1/2005 ; accepted: 22/2/2005)

Abstract

In this paper we presented a version of the entropy estimator in view of Vasicek (1976) and Ebrahimi *et al.*, (1994). Some of its properties and a comperative study of this estimators are considered.

Keywords: *Shannon Entropy, Kullback Leibler Information, Entropy Estimator.*

Introduction

Shannon entropy has an important role in information theory. In fact, the amount of uncertainty related to the observation of the random variable X with pdf f is measured by Shannon entropy which is defined as $H(f) = E(-\log f(X))$.

There is an extensive literature on the nonparametric estimation of the Shannon entropy, for instance Ahmad *et al.*, (1976), Vasicek (1976), Joe (1987), Arizono *et al.*, (1989), Makkadem (1989), Van Es (1992), Ebrahimi *et al.*, (1994), Correa (1995) and Beirlant *et al.*, (1997).

One method for the nonparametric estimation of entropy which is considered by many authors, is the estimation of entropy as $H_n(f) = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_n(X_i)$ based on the random sample $X_1,...,X_n$ from continuous probability density function or pdf f with some estimation of density function.

On the other hand, from the definition of entropy $H(f)$, it can be easily seen that

$$H(f)=\int_0^1\left(\log\frac{d}{dp}F^{-1}(p)\right)dp,$$

and the estimation of $\frac{d}{dp}F^{-1}(p)$ can be derived by empirical distribution function via replacement of the derivation operator by a difference operator.

Based on order statistics, Vasicek (1976) proposed the entropy estimator as

$$H_v(m,n)=\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{Y_{i+m}-Y_{i-m}}{2\frac{m}{n}}\right\},$$

such that

$$\begin{cases}Y_{i+m}=Y_n; & i+m\geq n\\ Y_{i-m}=Y_1 & i-m\leq 1\end{cases}$$

where $Y_1<\ldots<Y_n$ are order statistics of the random variables $X_1,\ldots,X_n$ and m is a positive integer less than or equal to $\frac{n}{2}$, and is called a window size.

Ebrahimi *et al.*, (1994), expressed that the Vasicek estimator in the state of $i\leq m$ and $i\geq n-m+1$ is not a suitable formula for $\frac{d}{dp}F^{-1}(p)$. Therefore, they have presented two modified estimators, one of them is better and as follows

$$H_d(m,n)=\frac{1}{n}\sum_{i=1}^{n}\log\left\{\frac{Z_{i+m}-Z_{i-m}}{d_i\frac{m}{n}}\right\}$$

where $d_i = \begin{cases} 1+\dfrac{i+1}{m}-\dfrac{i}{m^2} & 1 \le i \le m \\ 2 & m+1 \le i \le n-m-1 \\ 1+\dfrac{n-i}{m+1} & n-m \le i \le n \end{cases}$,

and $\begin{cases} Z_{i-m} = a+\dfrac{i-1}{m}(Y_1-a) & 1 \le i \le m \\ Z_i = Y_i & m+1 \le i \le n-m-1 \\ Z_{i+m} = b-\dfrac{n-i}{m}(b-Y_n) & n-m \le i \le n \end{cases}$

that *a*, *b* are two known constants such that $P(a \le X \le b) \approx 1$.

In the next sections of this paper, we have modified the estimator H_d, and derived some characterization results of it. Finally, we have done a comparative study of the entropy estimators on using a simulation results.

Properties of the Modified Estimator

In Ebrahimi *et al.*, (1994) in which the estimator H_d has been introduced, it is assumed that the random variable X has bounded support, *a* and *b* are the lower and upper bounds of the support, respectively (for Uniform (0, 1) distribution, $a=0$, $b=1$); for the case X has lower (upper) bound, then $a(b)$ is the lower (upper) bound of the support (for Exponential (1) distribution, $a=0$, $b=\bar{x}+ks$). Also for the case X has unbounded support, $a=\bar{x}-ks$ and $b=\bar{x}+ks$, in which $\bar{x}$ and s are mean and standard deviation of sample, respectively.

Considering a, b as above has some problems which are listed as follows:

First, because the estimation is nonparametric, no information about the data distribution form and the support of it should be used, and in the situation where the distribution form and the context of data are unknown, therefore its support is unknown too.

