The Korean Communications
in Statistics Vol. 9, No. 1, 2002
pp. 13-19

Modified Mass-Preserving Sample Entropy

Chul Eung Kim[1] and Sang Un Park[2]

Abstract

In nonparametric entropy estimation, both mass and mean-preserving maximum entropy distribution (Theil, 1980) and the underlying distribution of the sample entropy (Vasicek, 1976), the most widely used entropy estimator, consist of n mass-preserving densities based on disjoint intervals of the simple averages of two adjacent order statistics. In this paper, we notice that those nonparametric density functions do not actually keep the mass-preserving constraint, and propose a modified sample entropy by considering the generalized O-statistics (Kaigh and Driscoll, 1987) in averaging two adjacent order statistics. We consider the proposed estimator in a goodness of fit test for normality and compare its performance with that of the sample entropy.

I. Introduction

Suppose that a random variable X has a distribution function $F(x)$ with a continuous density function $f(x)$. The differential entropy $H(F)$ of the random variable is defined by Shannon (1948) to be

$$H(F) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx. \tag{1}$$

The entropy difference between two distribution functions F and G is defined to be

$$\Delta H(F, G) = H(F) - H(G),$$

which is nonnegative if F and G are in the same moment class and F is the maximum entropy (ME) distribution in the class.

Suppose that we have an independently identically distributed (i.i.d.) sample, $X_1, X_2, \cdots, X_n$. We define $X_{(r,n)}$ to be rth order statistic from a sample of size n and suppress the notation n for a sample of size n. In a goodness of fit test for $H_0: f = f_0$ based on the sample, $\Delta H(F_0, G)$ has been considered in establishing a goodness of fit test statistic

[1] Corresponding author. Associate Professor, Department of Applied Statistics, Yonsei University, Shinchon Dong 134, Seoul, Korea
E-mail ; cekim@yonsei.ac.kr

[2] Associate Professor, Department of Applied Statistics, Yonsei University, Shinchon Dong 134, Seoul, Korea.
E-mail : sangun@bubble.yonsei.ac.kr

by estimating $H(G)$ and the possible unknown parameters in F_0 (see, Vasicek, 1976; Gokhale, 1983). In estimating $H(G)$, the sample entropy (Vasicek, 1976) has been most widely used, which can be written as

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

where $X_{(i)} = X_{(1)}$ for $i < 1$ and $X_{(i)} = X_{(n)}$ for $i > n$.

Theil (1980) has given a mass and mean-preserving ME distribution as

$$\begin{aligned} g_t(x) = \; & n^{-1} \frac{4}{x_{(2)} - x_{(1)}} \exp\left(\frac{x - \frac{1}{2}(x_{(1)} + x_{(2)})}{\frac{1}{4}(x_{(2)} - x_{(1)})} \right) && \text{if} \, x \leq \xi_2 \\ & n^{-1} \frac{2}{x_{(i+1)} - x_{(i-1)}} \qquad \text{if} \, \xi_i < x \leq \xi_{i+1}, \, i = 2, \cdots, n-1 && \\ & n^{-1} \frac{4}{x_{(n)} - x_{(n-1)}} \exp\left(- \frac{x - \frac{1}{2}(x_{(n-1)} + x_{(n)})}{\frac{1}{4}(x_{(n)} - x_{(n-1)})} \right) && \text{if} \, x > \xi_n \end{aligned} \qquad (2)$$

where $\xi_i = (x_{(i-1)} + x_{(i)})/2$, $i = 2, \cdots, n$, $\xi_1 = x_{(1)}$ and $\xi_{n+1} = x_{(n)}$. In g_t, the mass-preserving constraint, $\int_{\xi_i}^{\xi_{i+1}} g_t(x)\,dx = 1/n$, holds and the mean-preserving constraint, $E_{g_t}(X) = \bar{x}$, holds. However, we note that g_t in $(-\infty, \xi_1)$ has been established based on the mean-preserving constraint by neglecting the mass-preserving information in the interval $(x_{(1)}, \xi_1)$. It has been recently shown in Park and Park (2001) that the sample entropy for m=1 can be viewed as the entropy of the nonparametric density function,

$$\begin{aligned} g_v(x) = \; & 0 && \text{if} \, x < \xi_1 \text{or } \, x > \xi_{n+1} \\ & n^{-1} \frac{2}{x_{(i+1)} - x_{(i-1)}} && \text{if} \, \xi_i < x \leq \xi_{i+1}, \, i = 1, \cdots, n \end{aligned} \qquad (3)$$

