Thailand Statistician
July 2019; 17(2): 190-197
http://statassoc.or.th
Contributed paper

A New Estimator for Shannon Entropy

Havva Alizadeh Noughabi [a] and Jalil Jarrahiferiz* [b]

[a] Department of Computer Engineering, University of Gonabad, Gonabad, Iran.

[b] Department of Mathematics, Birjand Branch, Islamic Azad University, Birjand, Iran.

*Corresponding author; e-mail: jarrahi@iaubir.ac.ir

Received: 28 May 2018
Revised: 15 September 2018
Accepted: 2 December 2018

Abstract

In this paper, we propose a new estimator for Shannon entropy of a continuous random variable. Consistency and other properties of the new estimator is stated. Through a Monte Carlo simulation, the mean squared error of the proposed estimator is compared with some prominent estimators, namely Vasicek's estimator, Van Es's estimator, and Correa's estimator. Finally, it is shown that the proposed estimator has smaller mean squared error than the other estimators.

Keywords: Information theory, sample entropy, spacings, local linear model, mean squared error.

1. Introduction

Shannon (1948) defined entropy of a continuous random variable X with density function $f(x)$ as

$$H(f) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx.$$

Shannon entropy has been widely applied in many fields such as model selection, goodness of fit tests and etc. We interest to estimate the entropy of a random variable based on a random sample. This problem has been considered by some authors such as Vasicek (1976), Van Es (1992), and Correa (1995). They proposed solutions for the problem of estimating the entropy and then introduced estimators for the population entropy.

Vasicek (1976) first expressed Shannon entropy as

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

and then by replacing the distribution function F by the empirical distribution function F_n, and using a difference operator instead of the differential operator proposed his estimator. By a function of the order statistics, Vasicek (1976) estimated the derivative of $F^{-1}(p)$. Vasicek's estimator based on a random sample $X_1,\ldots,X_n$ is as

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

where the window size m is a positive integer smaller than $n/2$, $X_{(i)} = X_{(1)}$ if $i < 1$, $X_{(i)} = X_{(n)}$ if $i > n$ and $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$ are order statistics based on a random sample of size n. Consistency of the estimator is proved by Vasicek (1976). He showed that $HV_{mn} \xrightarrow{\Pr.} H(f)$ as $n \to \infty$, $m \to \infty$, while $\frac{m}{n} \to 0$.

After Vasicek (1976), Van Es (1992) proposed an estimator of entropy given by

$$HVE_{mn} = \frac{1}{n-m}\sum_{i=1}^{n-m}\left(\frac{n+1}{m}(X_{(i+m)} - X_{(i)}) \right) + \sum_{k=m}^{n}\frac{1}{k} + \log(m) - \log(n+1).$$

He proved the consistency of the estimator and also asymptotic normality of this estimator under some conditions.

Based on a local linear model, Correa (1995) proposed an entropy estimator. He first considered the sample information as

$$(F_n(X_{(1)}), X_{(1)}), (F_n(X_{(2)}), X_{(2)}), ..., (F_n(X_{(n)}), X_{(n)})$$

and written HV_{mn} as

$$HV_{mn} = -\frac{1}{n}\sum_{i=1}^{n} \log\left\{ \frac{(i+m)/n - (i-m)/n}{X_{(i+m)} - X_{(i-m)}} \right\}.$$

Then he noted that the argument of log is the equation of the slope of the straight line that joins the points $(F_n(X_{(i+m)}), X_{(i+m)})$ and $(F_n(X_{(i-m)}), X_{(i-m)})$. Therefore, he used a local linear model based on $2m+1$ points to estimate the density of $F(x)$ in the interval $(X_{(i+m)}, X_{(i-m)})$,

$$F(x_{(j)}) = \alpha + \beta x_{(j)} + \varepsilon, \quad j = m-i, ..., m+i.$$

Through the least square method, he obtained an estimator of entropy as

$$HC_{mn} = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\sum_{j=i-m}^{i+m}(X_{(j)} - \bar{X}_{(i)})(j-i)}{n\sum_{j=i-m}^{i+m}(X_{(j)} - \bar{X}_{(i)})^2} \right),$$

where

$$\bar{X}_{(i)} = \frac{1}{2m+1}\sum_{j=i-m}^{i+m} X_{(j)}.$$

Some authors have used the entropy estimators and developed statistical procedures based on entropy. One can see for example, Arizono and Ohta (1989), Ebrahimi (1998), Esteban et al. (2001), Balakrishnan et al. (2007), Rad et al. (2011), Pakyari and Balakrishnan (2012, 2013), Noughabi and Arghami (2013a, b), and Noughabi and Balakrishnan (2015). Therefore, it would be interested to propose an efficient entropy estimator in practice. Our goal in this paper is to present a desirable entropy estimator.

