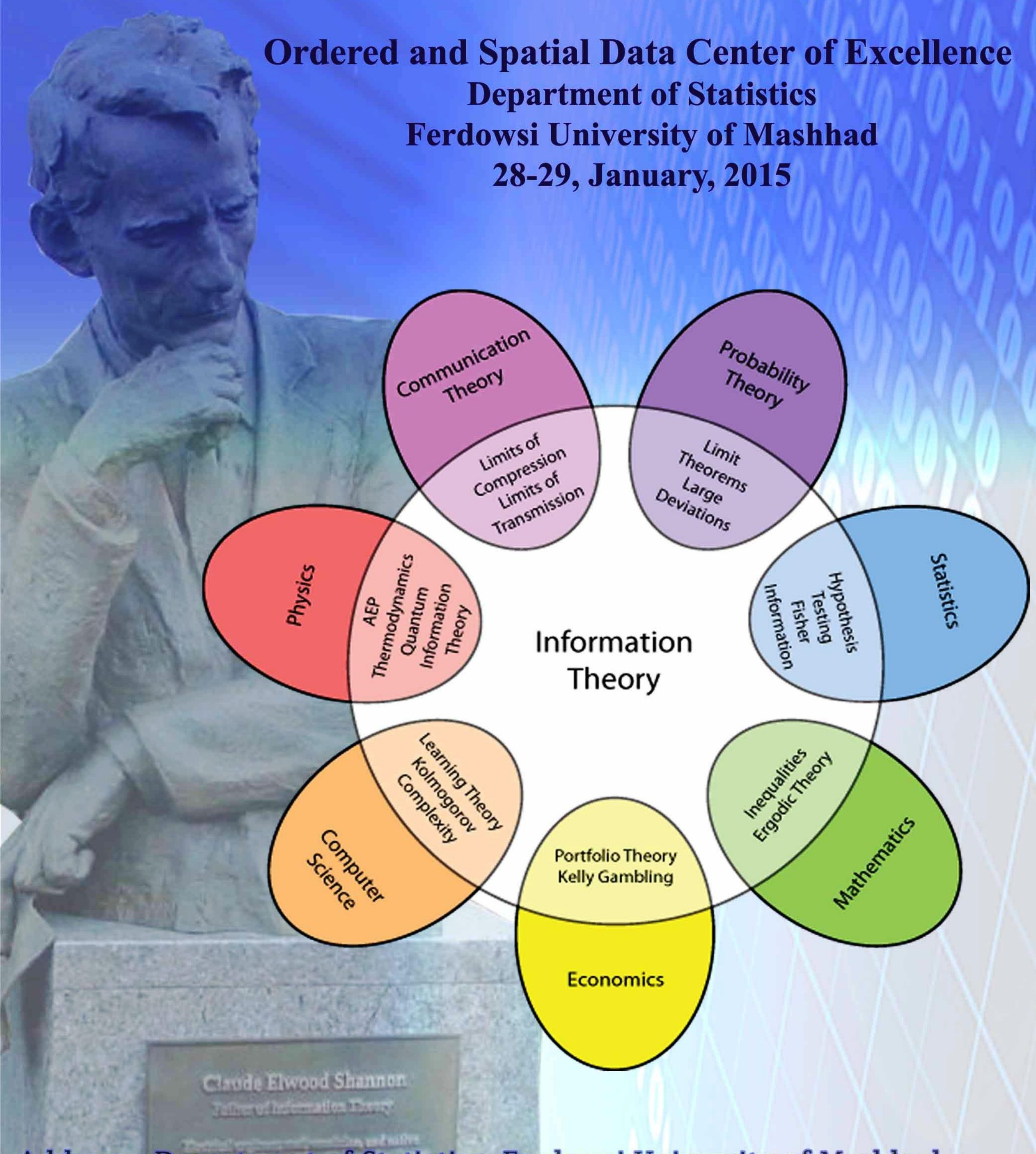
The 2nd Workshop on

Information Measures and Their Applications

Ordered and Spatial Data Center of Excellence
Department of Statistics
Ferdowsi University of Mashhad
28-29, January, 2015



Communication Theory

Limits of Compression Limits of Transmission

Probability Theory

Limit Theorems Large Deviations

Statistics

Hypothesis Testing Fisher Information

Mathematics

Inequalities Ergodic Theory

Economics

Portfolio Theory Kelly Gambling

Computer Science

Learning Theory Kolmogorov Complexity

Physics

AEP Thermodynamics Quantum Information Theory

Information Theory

Address : Department of Statistics, Ferdowsi University of Mashhad,
P.O. Box 1159-91775, Mashhad, Iran. Tel. and Fax : 051-38828605
Website : http://osdce.um.ac.ir E-Mail : osdce@um.ac.ir

In The Name of Allah

Proceedings of

The 2nd Workshop on Information Measures and Their Applications

Ordered and Spatial Data Center of Excellence

Department of Statistics
Faculty of Mathematical Sciences
Ferdowsi University of Mashhad

Mashhad, Iran

28-29 January, 2015

Scientific Committee:

Mohtashami Borzadaran, G. R. (Chairman)	Ferdowsi University of Mashhad
Abed Hodtani, G.	Ferdowsi University of Mashhad
Asadi, M.	University of Isfahan
Khaledi, B.	Razi University
Longobardi, M.	University of Naples Federico II

Organizing Committee:

Jabbari Nooghabi, H.	Ferdowsi University of Mashhad
Ahmadi, J.	Ferdowsi University of Mashhad
Emadi, M.	Ferdowsi University of Mashhad
Fashandi, M.	Ferdowsi University of Mashhad

Student Committee:

Mohtashami Borzadaran, H. A.	Ferdowsi University of Mashhad
Pakdaman, Z.	Ferdowsi University of Mashhad
Safaei, F.	Ferdowsi University of Mashhad

Preface

On behalf of the organizing and scientific committees, we would like to extend a very warm welcome to all the participants of the 2nd workshop on Information Measures and Their Applications. We hope this wokshop provides an environment of useful discussions and would also exchange scientific ideas through opinions.

We wish to express our gratitude to the numerous individuals and organizations that have contributed to the success of this workshop, in which more than 60 colleagues, researchers, and postgraduate students have participated.