Then we can see that g_t is just a modification of g_v in the end-intervals by taking the exponential smoothing to maximize the entropy under both end-intervals where the exponential distribution is fitted to satisfy the mean-preserving constraint. It has been shown in Theil (1980) that the entropy of H_t can be obtained as

$$H_t(n) = \frac{2}{n}(1 - \log 2) + H_v(n, 1).$$

In this paper,

we replace $\xi_i = (x_{(i-1)} + x_{(i)})/2$ in g_t with $\xi'_i = ((n - i + 1)x_{(i-1)} + (i-1)x_{(i)})/n$ to conform better to the mass-preserving constraint and also modify the exponential smoothing of both end-intervals by considering the fact that $F(X_{(1;n-1)}) - F(X_{(1;n)}) \approx 1/(n(n+1))$. Thus the proposed sample entropy is mass-preserving but not mean-preserving. We consider our modified sample entropy in a goodness of fit test and compare its performance with the test statistic based on the sample entropy of m=1 for a normal distribution. Monte Carlo simulation shows that our test statistic performs better that the competing test statistic against almost all alternatives except the uniform distribution.

2. Modified nonparametric entropy estimation

The nonparametric estimators of $H(F)$ have been suggested by many authors including Vasicek (1976), Theil (1980), Ebrahimi et al (1994) which are all based on gaps of order statistics. It has been shown in Cressie (1978) and Antille et al (1982) that test statistics based on gaps of order statistics perform well in a goodness of fit test and symmetric test, respectively. The test statistics based on the Kullback-Leibler information (Kullback and Leibler, 1951) and the entropy difference have been employing the sample entropy (Vasicek, 1976) as a nonparametric entropy estimator and show good performances in a goodness of fit test (see, Arizono and Ohta (1989), Ebrahimi et al (1992), etc). The g_v of the sample entropy and g_t of the mean and mass preserving maximum entropy (ME) distribution are based on second-order gaps of order statistics, where m-order gaps of order statistics is defined to be $x_{(i+m)}-x_{(i)}$.

We can see from (2) and (3) that g_v is actually based on n disjoint intervals composed of $(x_{(1)}, \xi_2, \cdots, \xi_n, x_{(n)})$ and g_t is based on n disjoint intervals composed of $(-\infty, \xi_2, \cdots, \xi_n, \infty)$. Their difference is that g_v is based on the bounded interval while g_t is based on the unbounded interval. In each interval, (ξ_i, ξ_{i+1}), $i=2, \cdots, n-1$, the mass-preserving constraint has its meaning if the mass-approximation holds as $\int_{\xi_i}^{\xi_{i+1}} f(x)dx \approx 1/n, i=2, \cdots, n-1$. In view of this, we think that the closer value to $x_{(i;n-1)}$ is better than the simple average in averaging $x_{(i)}$ and $x_{(i+1)}$. Kaigh and Driscoll (1987) suggested the O-statistics of order statistics as

$$m_{r,d} = \sum_{i=r}^{n+r-d} \frac{C_{j-1,r-1} C_{n-j,d-r}}{C_{n,d}} x_{(j;n)}$$

where $E(m_{r,d}) = \mu_{r,d}$ and $C_{n,d} = n!/((n-d)!d!)$. In view of Kaigh and Driscoll (1987), it is clear that $\xi_i' = ((n-i+1)x_{(i-1)} + (i-1)x_{(i)})/n$ is better fitting the mass-approximation than just the simple average. Thus our modified intervals are

$$A_n = (x_{(1)}, \xi_2', \cdots, \xi_n', x_{(n)}) \tag{4}$$

or

$$B_n = (-\infty, \xi_2', \cdots, \xi_n', \infty)$$

If we consider the fact that $F(X_{(1;n-1)})-F(X_{(1;n)})\approx 1/(n(n+1))$, we need to adjust the mass-preserving adjustment n for the interval $(x_{(1)},\xi_2^{'})$ in (4) with $n(n+1)$. However, the resulting g_v is no more probability density function. Ebrahimi et al (1994) also have suggested a modified sample entropy by adjusting the order of gaps of order statistics in both end-intervals, but their modified sample entropy is also not based on a probability density function in a similar context. Thus our interest here is on the modification of B_n and the exponential density in the interval $(-\infty,\xi_2^{'})$, which satisfies $F(X_{(1;n-1)})-F(X_{(1;n)})\approx 1/(n(n+1))$, can be determined to be

$$n^{-1}\frac{\log((n+1)/n)}{\xi_2^{'}-x_{(1)}}\exp(\frac{\log((n+1)/n)}{\xi_2^{'}-x_{(1)}}(x-\xi_2^{'})).$$