Second, from our simulation results, if the data is generated from a longer tailed distribution, the amount of k should be increased as sample size increases (for example in lognormal(0,1) distribution, for sample size 30, k=5 is appropriate but for sample size 100, k should be selected at least 7). Specially, if underling distribution is skewed Normal, even with known context of data, the interval $(\bar{x}-ks\,,\bar{x}+ks)$ is not suitable.

Therefore, on assuming that the extreme values Y_n and Y_1 are not outliers, we propose range of the observations as follows:

$$a_n = Y_1 - \frac{Y_n - Y_1}{n-1} \quad \text{and} \quad b_n = Y_n + \frac{Y_n - Y_1}{n-1}.$$

For the case that there are outliers in the sample, we can use the methods which can tackle the outliers (if we look at the entropy estimator as arithmetic mean of the values $\{-\log \hat{f}(x_1),...,-\log \hat{f}(x_n)\}$, we can use robustness methods).

Thirdly, according to the results of our simulation, we have observed that the behavior of window size in H_d and H_v are noticeable. In some distributions, there are a big difference between the window size which causes minimum bias and minimum MSE (mean square errors). So, for solving this problem, we propose the following factors:

$$d_i = \begin{cases} 1+\dfrac{i}{i-1+m} & 1 \le i \le m \\ 2 & m+1 \le i \le n-m \\ 1+\dfrac{n-i+1}{n-i+m} & n-m+1 \le i \le n \end{cases} \qquad \text{if} \quad m < \frac{n}{2}$$

$$d_i = \begin{cases} 1+\dfrac{i}{i-1+m} & 1 \le i \le m \\ 1+\dfrac{n-i+1}{n-i+m} & n-m+1 \le i \le n \end{cases} \qquad \text{if} \quad m = \frac{n}{2}.$$

Our modified estimator in this note is H_k. The following theorems gives, constancy of MSE under linear transformation of the data, consistency and asymptotic unbias.

Theorem 1. Let $X_1,...,X_n$ be a sequence of *iid* random variables with entropy $H^{(X)}(f)$, and let $W_i = tX_i + l$ where $t > 0, -\infty < l < +\infty$ and i=1,…n. Let $H_k^{(X)}(m,n)$ and $H_k^{(W)}(m,n)$ be entropy estimators for $H^{(X)}(f)$, $H^{(W)}(g)$ respectively (here g is pdf of W). Then the following properties hold:

i) $E(H_k^{(W)}(m,n)) = \log(t) + E(H_k^{(X)}(m,n))$,
ii) $MSE(H_k^{(W)}(m,n)) = MSE(H_k^{(X)}(m,n))$.

Proof: On using the definition of $H_k(m,n)$, we have;

$$a_n^{(W)} = ta_n^{(X)} + l \qquad b_n^{(W)} = ta_n^{(X)} + l$$
$$Z_{i-m}^{(W)} = tZ_{i-m}^{(X)} + l \qquad Z_{i+m}^{(W)} = tZ_{i+m}^{(X)} + l$$
$$H_k^{(W)}(m,n) = \frac{1}{n}\sum_{i=1}^{n} \log\left\{\frac{Z_{i+m}^{(W)} - Z_{i-m}^{(W)}}{d_i \frac{m}{n}}\right\}$$
$$= \log(t) + H_k^{(X)}(m,n)$$

Therefore,

$$E(H_k^{(W)}(m,n)) = \log(t) + E(H_k^{(X)}(m,n))$$
$$Var(H_k^{(W)}(m,n)) = Var(H_k^{(X)}(m,n)).$$

Theorem 2. Let C be the class of continuous density functions with finite expectation and entropy and let $X_1,...,X_n$ be a random sample from $f \in C$, then $H_k(m,n) \longrightarrow H(f)$ in probability as

$$n, m \longrightarrow \infty \quad \text{and} \quad \frac{m}{n} \longrightarrow 0.$$

Proof. From Vasicek (1976), $H_v(m,n) \longrightarrow H(f)$ in probability, on the other hand, it can be easily shown that $H_k(m,n) \geq H_v(m,n)$, therefore, we will show that

$$H_k(m,n) - H_v(m,n) \longrightarrow 0 \quad \text{as} \quad n,m \longrightarrow \infty \quad \text{and} \quad \frac{m}{n} \longrightarrow 0.$$

