A Goodness of Fit Tests Based on the Partial Kullback-Leibler Information with the Type II Censored Data

Sangun Park*and Jonggun Lim†

Abstract

Goodness of fit test statistics based on the information discrepancy have been shown to perform very well (Vasicek 1976, Dudewicz and van der Meulen 1981, Chandra et al 1982, Gohkale 1983, Arizona and Ohta 1989, Ebrahimi et al 1992, etc). Although the test is well defined for the non-censored case, censored case has not been discussed in the literature. Therefore we consider a goodness of fit test based on the partial Kullback-Leibler(KL) information with the type II censored data. We derive the partial KL information of the null distribution function and a nonparametric distribution function, and establish a goodness of fit test statistic. We consider the exponential and normal distributions and made Monte Calro simulations to compare the test statistics with some existing tests.

Key Words: Entropy difference, Maximum entropy distribution, Minimum discrimination information loss estimation, Order statistics, Sample entropy.

1 Introduction

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

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

We denote Ω_θ to be the moment class distributions, $\{f(x;\theta) : E_F(T_i(X)) = \theta_i, i = 1, 2, \cdots, k\}$. Then it is well-known that all of the well-known distribution are the maximum entropy (ME) distributions in the appropriate moment class Ω_θ (Soofi et al. 1995). The entropy difference is defined to be

$$\Delta H(f, g) = H(f) - H(g),$$

*Associate professor, Department of Applied Statistics, Yonsei University.
†Ph.D. Student, Department of Applied Statistics, Yonsei University.

which is nonnegative if $f(x)$ and $g(x)$ are in the same moment class and $f(x)$ is the ME distribution in the class. The Kullback-Leibler information is defined by Kullback-Leibler (1951) as

$$KL(g,f) = \int_{-\infty}^{\infty} g(x) \log \frac{g(x)}{f(x)} dx. \tag{2}$$

It is well known that $KL(g,f) \geq 0$ and the equality holds for $f(x) = g(x)$. It is shown in Soofi et al. (1995) that $\Delta H(f,g) = KL(g,f)$ if $g(x)$ is the same moment class with the ME distribution $f(x)$. In the statistical test of goodness of fit, the entropy difference, $\Delta H(f,g)$, and the KL information, $KL(g,f)$, have drawn much attention because of their nonnegativeness along with the Kolmogorov-Smirnov and Cramér-von Mises distances (Vasicek 1976, Dudewicz and van der Meulen 1981, Chandra et al 1982, Gohkale 1983, Arizona and Ohta 1989, Ebrahimi et al 1992, etc).

Let $X_{(1:n)}, X_{(2:n)}, \cdots, X_{(n:n)}$ be the order statistics of an independently identically distributed (i.i.d) sample of size n from $f(x)$. Suppose that we are interested in a goodness of fit test for an ME distribution $f_0(x;\theta)$ with the type II right censored data $X_{(1:n)}, X_{(2:n)}, \cdots, X_{(r:n)}$ for $r \leq n$. Thus we consider the censored KL information as

$$KL(g_n, f_0 : c) = \int_{-\infty}^{c} g_n(x) \log \frac{g_n(x)}{f_0(x)} dx. \tag{3}$$

In non-censored case, it is well known that the KL information $KL(g_n, f_0 : \infty) \geq 0$, and the equality holds for $f_0(x) = g_n(x)$. However the censored KL information does not satisfy the nonnegativity any more. So we consider for the first time the partial KL information as

$$KL^*(g_n, f_0 : c) = \int_{-\infty}^{c} g_n(x) \log \frac{g_n(x)}{f_0(x)} dx + F_0(c) - G_n(c) \tag{4}$$

where $dG_n(x) = g_n(x)$, $dF_0(x) = f_0(x)$ and $-\infty < c < \infty$. We consider some well-known distribution, and compare the performance of the test statistic with those of some existing test statistics. We consider the exponential and normal null distribution.