Thus the modified nonparametric density function can be written as

$$\begin{aligned} g_{mt}(x)= \;& n^{-1}\frac{\log((n+1)/n)}{\xi_2^{'}-x_{(1)}.}\exp(\frac{\log((n+1)/n)}{\mathrm{xi}_2^{'}-x_{(1)}}(x-\xi_2^{'})) && \text{if}\quad x\leq\xi_2^{'} \\ & n^{-1}\frac{1}{\xi_{i+1}^{'}-\xi_i^{'}} && \text{if}\quad \xi_i^{'}< x\leq\xi_{i+1}^{'},\, i=2,\cdots,n-1 \\ & n^{-1}\frac{\log((n+1)/n)}{x_{(n)}-\xi_n^{'}}\exp(-\frac{\log((n+1)/n)}{x_{(n)}-\xi_n^{'}}(x-\xi_n^{'})) && \text{if}\quad x>\xi_n^{'}. \end{aligned}$$

Then we can obtain the entropy of g_{mt} as

$$H_{mt}(n)=\frac{2}{n}(1-\log(\frac{n+1}{n}))+n^{-1}\sum_{i=1}^{n}\log(\xi_{i+1}^{'}-\xi_i^{'}).$$

where $\xi_1^{'}=x_{(1)}$ and $\xi_{n+1}^{'}=x_{(n)}$.

3. Test for normality

Suppose that we are interested in a goodness of fit test for,

$H_0: f_0(x;\mu,\sigma)=\exp(-(x-\mu)^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$, where μ and σ are unknown. The normal distribution is the ME distribution under the moment constrains $E(X)=\mu$ and $E(X^2)=\mu^2+\sigma^2$, and $\varDelta H(F_0,G)$ can be written as $\log\sqrt{2\pi\sigma^2}+0.5-H(G)$. Thus we can establish a test statistic by estimating μ,σ and $H(G)$. In estimating μ and σ^2, we use the method of moments estimation where the moments are the ME characterizing moments by following the lines of Soofi et al (1995) and Park and Park (2001) as $E_{G_{mt}}(X)$ and $E_{G_{mt}}(X^2)-E_{G_{mt}}(X)^2$, respectively. Thus we consider here two test statistics for normality as follow.

1. T_v (m=1 in Vasicek(1976) and Arizono and Ohta (1989)).

$$T_v = \log\sqrt{2\pi\widehat{\sigma_v^2}} + 0.5 - H_v(n,1),$$

where $\widehat{\sigma_v^2} = E_{G_v}(X^2) - E_{G_v}(X)^2$.