Finally, we would like to extend our sincere gratitude to the students of the Department of Statistics at Ferdowsi University of Mashhad for their kind cooperation. We wish them all the best.

Ordered and Spatial Data
Center of Excellence

Contents

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

Entropy of some integer valued stochastic processes

Aghababaei Jazi, M.

Department of Statistics, University of Sistan and Baluchestan, Zahedan, Iran
Email: aghababaei@math.usb.ac.ir

Abstract

In this paper, we consider the entropy and entropy rate of some integer valued stochastic processes. We illustrate and compare the entropy (rate) of thinned values of Poisson and geometric processes. Also, we study the entropy (rate) of integer valued AR(1) process with Poisson innovations.

Keywords: Entropy, Integer valued stochastic processes, Markov chain.

1 Introduction

A stochastic process $X_1, X_2, ...$ is an indexed sequence of random variables and is characterized by the joint probability distribution function

$$\begin{aligned} Pr \quad & (X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \\ & = \prod_{k=1}^{n} Pr(X_k = x_k | X_1 = x_1, \cdots, X_{k-1} = x_{k-1}), \;\; n = 1, 2, ... \end{aligned} \tag{1}$$

This process is said to be stationary if the joint probability distribution function (1) of any subset of the sequence of random variables is invariant with respect to shifts in the time index, i.e.,

$$Pr(X_1 = x_1, \cdots, X_n = x_n) = Pr(X_{1+k} = x_1, \cdots, X_{n+k} = x_n),$$

for every shift k and for all $x_1, x_2, ..., x_n$.
$X_1, X_2, ...$ are said to be a Markov chain (process) if every random variable (r.v.) X_n depends on the one preceding it (X_{n-1}) and is conditionally independent of all the other preceding random variables $X_1, X_2, ..., X_{n-2}$, i.e., if

$$Pr(X_n = x_n | X_1 = x_1, \cdots, X_{n-1} = x_{n-1}) = Pr(X_n = x_n | X_{n-1} = x_{n-1}),$$

for all $x_1, x_2, ..., x_n$ and $n = 2, 3, ...$ In this case, the joint probability distribution function (pdf) of the r.v.'s $X_1, X_2, ..., X_n$ can be written as

$$Pr(X_1 = x_1, \cdots, X_n = x_n) = \prod_{k=1}^{n} Pr(X_k = x_k | X_{k-1} = x_{k-1}).$$

If $X_1, X_2, ...$ are Markov chain and $P_{ij} \equiv Pr(X_n = j | X_{n-1} = i)$ does not depend on n, i.e.,

$$Pr(X_n = j | X_{n-1} = i) = Pr(X_2 = j | X_1 = i), \quad n = 2, 3, ...$$

then the chain is said to be time invariant (or homogeneous). In this paper, Markov chains are assumed to be time invariant, so that, they can be characterized by its initial state and a probability transition matrix $P = [P_{ij}]$. Also, the pdf of the process at time n, i.e., $\pi_i^{(n)} \equiv Pr(X_n = i)$ is characterized by the pdf at time $n-1$ and P_{ij} such that

$$\begin{aligned} \pi_i^{(n)} &= \sum_i Pr(X_n = j, X_{n-1} = i) \\ &= \sum_i Pr(X_{n-1} = i).Pr(X_n = j | X_{n-1} = i) \\ &= \sum_i Pr(X_{n-1} = i) P_{ij}. \end{aligned}$$

If the pdf $\pi_i^{(n)}$ does not depend on n, i.e.

$$\pi_i^{(n)} = \pi_i^{(1)}, \quad \forall n = 2, 3, ...,$$

then the pdf is said to be stationary. In this case, the stationary pdf $\pi_i \equiv Pr(X_n = j)$ is the solution to the equation system

$$\pi_j = \sum_i \pi_i P_{ij}$$

The entropy of r.v. X with pdf $f(.)$ has been defined to be $H(X) = -E(\log f(X))$. It implies that the entropy of r.v.'s $X_1, X_2, ..., X_n$ with joint pdf (1.1) is given by the following chain rule

$$H(X_1, X_2, \cdots, X_n) = \sum_{k=1}^{n} H(X_k | X_1, X_2, \cdots, X_{k-1}). \tag{2}$$

Specially, the entropy of a sequence of independent r.v.'s $X_1, X_2, ..., X_n$ is $\sum_{i=1}^n H(X_i)$, where $H(X_i)$ is the entropy of r.v. X_i, $i = 1, 2, ...$ Consequently, the entropy of iid r.v.'s $X_1, X_2, ..., X_n$ is $nH(X_1)$. This implies that the average length of a code can increase linearly with the sequence length (n), and the slope (rate) of $H(X_1)$.

The entropy rate of a stochastic process $X_1, X_2, ...$ is defined by

$$H(\chi) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n),$$

when the limit exists. The rate obviously is constant $H(X_1)$ for iid r.v.'s $X_1, X_2, ...$. Also, the entropy rate of independent r.v.'s $X_1, X_2, ...$ given by

$$H(\chi) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i)$$

may or may not exist.

The entropy rate has also been defined by

$$H'(\chi) = \lim_{n\to\infty} H(X_n | X_1, X_2, \cdots, X_{n-1}),$$

when the limit exists. This rate is actually the entropy of the most recent output given all past outputs. Obviously, $H(X\chi) = H'(\chi)$ for a sequence of iid r.v.'s. This equality also holds for stationary stochastic processes. To see this, let $X_1, X_2, ...$ be an stationary stochastic process. Then,

$$\begin{aligned} H(X_{n+1} | X_1, X_2, \cdots, X_n) &\leq H(X_{n+1} | X_2, X_3, \cdots, X_n) \\ &= H(X_n | X_1, X_2, \cdots, X_{n-1}), \end{aligned}$$

where the inequality follows from the fact that conditioning reduces entropy and the equality follows from the stationarity of the process. Hence, $H(X_n|X_1, X_2, ..., X_{n-1})$ is a decreasing sequence in n and, consequently,

$$H'(\chi) = \lim_{n\to\infty} H(X_n|X_1, X_2, \cdots, X_{n-1})$$

exists. Also, from the chain rule (1.2), we obtain

$$H(\chi) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(X_i) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_1, X_2, \cdots, X_{i-1})$$

where, by Cesaro mean, we have

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_1, X_2, \cdots, X_{i-1}) &= \lim_{n\to\infty} H(X_n|X_1, X_2, \cdots, X_{n-1}) \\ &= H'(\chi). \end{aligned}$$