2 Test Statistic

The nonparametric estimation of $H(f)$ have been discussed by many authors including Vasicek (1976), Theil (1980), Dudewicz and van der Meulen (1987), Bowman (1992), Ebrahimi et al (1994) and Park and Park (2003). Among these various entropy estimators, Vasicek's sample entropy has been most widely used in developing entropy-based statistical procedures (Dudewicz and van der Meulen 1981, Gohkale 1983, Arizona and Ohta 1989, Ebrahimi et al 1992, etc). Vasicek (1976)'s estimator is then given by

$$H(n,m) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \log \frac{n}{2m} \left(x_{(i+m:n)} - x_{(i-m:n)} \right) \right\}, \tag{5}$$

where the window width m is positive integer smaller than $n/2$, $x_{(1:n)}, x_{(2:n)}, \cdots, x_{(n:n)}$ denote the order statistics of the sample, and $x_{(i:n)} = x_{(1:n)}$ for $i < 1$ and $x_{(i:n)} = x_{(n:n)}$ for $i > n$.

Suppose that we have only a part of sample $x_{(1:n)}, x_{(2:n)}, \cdots, x_{(r:n)}$ and are interested in a goodness of fit test $H_0 : f = f_0(x;\theta)$ vs $H_a : f \neq f_0(x;\theta)$ where $f_0(x;\theta)$ is an ME distributions. In the type II right censored sample $x_{(1:n)}, x_{(2:n)}, \cdots, x_{(r:n)}$ for $x \leq x_{(r:n)}$, it is difficult to establish a nonparametric distribution function for $x > x_{(r:n)}$. Although it may be possible to establish a nonparametric distribution function for $x > x_{(r:n)}$, it is not reasonable to consider the information discrepancy for $x > x_{(r:n)}$. To overcome this difficulty, it is assumed that a nonparametric distribution function G_n to be F_0 for $x > x_{(r:n)}$. Thus the nonparametric density function can be obtained in view of Park and Park (2003) as

$$g_n(x) = \begin{cases} 0 & \text{if } x < \xi_1 \\ n^{-1}\frac{2m}{x_{(i+m:n)} - x_{(i-m:n)}} & \text{if } \xi_i < x \leq \xi_{i+1}, i = 1, \cdots, r \\ f_0(x) & \text{if } x > \xi_{r+1} \end{cases} \tag{6}$$

where $\xi_i = (x_{(i-m:n)} + \cdots + x_{(i+m-1:n)})/2m$, and $x_{(i:n)} = x_{(1:n)}$ for $i < 1$ and $x_{(i:n)} = x_{(r:n)}$ for $i > r$. For this nonparametric density function, the partial KL information can be established as

$$\begin{aligned} KL^*(g_n, f_0 : c) &= \int_{-\infty}^{c} g_n(x) \log \frac{g_n(x)}{f_0(x)} dx + F_0(c) - G_n(c) \\ &= \int_{\xi_1}^{\xi_{r+1}} g_n(x) \log \frac{g_n(x)}{f_0(x)} dx + F_0(\xi_{r+1}) - \frac{r}{n}. \end{aligned} \tag{7}$$

Thus the test statistic based on the partial KL information can be written as

$$\begin{aligned} T(n,m,r) &= -H(n,m,r) - \int_{\xi_1}^{\xi_{r+1}} g_n(x) \log f_0(x;\hat{\theta}) dx \\ &\quad + F_0(\xi_{r+1};\hat{\theta}) - \frac{r}{n} \end{aligned} \tag{8}$$

where the estimate $-\int_{\xi_1}^{\xi_{r+1}} g_n(x) \log g_n(x) dx$ as $(1/n)\sum_{i=1}^{r}\{\log(n/2m)(x_{(i+m:n)} - x_{(i-m:n)})\}$ and $\hat{\theta}$ is an estimator of θ. It is natural to estimate θ so that the partial KL information is minimized where such an estimator is called the minimum discriminant information loss (MDI) estimate (Soofi, 2000).

$$\hat{\theta}_{MDI} = \arg\min_{\theta} KL^*(g_n, f_0 : \xi_{r+1}) \tag{9}$$

Under the null hypothesis, $T(n,m,r)$ will be close to 0.