In Section 2, we present a new entropy estimator and scale invariance of variance and mean squared error of it is proven. In Section 3, through a Monte Carlo simulation we compute root

mean square error (RMSE) of estimators under different distributions and finally the proposed estimator is compared with the existing estimators in terms of RMSE.

2. The Proposed Estimator and Its Properties

Suppose $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ are order statistics of a random sample of size n from an unknown continuous distribution F with a probability density function $f(x)$. We interest to estimate of entropy $H(f)$ of an unknown continuous probability density function f.

2.1. The new estimator

In Vasicek's estimator, since

$$\frac{X_{(i+m)} - X_{(i-m)}}{2m/n}$$

is not a correct formula for the slope when $i \le m$ or $i \ge n-m+1$, we must correct this formula. In order to a correct estimate the slopes at these points the denominator and/or the numerator should be modified for $i \le m$ or $i \ge n-m+1$. Toward this end, we construct the following estimator.

We propose to estimate the entropy $H(f)$ of an unknown continuous probability density function f by

$$HM_{mn} = -\frac{1}{n}\sum_{i=1}^{n} \log\{S_i(n,m)\},$$

where

$$S_i(m,n) = \begin{cases} \dfrac{m/n}{X_{(i+m)} - X_{(1)}}, & 1 \le i \le m, \\ \dfrac{\sum_{j=i-m}^{i+m} (X_{(j)} - \bar{X}_{(i)})(j-i)}{n\sum_{j=i-m}^{i+m} (X_{(j)} - \bar{X}_{(i)})^2}, & m+1 \le i \le n-m, \\ \dfrac{m/n}{X_{(n)} - X_{(i-m)}}, & n-m+1 \le i \le n, \end{cases}$$

and

$$\bar{X}_{(i)} = \frac{1}{2m+1}\sum_{j=i-m}^{i+m} X_{(j)}.$$

2.2. Properties of the Estimator

In the next theorem, we show that the scale of the random variable X has no effect on the accuracy of the proposed estimator in estimating $H(f)$.

Theorem 1 *Let $X_1,\ldots,X_n$ be a sequence of i.i.d. random variables with entropy $H^X(f)$ and $Y_i = cX_i$, $i=1,...,n$, where $c>0$. Also, let HM_{mn}^X and HM_{mn}^Y be entropy estimators for $H^X(f)$ and $H^Y(g)$, respectively (here g is pdf of $Y=cX$). Then the following properties hold.*

i) $E(HM_{mn}^Y) = E(HM_{mn}^X) + \log c$,

ii) $Var(HM_{mn}^Y) = Var(HM_{mn}^X)$,

iii) $MSE(HM_{mn}^Y) = MSE(HM_{mn}^X)$.

Proof: We have

$$S_i^Y(m,n) = \left\{ \begin{array}{ll} \dfrac{m/n}{Y_{(i+m)}-Y_{(1)}} = \dfrac{1}{c}\dfrac{m/n}{X_{(i+m)}-X_{(1)}}, & 1\le i\le m, \\ \dfrac{\sum_{j=i-m}^{i+m}(Y_{(j)}-\bar{Y}_{(i)})(j-i)}{n\sum_{j=i-m}^{i+m}(Y_{(j)}-\bar{Y}_{(i)})^2} = \dfrac{1}{c}\dfrac{\sum_{j=i-m}^{i+m}(X_{(j)}-\bar{X}_{(i)})(j-i)}{n\sum_{j=i-m}^{i+m}(Y_{(j)}-\bar{X}_{(i)})^2}, & m+1\le i\le n-m, \\ \dfrac{m/n}{Y_{(n)}-Y_{(i-m)}} = \dfrac{1}{c}\dfrac{m/n}{X_{(n)}-X_{(i-m)}}, & n-m+1\le i\le n, \end{array} \right\} = \frac{1}{c}S_i^X(m,n).$$