The above equalities implies that $H(\chi) = H'(\chi)$ for stationary stochastic processes. Specially, the entropy rate of a stationary Markov chain $X_1, X_2, ...$ is given by

$$\begin{aligned} H(\chi) &= H'(\chi) \\ &= \lim_{n\to\infty} H(X_n|X_1, X_2, \cdots, X_{n-1}) \\ &= \lim_{n\to\infty} H(X_n|X_{n-1}) \\ &= H(X_2|X_1). \end{aligned}$$

Hence, for the stationary Markov chain with transition matrix $P = [P_{ij}]$ and $X_1 \sim \pi_i$, the entropy rate of the Markov chain is then given by

$$\begin{aligned} H(\chi) &= H(X_2|X_1) \\ &= -\sum_{ij} \pi_i P_{ij} \log P_{ij} \\ &= \sum_{i} \pi_i (-\sum_{j} P_{ij} \log P_{ij}), \end{aligned} \tag{3}$$

i.e., a weighted sum of the entropy values for each state. [2]

2 Main results

Let X be a nonnegative integer valued r.v. with pdf $\pi_i = Pr(X = i)$, $i = 0, 1, ...$. According to Steutel and van Harn (1979), the binomial thinning operator $\circ$ is defined as:

$$\alpha \circ X = \sum_{i=1}^{X} B_i(\alpha),$$

where counting series $B_i(\alpha)$ is a sequence of iid binary r.v.'s with $P(B_i(\alpha) = 1) = 1 - P(B_i(\alpha) = 0) = \alpha$ and $\alpha \in [0, 1]$. It can be easily shown that the pdf of $\alpha \circ X$ is given by

$$\begin{aligned} \varphi_k \equiv Pr(\alpha \circ X = k) &= \sum_{i} Pr(X = i) Pr(\alpha \circ X = k|X = i) \\ &= \sum_{i} \pi_i \left(\begin{array}{c} i \\ k \end{array} \right) \alpha^k (1 - \alpha)^{i-k}, \end{aligned}$$

i.e., a mixture of binomial distributions Bin(i, α) with mixing pdf π_i. (see e.g., [1], [3] and [4])

This implies that if $X_1, X_2, ...$ is an integer valued stochastic process with stationary pdf $\pi_i \equiv Pr(X_n = i)$, $n = 1, 2, ...$, then the stationary pdf of thinned process $\alpha \circ X_1, \alpha \circ X_2, ...$ is φ_k and the entropy of $\alpha \circ X_n$ is a weighted sum of the binomial entropy given by

$$\begin{aligned} H(\alpha \circ X_n) &= -E(\log \varphi_{\alpha \circ X_n}) \\ &= -\sum_k \varphi_k \log \varphi_k \\ &= \sum_i \pi_i H(S_i), \end{aligned} \tag{4}$$

where $S_i \sim \text{Bin}(i, \alpha)$, $i = 1, 2, ...$. Furthermore, the entropy rate of stationary integer valued thinned process $\alpha \circ X_1, \alpha \circ X_2, ...$ is given by

$$\begin{aligned} H(\alpha \circ \chi) &= H'(\alpha \circ \chi) \\ &= \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} H(\alpha \circ X_i). \end{aligned} \tag{5}$$

The entropy of $\alpha \circ X_n$'s given by (2.1) have been shown in Figure 1 for $\alpha = 0.01, 0.02, ..., 1$ and $X_n \sim \pi_i$ is Poisson(λ) and Geometric(p) distributed r.v. with $\lambda = 1, 50$ and $p = 0.1, 0.9$. We see two different result for small and large values of λ and p. For the large value of λ we see more entropy, and with both small and large value of λ we see the maximum entropy for $\alpha = 0.5$. In contrast, for the small value of p we see more entropy, but again with both small and large value of p we see the maximum entropy for $\alpha = 0.5$.

The entropy rate of $\alpha \circ X_n$'s given by (2.2) have been shown in Figure 2 for $\alpha = 0.1, 0.9$ and through $n = 1000$ simulated r.v.'s from Poisson(λ) and Geometric(p) with $\lambda = 1, 2, ..., 100$ and $p = 0.01, 0.02, ..., 1$. Again we see two different entropy rates for small and large values of λ and p. For both small and large values of α, thinned Poisson(λ) processes have the same increasing concave entropy rate. In contrast, for both small and large values of α, thinned geometric(p) processes have the same decreasing convex entropy rate.

The integer valued first-order autoregressive (INAR(1)) process was independently introduced by McKenzie (1985) and Al-Osh and Alzaid (1987) based on the binomial thinning $\circ$ operator. Integer valued stochastic process $X_1, X_2, ...$ follows INAR(1) model, if

$$X_n = \alpha \circ X_{n-1} + \varepsilon_n, \quad n = 2, 3, ..., \tag{6}$$

where ε_n's known as innovations are iid nonnegative integer valued r.v.'s and independent of all the binary iid r.v.'s $B_i(\alpha)$ and X_{n-1}. If X_n is the population at time n then $\alpha \circ X_{n-1}$ and ε_n can be respectively interpreted as the number of survivors and the number of immigrants at time n from the previous period. It has been shown that an stationary INAR(1) process (2.3) has a Poisson marginal distribution (with mean λ) if and only if the innovations also follow a Poisson distribution (with mean $\lambda(1-\alpha)$). (see, e.g., [1] and [3])

Let $X_1, X_2, ...$ follow a INAR(1) process (2.3) wherein innovations ε_n's follow stationary pdf $q_k \equiv Pr(\varepsilon_n = k)$, $k = 0, 1, ...$ Then, the process forms a stationary Markov chain with transition probabilities

$$\begin{aligned} p_{ij} &\equiv P(X_n = j | X_{n-1} = i) \\ &= \sum_{k=0}^{\min(i,j)} \left(\begin{array}{c} i \\ k \end{array} \right) \alpha^k (1-\alpha)^{i-k} q_{j-k}, \quad i, j = 0, 1, \cdots , \end{aligned}$$

giving the probability of going from state i to state j in a single step Hence, the entropy rate is given by (1.3), wherein $\pi_i \equiv Pr(X_n = i)$ is the stationary pdf of X_n.