3 Numerical Examples

3.1 Test for Exponentiality

Suppose that we are interested in a goodness of fit test for $H_0 : f_0(x;\theta) = \exp(-x/\theta)/\theta$ vs. $H_A : f_0(x;\theta) \neq \exp(-x/\theta)/\theta$ where θ is unknown. Then the partial KL information can be

Table 1: Power estimate of 0.1 tests against of exponential distribution based on 10,000 simulations when $n = 30, r = 20$

Types	Alternative	T_{MDI}	T_{MLE}	T_P	z	Z	W
Type I error	Exp(1)	10.04	10.08	9.10	10.33	9.65	9.73
Monotone decreasing hazard	Gamma(0.5)	22.83	25.66	29.03	58.36	53.85	49.88
	Weibull(0.5)	45.92	51.03	56.31	82.21	79.10	74.47
	Weibull(0.8)	5.89	6.38	6.83	22.36	18.81	19.70
	Chi-square(1)	23.34	26.10	29.44	58.50	53.64	49.89
Monotone increasing hazard	Uniform	35.23	35.71	32.07	32.08	25.10	34.63
	Gamma(1.5)	40.17	40.41	37.75	19.78	18.06	27.81
	Gamma(2)	71.42	71.82	69.55	34.07	30.63	55.03
	Weibull(2)	94.08	94.33	93.13	63.32	55.72	89.09
	Chi-square(3)	38.77	39.20	36.39	19.13	17.91	26.24
	Chi-square(4)	72.22	72.74	70.64	34.38	30.59	55.86
	Beta(1 and 2)	19.35	19.56	17.07	15.29	12.70	15.90
	Beta(2 and 1)	99.45	99.48	99.29	88.74	81.43	99.55
None-monotone hazard	Lognormal(0.6)	99.82	99.82	99.77	41.38	38.97	91.49
	Lognormal(1.0)	55.28	54.92	53.62	12.44	14.84	22.63
	Lognormal(1.2)	26.77	26.60	26.53	13.37	14.76	13.79
	Beta (0.5 and 1)	16.65	17.56	18.01	36.03	37.28	28.35

written as

$$KL^*(g_n, f_0 : \xi_{r+1}) = \int_{\xi_1}^{\xi_{r+1}} g_n(x) \log g_n(x) dx + \frac{r}{n} \log \theta + \frac{1}{\theta} \int_{\xi_1}^{\xi_{r+1}} x g_n(x) dx + F_0(\xi_{r+1}) - \frac{r}{n}.$$

where θ need to be estimated with different methods, then we have the test statistics considered here are as follow.

1. $T_{MDI}(n, m, r)$: Based on MDI estimator $(\hat{\theta}_{MDI})$
2. $T_{MLE}(n, m, r)$: Based on MLE estimator $(\hat{\theta}_{MLE} = (\sum_{i=1}^{r} x_{(i:n)} + (n-r)x_{(r:n)})/r)$

We made 10,000 Monte Carlo simulations for $n = 30$ to estimate the powers of our proposed test statistic and the competing test statistics. The simulation results are summarized in Table 1. We can see from the Tables that any test statistic does not beat others against all alternatives, but it is notable that the proposed test statistic shows better powers than the competing test statistics against the alternatives with monotone increasing hazard functions(see also Park 2003).