Therefore,

$$HM_{mn}^Y = -\frac{1}{n}\sum_{i=1}^n \log\left\{S_i^Y(m,n)\right\} = -\frac{1}{n}\sum_{i=1}^n \log\left\{\frac{1}{c}S_i^X(m,n)\right\} = HM_{mn}^X + \log c,$$

and consequently

$$E(HM_{mn}^Y) = E(HM_{mn}^X) + \log c,$$

$$Var(HM_{mn}^Y) = Var(HM_{mn}^X + \log c) = Var(HM_{mn}^X),$$

$$\begin{aligned} MSE(HM_{mn}^Y) &= Var(HM_{mn}^Y) + Bias^2(HM_{mn}^Y) = Var(HM_{mn}^X) + \left\{E(HM_{mn}^Y) - H^Y(g)\right\}^2 \\ &= Var(HM_{mn}^X) + \left\{E(HM_{mn}^X) + \log c - H^X(g) - \log c\right\}^2 \\ &= Var(HM_{mn}^X) + \left\{E(HM_{mn}^X) - H^X(g)\right\}^2 \\ &= Var(HM_{mn}^X) + Bias^2(HM_{mn}^X) \\ &= MSE(HM_{mn}^X). \end{aligned}$$

The following theorem establishes the consistency of the proposed estimator.

Theorem 2 *Let C be the class of continuous densities with finite entropies and let $X_1,\ldots,X_n$ be a random sample from $f\in C$. If $n\to\infty$, then*

$$HM_{mn} \xrightarrow{\Pr.} H(f).$$

Proof: Since Vasicek's and Correa's estimators are consistent and also by consistency of $\hat{f}(x)$ and continuity of $\hat{f}(x)$, the result is hold.

3. Simulation Study

In this section, we compared the new estimator with the other existing estimators. A Monte Carlo simulation is performed to analyze the behavior of the proposed estimator and other estimators. Here, the proposed estimator is compared with some prominent estimators, namely Vasicek's estimator, Van Es's estimator, and Correa's estimator. Correa (1995) considered three distributions normal, exponential and uniform for their comparisons, therefore we also consider these distributions for our comparisons.

We generate random samples from these distributions and then compute the values of estimators and finally based on 10,000 replications RMSE of the estimators are obtained. For the considered estimators, we chose value of m using the following heuristic formula (see Grzegorzewski and Wieczorkowski 1999):

$$m = \left\lfloor \sqrt{n} + 0.5 \right\rfloor.$$

Tables 1 to 3 show the values of root of mean squared error (RMSE) and standard deviation (SD) of the estimators at different sample sizes. We chose small and moderate sample sizes $n = 5, 10, 15, 20, 30, 40, 50$ and computed the root of mean squared error of the estimators. We can see from the values of tables that the RMSE tends to zero when the sample size increasing. Also, for large sample sizes, the values of RMSE are approximately the same. Therefore, we considered the sample sizes 5 to 50.

Table 1 Root of mean squared error (and standard deviation) of estimators in estimate of entropy $H(f)$ for standard normal distribution

n	m	RMSE (SD)			
		HV_{mn}	HVE_{mn}	HC_{mn}	HM_{mn}
5	2	0.994 (0.413)	0.509 (0.452)	0.793 (0.418)	0.536 (0.418)
10	3	0.618 (0.264)	0.366 (0.283)	0.470 (0.271)	0.292 (0.267)
15	4	0.474 (0.211)	0.318 (0.220)	0.348 (0.213)	0.207 (0.205)
20	4	0.373 (0.178)	0.276 (0.185)	0.265 (0.182)	0.179 (0.178)
30	5	0.282 (0.144)	0.243 (0.148)	0.194 (0.146)	0.144 (0.143)
50	7	0.198 (0.109)	0.212 (0.110)	0.135 (0.112)	0.120 (0.109)

Table 2 Root of mean squared error (and standard deviation) of estimators in estimate of entropy $H(f)$ for exponential distribution with mean one

n	m	RMSE (SD)			
		HV_{mn}	HVE_{mn}	HC_{mn}	HM_{mn}
5	2	0.930	0.596	0.743	0.580
		(0.559)	(0.586)	(0.554)	(0.555)
10	3	0.570	0.392	0.435	0.358
		(0.360)	(0.373)	(0.361)	(0.357)
15	4	0.421	0.310	0.328	0.296
		(0.282)	(0.290)	(0.290)	(0.284)
20	4	0.356	0.274	0.272	0.253
		(0.242)	(0.250)	(0.247)	(0.245)
30	5	0.276	0.227	0.208	0.210
		(0.198)	(0.201)	(0.197)	(0.194)
50	7	0.194	0.179	0.155	0.178
		(0.148)	(0.149)	(0.152)	(0.150)