We can write; $H_k(m,n) - H_v(m,n) = A_{mn} + B_{mn} + C_{mn} + D_{mn}$

where

$$A_{mn} = \frac{1}{n}\sum_{i=1}^{m} \log\left\{\frac{Z_{i+m} - Z_{i-m}}{Y_{i+m} - Y_{i-m}}\right\}, \qquad C_{mn} = \frac{1}{n}\sum_{i=1}^{m} \log\frac{2}{d_i}$$

$$B_{mn} = \frac{1}{n}\sum_{i=n-m+1}^{n} \log\left\{\frac{Z_{i+m} - Z_{i-m}}{Y_{i+m} - Y_{i-m}}\right\}, \qquad D_{mn} = \frac{1}{n}\sum_{i=n-m+1}^{m} \log\frac{2}{d_i}.$$

Via some algebraic calculation, we have

$$0 \leq \frac{1}{n}\sum_{i=1}^{m} \log\frac{2}{d_i} \leq \frac{m}{n}\log 2.$$

Hence, $C_{mn} \longrightarrow 0$ as $\frac{m}{n} \longrightarrow 0$.

Now; we have

$$B_{mn} = \frac{1}{n}\sum_{i=n-m+1}^{n} \log\left\{\frac{Z_{i+m} - Z_{i-m}}{Y_{i+m} - Y_{i-m}}\right\}$$

$$= \frac{1}{n}\sum_{j=1}^{m} \log\left\{1 + \frac{j}{m}\left(\frac{b_n - Y_n}{Y_n - Y_{j+n-2m}}\right)\right\}.$$

For $1 \leq j \leq m$, we have

$$0 \leq \frac{j}{m}\frac{b_n - Y_n}{Y_n - Y_{j+n-2m}} \leq \frac{b_n - Y_n}{Y_n - Y_{j+n-2m}} \leq \frac{b_n - Y_n}{Y_n - Y_{n-m}} \qquad (a.s)$$

$$0 \leq B_{mn} \leq \frac{m}{n}\log\{1 + \frac{b_n - Y_n}{Y_n - Y_{j+n-2m}}\} \leq \frac{m}{n}\frac{b_n - Y_n}{Y_n - Y_{n-m}} \qquad (a.s)$$

and $E(b_n - Y_n) = E(\frac{Y_n - Y_1}{n-1}) \le \frac{n}{n-1} E(X) < \infty$.

Therefore, $P(b_n - Y_n = \infty) = 0$, also $P(Y_n - Y_{n-m} = 0) = 0$,

then $P(\frac{b_n - Y_n}{Y_n - Y_{n-m}} = \infty) = 0$ and $E(\frac{b_n - Y_n}{Y_n - Y_{n-m}}) < \infty$.

Now, via Markov inequality, $\left(\frac{m}{n}\left\{\frac{b_n - Y_n}{Y_n - Y_{n-m}}\right\}\right) \longrightarrow 0$ in probability as

$\frac{m}{n} \longrightarrow 0$. Consequently $B_{mn} \longrightarrow 0$.

Similarly $D_{mn} \longrightarrow 0$ and $A_{mn} \longrightarrow 0$.

Theorem 3. Under the assumptions of the theorem 2,

$$E(H_k(m,n)) \longrightarrow H(f) \text{ as } n,m \longrightarrow \infty \text{ and } \frac{m}{n} \longrightarrow 0.$$

Proof. In proof of theorem 2, we obtained $E(\frac{b_n - Y_n}{Y_n - Y_{n-m}}) < \infty, \forall n$

and

$$0 \le E(B_{mn}) \le \frac{m}{n} E(\frac{b_n - Y_n}{Y_n - Y_{n-m}}).$$

Therefore $E(B_{mn}) \longrightarrow 0$ as $\frac{m}{n} \longrightarrow 0$ and similarly $E(A_{mn}) \longrightarrow 0$.

Hence; $0 \le E(H_k(m,n) - H_v(m,n)) \longrightarrow 0$ as $n,m \longrightarrow \infty$ and $\frac{m}{n} \longrightarrow 0$.

From Vasicek (1976), for the uniform distribution we have

$$E(H_v(m,n)) \longrightarrow 0.$$

Note that from mean value theorem, we can write;

$$H_v^{(F(X))}(m,n) = \frac{1}{n}\sum_{i=1}^{n} \log\left\{\frac{F(Y_{i+m}) - F(Y_{i-m})}{2\frac{m}{n}}\right\}$$

$$= H_v^{(X)}(m,n) + \frac{1}{n}\sum_{i=1}^{n} \log f(c_i)$$

where F is cdf of X, and for some $c_i \in (Y_{i-m}, Y_{i+m})$.