The entropy rate (1.3) of INAR(1) process (with $\pi_i \equiv$ Poisson(λ) marginal distribution) has been shown in Figure 3 for $\lambda = 1, 2, ..., 100$ and $\alpha = 0.01, 0.02, ..., 1$. The entropy rate is non-monotone and concave in λ and decreasing concave in α. For small values of α we have more entropy rate. However, for both small and large values of α we have maximum entropy for $\lambda \approx 80$.

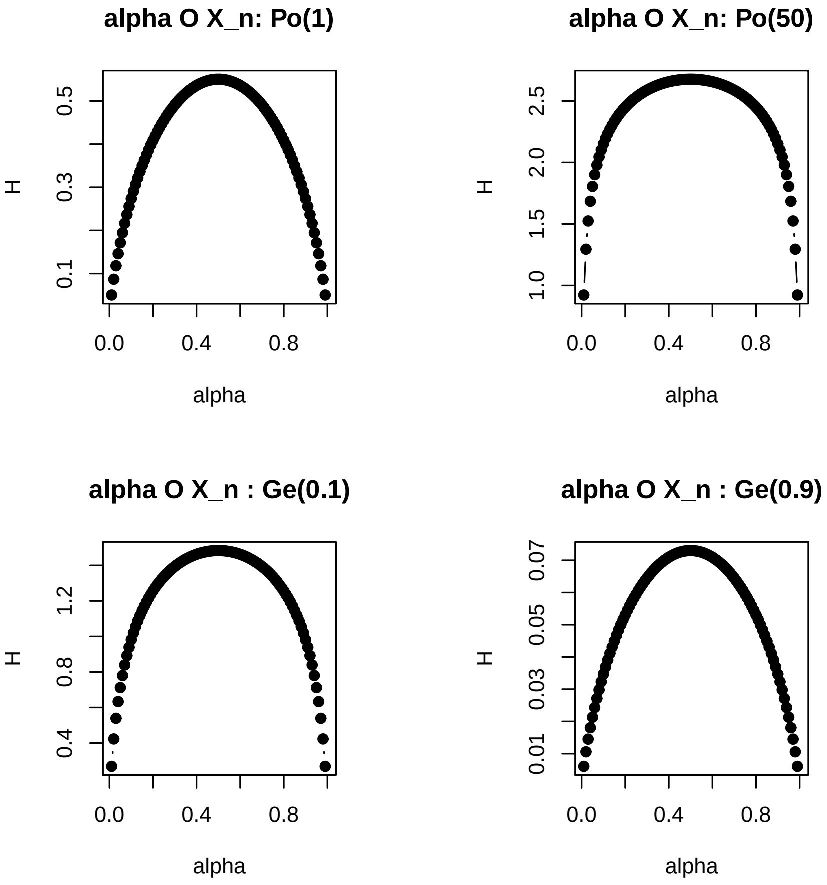


Figure 1: The entropy of thinned process $\alpha \circ X_n$.

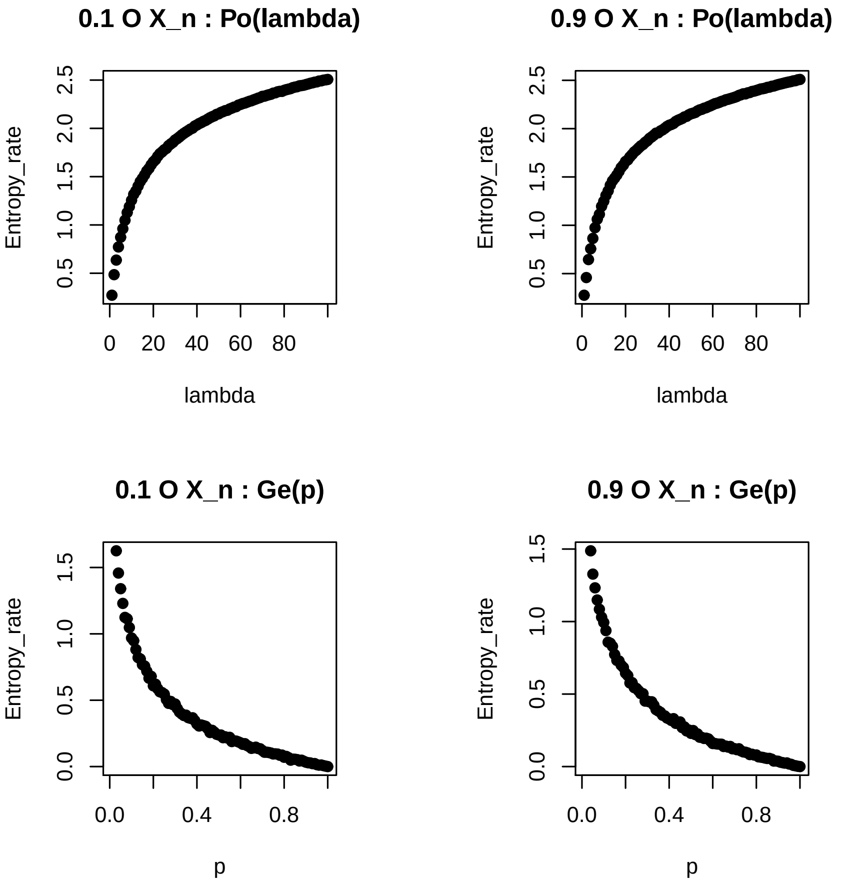


Figure 2: The entropy rate of thinned process $\alpha \circ X_n$.