3.2 Test for Normality

Suppose that we are interested in a goodness of fit test for $H_0 : f_0(x; \mu, \sigma^2) = \exp(-(x - \mu)^2/2\sigma^2)/(\sqrt{2\pi\sigma^2})$ vs. $H_A : f_A(x; \mu, \sigma^2) \neq \exp(-(x - \mu)^2/2\sigma^2)/(\sqrt{2\pi\sigma^2})$ where $\theta = (\mu, \sigma^2)$

Table 2: Power estimate of 0.1 tests against of normal distribution based on 10,000 simulations when $n = 30, r = 20$

Alternative	T_{MDI}	T_{BLUE}	T_{1n}	T_{2n}	T_{3n}	${}_rW^2$	${}_rA^2$	S
Type I error	10.22	10.14	9.98	8.98	8.68	10.71	10.60	10.39
Tukey(1.5)	62.65	65.41	47.59	51.16	34.30	38.69	46.27	38.27
Tukey(3.0)	30.34	31.80	17.95	16.77	11.58	16.00	18.87	11.79
Tukey(5.0)	25.73	24.90	17.06	11.93	20.34	32.45	31.32	21.76
Uniform	55.71	58.32	41.25	43.31	28.36	33.77	40.14	31.26
Weibull(2.0)	25.82	28.23	24.13	16.16	15.37	20.71	22.32	13.41
Exp(1.0)	85.14	87.98	83.50	83.60	71.26	71.83	78.46	72.31

is unknown. Then the partial KL information can be written as

$$KL^*(g_n, f_0 : \xi_{r+1}) = \int_{\xi_1}^{\xi_{r+1}} g_n(x) \log g_n(x) dx + \frac{r}{n} \log \sqrt{2\pi\sigma^2} + \frac{1}{2\sigma^2} \int_{\xi_1}^{\xi_{r+1}} (x-\mu)^2 g_n(x) dx + F_0(\xi_{r+1}) - \frac{r}{n}.$$

where θ need to be estimated with different methods, then we have the test statistics considered here are as follow.

1. $T_{MDI}(n, m, r)$: Based on MDI estimator $(\hat{\mu}_{MDI}, \hat{\sigma}^2_{MDI})$
2. $T_{BLUE}(n, m, r)$: Based on BLUE estimator $(\hat{\mu}_{BLUE}, \hat{\sigma}^2_{BLUE})$

We made 10,000 Monte Carlo simulations for $n = 30$ to estimate the powers of our proposed test statistic and the competing test statistics. The simulation results are summarized in Table 2. We can see from the Table that any test statistic does not beat others against all alternatives, but it is notable that the proposed test statistic shows better powers except for Tukey(5.0) distribution.

References

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

Bowman, A. W. (1992). Density based tests for goodness-of-fit. *Journal of Statistical Computation and Simulation*, **40**, 1–13.

Chandra, M., De Wet, T. and Sameniago, F. J. (1982). On the sample redundancy and a test for exponentiality. *Communication in Statistics, Part A - Theory and Methods*, **11**, 429–438.

Dudewicz, E. J. and E. C. Van Der Meulen (1981). Entorpy-based tests of uniformity. *Journal of the American Statistical Association*, **76**, 967–974.

Dudewicz, E. J. and E. C. Van Der Meulen (1987). Emiric entropy, a new approach to nonparametric entropy estimation. *New Perspectives in Theoretical and Applied Statistics*, Puir, M. J., Vilaplana, J. P., and Wrtz, W. (eds.), 202–207, New York : Wiley

Ebrahimi, N., Habibullah, M. and Soofi, E. S. (1992). Testing exponentiality based on Kullback-Leibler information. *Journal of the Royal Statistical Society*, **54**, 739–748.

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

Gokhale, D. V. (1983). On entropy-based goodness-of-fit tests. *Computational Statistics and Data Analysis*, **1**, 157–165.

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

Park, S. (2003). Testing exponentiality based on the Kullback-Leibler information. *To appear in IEEE Transactions on Reliability.*

Park, S. and Park, D. (2003). Correcting moments for goodness of fit tests based on two entropy estimates. *Journal of Statistical Computation and Simulation*, **73**, 685–694.

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

Soofi, E. S. (2000). Principal information theoretic approaches. *Journal of the American Statistical Association*, **95**, 1349–1353.

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.

Vasicek, O. (1976). A Test for normality based on sample entropy. *Journal of the Royal Statistical Society, Ser. B*, **38**, 54–59.