Consequently; $E(H_v(m,n)) \longrightarrow H(f)$ as $n,m \longrightarrow \infty$.

And via, $0 \le E(H_k(m,n) - H_v(m,n)) \longrightarrow 0$, we have

$$E(H_k(m,n)) \longrightarrow H(f).$$

Simulation Results

In this section, we give the results of simulation studies of the bias and MSE(mean square errors) for three entropy estimators H_v, H_d and H_k for some distributions such as Uniform (0,1), Exponential (1), Normal (0,1) and Lognormal (0,1). These results are based on more than 5000 samples of different sizes.

The best choice of the window size are given in the table 1. According to this table, we see that the window size which causes minimum bias is very different from the window size which causes minimum MSE for H_d in the uniform distribution and for H_v in the exponential, normal and lognormal distributions. This means that there is a big trade off between bias and variance (bias and variance move on opposite direction). The next point is that the window size which causes minimum bias in the H_d for uniform (0, 1) is equal to $\frac{n}{2}$ which increases as n increases that not suitable.

For campration the amount of reduction in bias and MSE, $m = [\sqrt{n} + 0.5]$ and $m = [\sqrt{n}]$ is used and results are given in tables 2 to 5. It can be compared with tables of Inverardi (2003), our modified estimator have smaller MSE with respect to the Vasicek(1976), van Es (1992), Ebrahimi *et al.*, (1994) and Correa (1995).

Here, because the appropriate selection problem of k, we don't consider H_d for exponential and lognormal distributions, and for normal distribution k=5 is selected, as considered in Ebrahimi *et al.*, (1994).

Table 1

	U(0,1)			Exp(1)		N(0,1)			LN(0,1)	
	H_k	H_d	H_v	H_k	H_v	H_k	H_d	H_v	H_k	H_v
n	**a**	**b**	**a**	**a**	**b**	**a**	**a**	**b**	**a**	**b**
20	8	10(10)	3	5	4(4)	10	2	3(3)	4	10(4)
30	9	15(15)	4	5	6(5)	13	2	4(4)	4	15(6)
40	9	20(15)	4	6	7(6)	12	3	4(4)	4	20(7)
50	10	25(17)	5	6	25(7)	11	3	5(5)	5	21(8)
60	11	30(21)	5	6	30(8)	11	3	6(6)	6	21(9)
70	12	35(25)	6	7	35(9)	11	4	7(7)	6	20(10)
80	13	40(27)	6	7	40(11)	11	4	7(7)	6	20(10)
90	13	45(30)	6	7	45(11)	11	4	8(8)	6	20(12)
100	13	50(31)	14	8	50(13)	11	4	9(9)	7	20(11)
300	18	150(45)	17	12	55(23)	14	6	57(14)	10	22(15)
500	24	250(99)	20	14	66(39)	16	9	88(37)	12	24(19)

a: Best window size to obtain minimum bias absolute and MSE entropy estimation
b: Best window size to obtain minimum bias absolute (MSE) entropy estimation.

Table 2. Uniform (0,1)

		H_v		H_d		H_k	
n	**m**	**|Bias|**	**MSE**	**|Bias|**	**MSE**	**|Bias|**	**MSE**
20	4	0.2603	0.0754	0.0859	0.0106	0.0201	0.0121
	5	0.2772	0.0838	0.0753	0.0080	0.0012	0.0080
30	5	0.2011	0.0438	0.0665	0.0061	0.0173	0.0050
	6	0.2122	0.0483	0.0590	0.0050	0.0023	0.0030
40	6	0.1712	0.0314	0.0559	0.0040	0.0125	0.0030
	7	0.1801	0.0344	0.0504	0.0030	0.0010	0.0020
50	7	0.1508	0.0241	0.0472	0.0020	0.0076	0.0021
	8	0.1583	0.0263	0.0429	0.0020	0.0013	0.0010
60	8	0.1371	0.0197	0.0409	0.0020	0.0119	0.0010
80	9	0.1144	0.0136	0.0351	0.0015	0.0039	0.0008
100	10	0.1015	0.0105	0.0311	0.0010	0.0037	0.0005

Table 3. Normal(0,1)