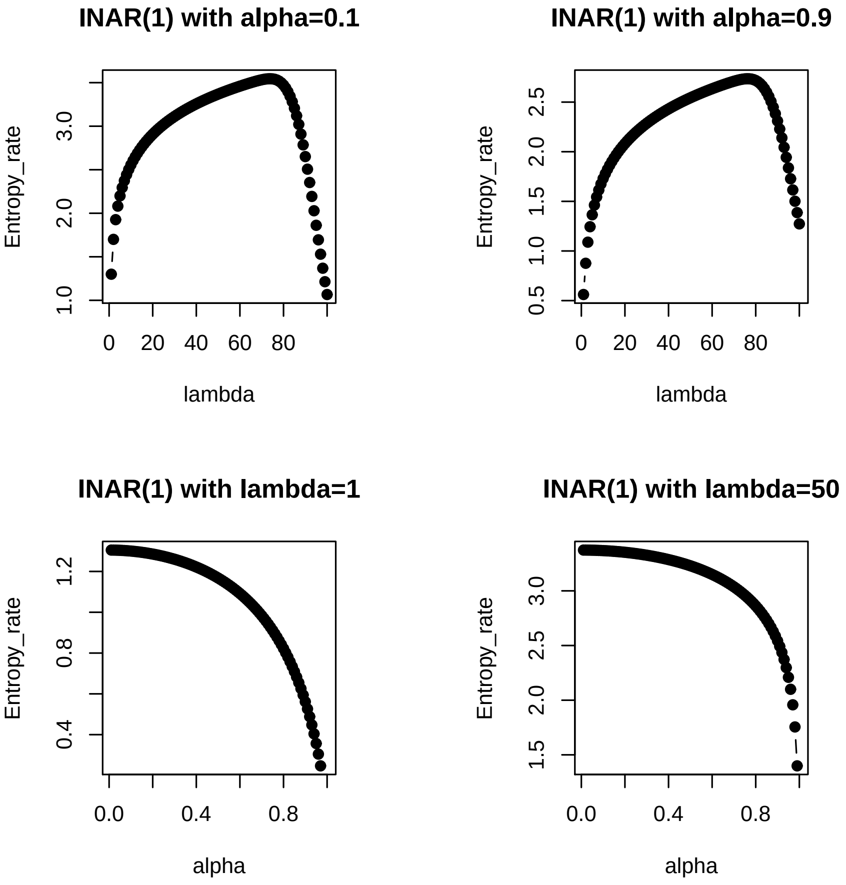


Figure 3: The entropy rate of Poisson INAR(1) process.

References

[1] Al-Osh, M.A. and Alzaid, A.A. (1987), First-order integer-valued autoregressive (INAR(1)) process, *Journal of Time Series Analysis*, **8**, 261-275.

[2] Cover, T. M. and Thomas, J. A. Thomas (2006), *Elements of Information Theory*, John Wiley & Sons, New York.

[3] McKenzie, E. (1985), Some simple models for discrete variate time series. *Water Resources Bulletin*, **21**, 645-650.

[4] Steutel, F.W. and van Harn, K. (1979), Discrete analogues of self-decomposability and stability, *The Annals of Probability*, **7**(5), 893-899.

Topics in maximum entropy modeling in continuous and discrete setting

Asadi, M.

Department of Statistics, University of Isfahan, Isfahan, Iran
Email: asadi4@hotmail.com

Abstract

In this talk we address some new maximum entropy (ME) methods for modeling the distribution of time to an event in both continuous and discrete setting. In continuous case, after reviewing some of the existing results in the literature, we talk a new method which provides characterizations of change point models such as the piecewise exponential distribution. In discrete case it is proved by authors that within the class ultra log-concave distributions (of order n) the Poisson (binomial) distribution is ME. We define a new class of discrete distributions called the weak ultra log-concave (WULC) distributions which includes the negative binomial distribution as boundary. We rise the conjecture that the negative binomial distribution is ME in the class of WULC distributions.

Keywords: Maximum entropy, Log-concave, WULC distributions.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

On robustness of entropy-based goodness of fit tests

Alizadeh Noughabi, R., Aminghafari, M. and Mohammadpour, A.

Department of Statistics, Faculty of Mathematics and Computer Science,
Amirkabir University of Technology, Tehran, Iran
Email: adel@aut.ac.ir

Abstract

We study the robustness of a few entropy and classic goodness of fit tests. We investigate robustness of normality and exponentially tests with respect to heavy tailness and dependence. Non-Gaussian stable distributions are considered for the alternative hypotheses and exchangeability assumption is assumed under the null hypothesis. Our simulation results show that classical tests are more sensitive with respect to a few entropy tests for heavy tail data. However, the entropy and classical tests are robust with respect to the exchangeability assumption.

Keywords: Goodness-of-fit, Entropy, Power of test, Stable distribution.

1 Introduction

The goodness of fit tests had a long history. Well know goodness of fit tests are nonparametric and asymptotically efficient. That is, they can apply for testing almost all family of distributions under a few conditions, without any changes in the test statistic of the goodness of fit tests, and critical value can be computed for large sample sizes. The usual best testing parametric methods cannot be applied to a goodness of fit test problem. This is the main reason that, recently many authors focused on introducing goodness of fit tests for a known density function under simple null hypothesis $H_0 : F = F_0$ (versus $H_1 : F \neq F_0$) to improve power of the well-known goodness of fit tests. That is, their proposed test function can be applied just for a parametric distribution F_0. So, for another distribution, we need a major change in the test statistic or testing method. The idea of introducing a new test function can be conducted in several ways: using the characterization results, introducing efficient estimation of the test function, or proposing new tools to make a sensitive test statistic. These methods usually have not been analytical solution and need many simulation studies to find out the critical values and checking its robustness or sensitiveness with respect to different alternative hypothesis distributions. This is the main drawback of such a goodness of fit tests. Because we cannot check all alternative family of distributions or check their robustness with respect to the assumptions of a goodness of fit test. However, we can compare them with the best parametric test.

Unfortunately, we have many goodness of fit tests without any mathematical justification of power function behavior. This article proposes considering three mentioned directive points. We consider a flexible alternative family of distributions by the name of stable laws and propose to reduce the independence assumption to exchangeability assumption, as an example. Finally, we suggest to consider a benchmark for the best power that is the Neyman-Pearson power.