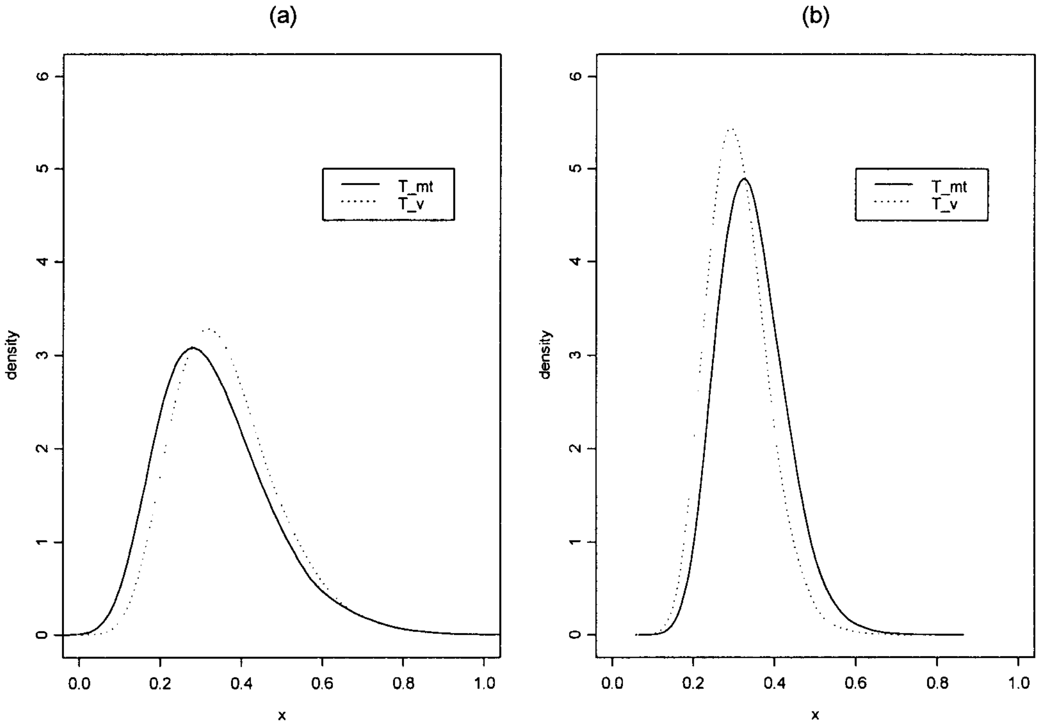
2. T_{mt}

$$T_{mt} = \log\sqrt{2\pi\widehat{\sigma_{mt}^2}} + 0.5 - H_{mt}(n),$$

where $\widehat{\sigma_{mt}^2} = E_{G_{mt}}(X^2) - E_{G_{mt}}(X)^2$.

We made 20,000 Monte Carlo simulations and obtained the percentage points of each statistic for the normal distribution. In Figure 1, we present the distributions of two statistic based on the Monte Carlo samples where the sample sizes are 20, 50.

Figure 3. Normal Null Distribution : Distributions of two test statistics based on 20,000 simulations (a) n=20 (b) n=50.



We also made 10,000 Monte Carlo simulations on the powers of the test statistics against 8 alternatives for sample sizes 20, 50. The critical values for T_v and T_{mt} are determined from the previous simulations to be 0.5959 and 0.5936 for n=20, and 0.4391 and 0.4888 for n=50, respectively. In this power comparison, we see from Table 1 that T_{mt} has much better powers than T_v against alternatives with unbounded supports and similar powers against other alternatives with bounded supports except the uniform alternative. The poor performance for the exponential distribution comes from the fact that the uniform distribution has a bounded support.

4. Concluding remarks

We modify the sample entropy, which is most widely used nonparametric entropy estimator, and see that its performance as a goodness of fit test statistic is better than that of Vasicek (1976) against alternatives with unbounded supports but is not better against alternatives with bounded supports. We may consider including the endpoints in the sample for the alternatives with bounded supports. In modifying the sample entropy, we have considered two adjustments concerning the mass-preserving constrains : The first one is that we use the weighted average, instead of the simple average, which produces closer value to $x_{(i;n-1)}$ and the second one is that we modify the exponential fitting in Theil (1980) to satisfy the mass approximation. Our modification here is limited to $m=1$, and it needs to be further generalized for $m>1$.

Table 1. Power estimate of .05 tests against eight alternatives of the normal distribution based on 10,000 simulations (%)

Alternatives	n=20		n=50	
	T_m	T_{mt}	T_v	T_{mt}
Uniform	29.26	15.41	64.25	42.45
Exponential	64.13	65.57	97.77	97.03
$\chi^2(1)$	97.09	96.75	100.00	100.00
$\chi^2(2)$	64.13	65.57	97.77	97.03
$\chi^2(4)$	26.13	30.66	61.91	61.49
Cauchy	68.06	81.75	97.18	98.79
t(3)	12.40	27.69	29.34	51.79
t(5)	6.46	15.22	10.82	25.99

References

[1] Antille, A., Kersting, G. and Zucchini, W. (1982). Testing symmetry. *Journal of the American Statistical Association*, 77, 639-646.

[2] Arizono, I and Ohta, H. (1989). A test for normality based on Kullback-Leibler information. *American Statistician*, 43, 20-22.

[3] Cressie, N. (1978). Power results for tests based on high-order gaps. *Biometrika*, 65, 214-218.

[4] Ebrahimi, N., Habibullah and Soofi, E. S. (1992). Testing Exponentiality Based on Kullback-Leibler Information. *Journal of the Royal Statistical Society*, 54, 739-748.

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

[6] Kaigh, W. D. and Driscoll, M. F. (1987). Numerical and Graphical Data Summary Using O-Statistics. *American Statistician*, 41, 25-32.

[7] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79-86.

[8] Park, S and Park, D. (2001). On the goodness of fit test based on the Kullback-Leibler information. *submitted.*

[9] Shannon, C. E. (1948). A mathematical theory of communications. *Bell System Technical Journal*, 27, 379-423; 623-656.

[10] Soofi, E. S., Ebrahimi, N., and Habibullah, M. (1995). Information distinguishability with application to analysis of failure data. *Journal of the American Statistical Association*, 90, 657-668.

[11] Theil, H. (1980). The entropy of the maximum entropy distribution. *Economics Letters*, 5, 145-148.

[12] Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society*, Ser. B, 38, 54-59.

[2001년 6월 접수 , 2001년 11월 채택]