Table 3 Root of mean squared error (and standard deviation) of estimators in estimate of entropy $H(f)$ for uniform distribution on (0,1)

n	m	RMSE (SD)			
		HV_{mn}	HVE_{mn}	HC_{mn}	HM_{mn}
5	2	0.774	0.407	0.566	0.363
		(0.346)	(0.407)	(0.336)	(0.342)
10	3	0.455	0.216	0.295	0.168
		(0.166)	(0.216)	(0.169)	(0.165)
15	4	0.343	0.155	0.208	0.134
		(0.110)	(0.155)	(0.112)	(0.110)
20	4	0.274	0.121	0.157	0.104
		(0.087)	(0.121)	(0.088)	(0.086)
30	5	0.210	0.086	0.110	0.091
		(0.059)	(0.086)	(0.061)	(0.060)
50	7	0.156	0.058	0.076	0.085
		(0.037)	(0.058)	(0.039)	(0.038)

From Table 1, we can see that for the normal distribution the proposed estimator HM_{mn} has a smaller RMSE than the other estimators. For the exponential distribution, we can see from Table 2 that the proposed estimator HM_{mn} has a smaller RMSE than the competitors for small sample sizes and for $n = 40, 50$ the Correa's estimator HC_{mn} has the smallest RMSE. For the uniform distribution, we can see from Table 3 that the proposed estimator HM_{mn} has the smallest RMSE for small sample sizes and for $n = 40, 50$ the Correa's estimator HVE_{mn} has the smallest RMSE.

Generally, it is evident from Tables 1 and 3 that the proposed estimator has an excellent performance in compared to its competitors. It can be seen that they have a smaller RMSE than the other estimators for small and moderate sample sizes.

Based on our simulations, we can conclude that the proposed estimator has an excellent performance for small and moderate sample sizes and therefore it can be easily applied in practice.

4. Conclusions

In this paper, we have first described some prominent estimators for Shannon entropy and then propose a new entropy estimator of a continuous random variable. The proposed estimator has been constructed based on modification of Correa entropy estimator. We have presented the properties of the proposed estimator. We finally have compared the proposed estimator with some prominent existing estimators. We have shown that for small and moderate sample sizes the new estimator behave better than the competitors. Generally, the proposed estimator has a good performance and it can be easily applied in practice.

Acknowledgements

The authors would like to thank the reviewers and editor for their critical review and comments on the earlier version of the manuscript.

References

Arizono I, Ohta H. A test for normality based on Kullback-Leibler information. Am Stat. 1989; 43: 20-22.

Balakrishnan N, Rad AH, Arghami NR. Testing exponentiality based on Kullback-Leibler information with progressively type-II censored data. IEEE T Reliab. 2007; 56(2): 301-307.

Correa JC. A new estimator of entropy. Commun Stat-Theor M. 1995; 24(10): 2439-2449.

Ebrahimi N. Testing exponentiality of the residual life, based on dynamic Kullback-Leibler information. IEEE T Reliab. 1998; 47(2): 197-201.

Esteban MD, Castellanos ME, Morales D, Vajda I. Monte Carlo comparison of four normality tests using different entropy estimates. Commun Stat-Simul C. 2001; 30(4): 761-785.

Grzegorzewski P, Wieczorkowski R. Entropy-based goodness-of-fit test for exponentiality. Commun Stat-Theor M. 1999; 28(5): 1183-1202.

Noughabi HA, Arghami NR. General treatment of goodness-of-fit tests based on Kullback-Leibler information. J Stat Comput Sim. 2013a; 83(8): 1556-1569.

Noughabi HA, Arghami NR. Goodness-of-fit tests based on correcting moments of entropy estimators. Commun Stat-Simul C. 2013b; 42(3): 499-513.

Noughabi HA, Balakrishnan N. Goodness of fit using a new estimate of Kullback-Leibler information based on Type II censored data. IEEE T Reliab. 2015; 64(2): 627-635.

Pakyari R, Balakrishnan N. A general purpose approximate goodness-of-fit test for progressively Type-II censored data. IEEE T Reliab. 2012; 61(1): 238-244.

Pakyari R, Balakrishnan N. Goodness-of-fit tests for progressively Type-II censored data from location-scale distributions. J Stat Comput Sim. 2013; 83(1): 167-178.

Rad AH, Yousefzadeh F, Balakrishnan N. Goodness-of-fit test based on Kullback–Leibler information for progressively Type-II censored data. IEEE T Reliab. 2011; 60(3): 570-579.

Shannon CE. A mathematical theory of communications. Bell Syst Tech J. 1948; 27: 379-423, 623-656.

Van Es B. Estimating functionals related to a density by class of statistics based on spacings. Scand J Stat. 1992; 19: 61-72.

Vasicek O. A test for normality based on sample entropy. J Roy Stat Soc B Met. 1976; 38: 54-59.