In the next two sections, we focus on tests for normal and exponential distribution and consider a stable law as the alternative hypothesis. Stable law was introduced by Paul Lévy in his study of sums of independent and identically distributed terms in the 1920s. A stable distribution characterized by four parameters: an index of stability $\alpha \in (0, 2]$, a skewness parameter $\beta \in [-1, 1]$, a scale parameter $\gamma > 0$ and a location parameter $\delta \in \mathbb{R}$. A stable distribution specified by its characteristic function and its as follows

$$\varphi_X(t) = \begin{cases} \exp\left\{-\gamma^\alpha |t|^\alpha \left[1 - i\beta \left(\tan \frac{\pi\alpha}{2}\right)(sign\, t)\right] + i\delta t\right\} & \alpha \neq 1, \\ \exp\left\{-\gamma |t| \left[1 + i\beta \frac{2}{\pi} (sign\, t) \log |t|\right] + i\delta t\right\} & \alpha = 1. \end{cases}$$

where $sign\, t$ is sign function, see Nolan [3]. We denote it by $\mathrm{S}(\alpha, \beta, \gamma, \delta)$ and reduce it to $\mathrm{S}(\alpha, \beta)$ when $\gamma = 1, \delta = 0$.

In recent years, many researchers construct a test for exponentiality and normality based on various entropy estimators. The most popular estimator was introduced by Vasicek [5]. Also Ebrahimi et al. [12] and Correa [13] stated nonparametric entropy estimators. Testing exponentiality discussed by many authors. See, for example, [1, 2, 3, 5]. For test of normality based on entropy we refer to Arizono and Ohta [7], Mudholkar and Lin [8], Esteban et al. [2], Alizadeh Noughabi and Arghami[10] and Zamanzade and Arghami [8].

In section 2, we use Kullback Leibler divergence for testing exponentiality versus asymmetric positive stable distributions ($\alpha < 1, \beta = 1$) and compute the powers of test through a Mone-Carlo simulation study. In section 3, we compared the power of normality tests against the symmetric stable distributions ($\beta = 0$). In section 4, discussed about robustness of normality tests under the exchangeability assumption. Last section states a few concluding remarks.

2 Testing Exponentiality Based on Kullback-Leibler Information

In order to construct a test, we use the Kullback-Leibler discrimination function given by

$$KL(f, f_0) = \int_{-\infty}^{+\infty} f(x) \log\left(\frac{f(x)}{f_0(x)}\right) dx. \tag{1}$$

The evaluation of $KL(f, f_0)$ requires the knowledge of f and f_0, which is not operational. We use Vasicek's estimator H_{mn} to estimate entropy $H(f)$. Also, use the sample mean for estimating parameter of the exponential distribution λ by $\lambda = \frac{1}{\bar{x}}$ and estimate $KL(f : f_0)$ by

$$I_{mn} = -H_{mn} + \log(\bar{x}) + 1.$$

So, large values of I_{mn} indicate that the sample is from a non-exponential distribution.
Ebrahimi et al. [5] consider a monotone transformation of I_{mn}, i.e.

$$KL_{mn} = \exp(I_{mn}) = \exp(H_{mn} - \log(\bar{x}) - 1).$$

Positive stable distributions ($\alpha < 1, \beta = 1$) chosen as alternative hypothesis. The simulation results are reported in tables 1-3.

Table 1: Power comparison for the exponentiality hypothesis versus positive stable law with different index, $n = 10$, and $m = 3$, at significant level 0.05

Alt	T_V	KS	Kuiper	C-VM	AD	Watson
S(0.2,1)	0.973	0.996	0.994	0.997	0.999	0.994
S(0.4,1)	0.701	0.909	0.865	0.915	0.931	0.872
S(0.5,1)	0.521	0.790	0.737	0.806	0.816	0.749
S(0.7,1)	0.383	0.515	0.489	0.528	0.499	0.518
S(0.9,1)	0.979	0.940	0.969	0.937	0.910	0.968

Table 2: Power comparison for the exponentiality hypothesis versus positive stable law with different index, $n = 20$, and $m = 4$, at significant level 0.05

Alt	T_V	KS	Kuiper	C-VM	AD	Watson
S(0.2,1)	1	1	1	1	1	1
S(0.4,1)	0.979	0.995	0.990	0.996	0.997	0.991
S(0.5,1)	0.915	0.968	0.949	0.974	0.974	0.958
S(0.7,1)	0.842	0.786	0.831	0.811	0.808	0.845
S(0.9,1)	1	0.999	1	0.999	0.999	1

Table 3: Power comparison for the exponentiality hypothesis versus positive stable law with different index, $n = 40$, and $m = 6$, at significant level 0.05

Alt	T_V	KS	Kuiper	C-VM	AD	Watson
S(0.2,1)	1	1	1	1	1	1
S(0.4,1)	0.999	1	1	1	1	1
S(0.5,1)	0.997	0.999	0.998	0.999	0.999	0.999
S(0.7,1)	0.995	0.967	0.991	0.979	0.985	0990
S(0.9,1)	1	1	1	1	1	1

3 Testing normality based on entropy estimators

We compare the power of tests based on entropy and classical tests when alternative has a symmetric stable distribution. For this mean, we compare the powers of the tests based on TV_{mn}, $TE_{s_{mn}}$,TC_{mn} and some classical tests. We recall that

$$TV_{mn} = \frac{\exp(H_{mn})}{\hat{\sigma}},$$
$$TEs_{mn} = \frac{\exp(HEs_{mn})}{\hat{\sigma}},$$
$$TC_{mn} = \frac{\exp(HC_{mn})}{\hat{\sigma}}.$$

where $\hat{\sigma} = \sqrt{(1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2}$. Symmetric stable distributions ($\beta = 0$) are chosen as alternative hypothesis. In tables 4-6 simulation results are shown.