		H_v		H_d		H_k	
n	**m**	**|Bias|**	**MSE**	**|Bias|**	**MSE**	**|Bias|**	**MSE**
20	4	0.3314	0.1418	0.1631	0.0552	0.1433	0.0507
	5	0.3530	0.1566	0.2060	0.0710	0.1203	0.0449
30	5	0.2443	0.0801	0.1464	0.0401	0.0916	0.0283
	6	0.2517	0.0841	0.1798	0.0508	0.0730	0.0254
40	6	0.1959	0.0533	0.1357	0.0321	0.0653	0.0189
	7	0.2034	0.0565	0.1626	0.0410	0.0504	0.0173
50	7	0.1655	0.0392	0.1305	0.0278	0.0451	0.0136
	8	0.1706	0.0411	0.1530	0.0341	0.0323	0.0128
60	8	0.1423	0.0311	0.1414	0.0296	0.0275	0.0104
80	9	0.1137	0.0201	0.1137	0.0201	0.0185	0.0075
100	10	0.0927	0.0143	0.0297	0.0143	0.0085	0.0058

Table 4. Exponential(1)

		H_v		H_k	
n	**m**	**|Bias|**	**MSE**	**|Bias|**	**MSE**
20	4	0.2528	0.1215	0.0109	0.0552
	5	0.2553	0.1236	0.0248	0.0560
30	5	0.1905	0.0744	0.0013	0.0391
	6	0.1904	0.0749	0.0254	0.0405
40	6	0.1532	0.0506	0.0102	0.0286
	7	0.1527	0.0507	0.0317	0.0310
50	7	0.1312	0.0384	0.0235	0.0239
	8	0.1304	0.0385	0.0369	0.0242
60	8	0.1094	0.0306	0.0149	0.0188
80	9	0.0898	0.0215	0.0286	0.0144
100	10	0.0771	0.0171	0.0261	0.0119

Table 5. LogNormal(0,1)

		H_v		H_k	
n	**m**	**|Bias|**	**MSE**	**|Bias|**	**MSE**
20	4	0.2628	0.1607	0.0128	0.1020
	5	0.2549	0.1619	0.0378	0.1105
30	5	0.1894	0.0941	0.0221	0.0684
	6	0.1785	0.0923	0.0605	0.0756
40	6	0.1346	0.0632	0.0402	0.0514
	7	0.1265	0.0628	0.0710	0.0573
50	7	0.1085	0.0473	0.0504	0.0418
	8	0.1012	0.0471	0.0759	0.0467
60	8	0.0906	0.0388	0.0545	0.0365
80	9	0.0732	0.0289	0.0546	0.0265
100	10	0.0543	0.0209	0.0386	0.0204

Conclusions

In this paper, we have proposed a version of Shannon entropy estimator. Based on simulation results our estimator have smaller bias absolute and MSE than other estimators that mentioned and referred to them in this note.

References

Ahmad, I., and Lin, P. (1976) *A nonparametric estimation of the entropy for absolutely continuous distribution*. IEEE Trans. Inform. Theory IT-**22**, 327-375.

Arizono, I., and Ohta, H. (1989) *A test for normality based on Kullback-Leibler Information*. JASA, **43**(1), 20-22.

Beirlant, J., Dudewicz, E., Gyorfi, L., and Van Der Meulen, E. (1997) Nonparametric entropy estimation: An overview. Inter. J. Math. Stat. Sei., **6**, 17-39.

Correa, J. (1995) *A new estimator of entropy*. Commun. Statist. _Theory Meth. **24**, 2439-2449.

Ebrahimi, N., Pflughoeft, K., and Soofi, E. (1994) *Two measures of sample entropy*. Statistics & Probability Letters, **20**, 225-234.

Invrardi, A. (2003) *MSE comparison of some different estimators of entropy*. Commun. Statist. _Simu. Comp., **32**, 1, 17-30.

Joe, H. (1989) *Estimation of entropy and other functionals of a multivariate density*. Ann. Inst. Statist. Math, **41**, 683-697.

Makkadem, P. (1989) *Estimation of entropy and information of absolutely continuous random variables*. IEEE Trans. Inform. Theory IT- **35**, 193-196.

Shannon, C. (1949) *A mathematical theory of communication*. Bell System Tech. J., **22**, 379-423.

Van Es, B. (1992) Estimating *functionals related to a density by a class of statistics based on spacings*. Scand. J. Statist., **19**, 61-72.

Vasicek, O. (1976) *A test normality based on sample entropy*. J. Roy. Statist. Soc. ser B, **38**, 54-59.