Table 4: Power comparison for the normality hypothesis versus symmetric stable law with different index, $n = 10$, and $m = 3$, at significant level 0.05

	Entropy			Classic				
Alt	T_V	T_C	T_{Es}	KS	Kuiper	C-VM	AD	Watson
S(1.1,0)	0.368	0.320	0.557	0.489	0.487	0.524	0.529	0.507
S(1.2,0)	0.306	0.267	0.485	0.415	0.413	0.448	0.458	0.433
S(1.3,0)	0.257	0.232	0.425	0.349	0.345	0.378	0.386	0.367
S(1.4,0)	0.209	0.190	0.352	0.285	0.281	0.313	0.319	0.299
S(1.5,0)	0.169	0.161	0.292	0.235	0.229	0.254	0.264	0.242
S(1.6,0)	0.143	0.131	0.237	0.185	0.178	0.203	0.217	0.193
S(1.7,0)	0.108	0.103	0.176	0.142	0.139	0.159	0.167	0.147
S(1.8,0)	0.085	0.085	0.130	0.104	0.102	0.115	0.120	0.106
S(1.9,0)	0.063	0.066	0.088	0.069	0.071	0.080	0.081	0.075

4 Robustness of test for normality

In this section, we investigate about robustness of entropy and classical tests with respect to the exchangeability assumption. For identically distributed random variables, exchangeability is a weaker condition

Table 5: Power comparison for the normality hypothesis versus symmetric stable law with different index, $n = 20$, and $m = 4$, at significant level 0.05

Alt	T_V	KS	Kuiper	C-VM	AD	Watson
S(1.1,0)	0.601	0.765	0.781	0.802	0.809	0.798
S(1.2,0)	0.502	0.673	0.692	0.717	0.732	0.711
S(1.3,0)	0.425	0.588	0.605	0.632	0.650	0.627
S(1.4,0)	0.349	0.484	0.501	0.534	0.554	0.523
S(1.5,0)	0.285	0.397	0.409	0.442	0.466	0.431
S(1.6,0)	0.221	0.316	0.328	0.356	0.378	0.346
S(1.7,0)	0.166	0.226	0.235	0.255	0.276	0.246
S(1.8,0)	0.117	0.155	0.157	0.176	0.191	0.168
S(1.9,0)	0.081	0.099	0.098	0.107	0.118	0.103

Table 6: Power comparison for the normality hypothesis versus symmetric stable law with different index, $n = 40$, and $m = 6$, at significant level 0.05

Alt	T_V	KS	Kuiper	C-VM	AD	Watson
S(1.1,0)	0.868	0.952	0.963	0.970	0.971	0.971
S(1.2,0)	0.785	0.906	0.923	0.932	0.939	0.934
S(1.3,0)	0.678	0.831	0.854	0.872	0.884	0.872
S(1.4,0)	0.558	0.731	0.758	0.781	0.805	0.782
S(1.5,0)	0.451	0.618	0.643	0.674	0.703	0.674
S(1.6,0)	0.351	0.501	0.522	0.555	0.587	0.552
S(1.7,0)	0.250	0.368	0.392	0.422	0.454	0.417
S(1.8,0)	0.168	0.243	0.254	0.280	0.311	0.276
S(1.9,0)	0.098	0.142	0.142	0.154	0.170	0.151

then independence. That is, we generate identically normally distributed under the null hypothesis with correlation ρ. We investigate robustness of type one error, the probability of rejecting the null when it is true, with respect to the exchangeable normal data. We recall that, the exchangeable normal distribution can be characterized through the following equation

$$\mathbf{X} = (X_1, \ldots, X_n)' \sim N_n(\mu, \Sigma);\ \mu = (\mu, \ldots, \mu)',\ \Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \cdots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

if and only if

$$X_i = \sqrt{\rho} Z_0 + \sqrt{1-\rho} Z_i,\ i = 1, \ldots, n;$$

where $(Z_0, Z_1, \ldots, Z_n)$ is a standard normal random sample.

A part of simulation results are summarized in table 7. This is shown that the robustness of entropy and classical tests with respect to the exchangeability assumption.

5 Conclusion

- For exponential distribution, when the alternative hypothesis is a stable distribution; power of test based on entropy, is less than or equals to the classical tests. Tables 1-3 show that Watson is the best.

  From table 5 we can say that in entropy tests, we prefer the test based on Van Es estimator (TEs). This test also has the high performance among the others.

- When the alternative hypothesis is a stable distribution, the classical tests have a good performance, and we can use them.

Table 7: Percentage of rejecting the null hypothesis for exchangeable normal observations with location zero, scale 1, and correlation 0.2 with $m = 3$ at significant level 0.1

Test	n	0.2	0.5	0.8
Entropy	10	0.09810	0.09844	0.9916
	40	0.1024	0.10186	0.10472
AD	10	0.10116	0.10034	0.09914
	40	0.09932	0.1006	0.10138
KS	10	0.10162	0.10014	0.10058
	40	0.10000	0.10188	0.10332

- Table 7 shows that the entropy and classical are robust tests with respect to the exchangeability assumption.

References

[1] Lilliefors, H. W. (1969). On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. Journal of the American Statistical Association, 64(325), 387-389.

[2] Van Soest, J. (1969). Some goodness of fit tests for the exponential distribution. *Statist. Neerlandica*, **23**, 41-51.

[3] Finkelstein, J., & Schafer, R.E. (1971). Important goodness of fit tests. *Biometrika* **58**, 641–645.

[4] Ebrahimi, N., Habibullah, M., & Soofi, E. S. (1992). Testing exponentiality based on Kullback-Leibler information. *Journal of the Royal Statistical Society.* Series B (Methodological), 739-748.

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

[6] Nolan, J. (2003). Stable distributions: models for heavy-tailed data. Birkhauser.

[7] Arizono, I., & Ohta, H. (1989). A Test for Normality Based on Kullback Leibler Information. *The American Statistician*, **43(1)**, 20-22.

[8] Mudholkar, G. S., & Lin, C. T. (1984). On two applications of characterization theorems to goodness-of-fit. *In Colloq. Math. Soc. Janos Bolyai* **45**, 395-414.

[9] Esteban, M. D., Castellanos, M. E., Morales, D., & Vajda, I. (2001). Monte Carlo comparison of four normality tests using different entropy estimates. *Communications in Statistics–Simulation and computation*, **30(4)**, 761-785.

[10] Alizadeh Noughabi, H., & Arghami, N. R. (2011). Monte Carlo comparison of five exponentiality tests using different entropy estimates. *Journal of Statistical Computation and Simulation*, **81(11)**, 1579-1592.

[11] Zamanzade, E., & Arghami, N. R. (2012). Testing normality based on new entropy estimators. *Journal of Statistical Computation and Simulation*, **82(11)**, 1701-1713.

[12] Ebrahimi, N., Pflughoeft, K., & Soofi, E. S. (1994). Two measures of sample entropy. *Statistics & Probability Letters*, **20(3)**, 225-234.

[13] Correa, J. C. (1995). A new estimator of entropy. *Communications in Statistics–Theory and Methods*, **24(10)**, 2439-2449.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

Monte Carlo comparison of entropy tests with energy for stable distributions

Alizadeh Noughabi, R., Mohammadpour, A.
Department of Statistics, Faculty of Mathematics and Computer Science,
Amirkabir University of Technology, Tehran, Iran
Email: re_al@aut.ac.ir

Abstract

In this paper, we introduce a statistic by the name of energy statistic. It is used to construct a goodness of fit tests. Also, we present an application of the energy statistic as a test for stable distribution. In certain cases, normal and Cauchy, we have been compared power of them with the entropy and classical tests. Our simulation results show that a test based on energy is better than a few entropy tests in some cases.

Keywords: Goodness-of-fit, Energy statistic, Entropy, Power of test.

1 Introduction

One of the statistical problems is to test how applicable n independent measurments agree on a probability model for the experiment. A statistical test, which is designed for deciding whether or not a random sample may have been drawn by a specified distribution F_0 is called goodness of fit test.

Stable distributions are a rich class of probability distributions that allow skewness and heavy tails and have many intriguing mathematical properties. A stable distribution specified by four parameters, an index of stability $\alpha \in (0, 2]$, a skewness parameter $\beta \in [-1, 1]$, a scale parameter $\gamma > 0$ and finally a location parameter $\delta \in \Re$. A stable distribution determined by its characteristic function, that is $X \sim S(\alpha, \beta, \gamma, \delta)$ if and only if $\varphi_X(t)$ as follows

$$\varphi_X(t) = \begin{cases} \exp\left\{-\gamma^\alpha |t|^\alpha \left[1 - i\beta \left(\tan \frac{\pi\alpha}{2}\right)(sign\, t)\right] + i\delta t\right\} & \alpha \neq 1, \\ \exp\left\{-\gamma\, |t| \left[1 + i\beta \frac{2}{\pi}(sign\, t) \log |t|\right] + i\delta t\right\} & \alpha = 1. \end{cases}$$

where $sign\, t$ is sign function, see Nolan [3]. There is not a closed form for stable distributions and this issue has been overcome partly by the development of computer programs. In this paper for all parameterizations, the notation $S(\alpha, \beta) = S(\alpha, \beta, 1, 0)$ will be used.

Suppose that we have a random sample of size n with probability distribution F and density function f. Vasicek (1976) introduced the estimate of entropy based on order statistics. Many researchers have been interested in test normality based on entropy, see Vasicek [5], Ebrahimi et al. [5], Esteban et al. [2], Yousefzadeh and Arghami [7] and Zamanzade and Arghami [8].

In section 2, we introduce the energy statistic and in used to for goodness of fit for stable distributions. In section 3, compare the powers of test based on energy with the powers of entropy-based tests. In section 4, we present a test for Cauchy distribution, as a member of stable family distributions, based on entropy and then compare power of this test with test based on energy statistic. In last section, state some concluding remarks.

2 Energy statistic

The energy statistic discussed by Rizzo [4] for goodness of fit test. The original energy statistic is as follows

$$\begin{aligned} Q_n &= n \int_{-\infty}^{+\infty} \left|\varphi_{F_n}(t) - \varphi_{F_0}(t)\right|^2 \tfrac{1}{\pi t^2} dt \\ &= n \left\{ \tfrac{2}{n} \sum_{j=1}^{n} E\left|x_j - X\right| - E\left|X - X'\right| - \tfrac{1}{n^2} \sum_{j,k=1}^{n} \left|x_j - x_k\right| \right\}, \end{aligned} \tag{1}$$

where φ is characteristic function and X is a random variable from F_0, also X' is independent copy of X. The formula (1) is useful if $E\left|x_j - X\right| < \infty$, and if X has a stable distribution with $\alpha < 2$, $E\left|x_j - X\right|$ is not finite. A modified energy statistic for testing stable distribution proposed as follows

$$Q_{n,s} = n \left\{ \frac{2}{n} \sum_{j=1}^{n} E|x_j - X|^s - E|X - X'|^s - \frac{1}{n^2} \sum_{j,k=1}^{n} |x_j - x_k|^s \right\}, \tag{2}$$

where $X \sim S(\alpha, \beta, \gamma, \delta)$ and s is less than α. Yang [6] by some algebraic operations shows that $Q_{n,s}$ can be summarized as follows. If $\alpha \neq 1$

$$\begin{aligned} Q_{n,s} &= \tfrac{4}{\pi}\Gamma(1+s)\sin\left(\tfrac{\pi s}{2}\right) \sum_{j=1}^{n} \int_0^\infty \tfrac{1-e^{t^\alpha}\cos\left(\beta t^\alpha \tan\left(\frac{\pi\alpha}{2}\right) - x_j t\right)}{t^{s+1}} \\ &- \tfrac{n 2^{\frac{\alpha+s}{\alpha}}}{\pi}\Gamma\left(\tfrac{\alpha-s}{\alpha}\right)\Gamma(s)\sin\left(\tfrac{\pi s}{2}\right) - \tfrac{1}{n}\sum_{j,k=1}^{n} \left|x_j - x_k\right|^s. \end{aligned} \tag{3}$$

If $\alpha = 1$

$$\begin{aligned} Q_{n,s} &= \tfrac{4}{\pi}\Gamma(1+s)\sin\left(\tfrac{\pi s}{2}\right) \sum_{j=1}^{n} \int_0^\infty \tfrac{1-e^{t}\cos\left(\beta \frac{2t\log t}{\pi} + x_j t\right)}{t^{s+1}} \\ &- \tfrac{n 2^{1+s}}{\pi}\Gamma(1-s)\Gamma(s)\sin\left(\tfrac{\pi s}{2}\right) - \tfrac{1}{n}\sum_{j,k=1}^{n} \left|x_j - x_k\right|^s. \end{aligned} \tag{4}$$

3 Compare goodness of fit tests for normality

To specify that a random sample follows from H_0 with a probability density function f_0, we must have a test statistic. In this section, we compare the power of tests based on energy statistic and entropy. To do this, we calculate the powers of the tests based on TD, TV_{mn}, $TE_{s_{mn}}$ and TC_{mn}. Esteban et al. [2], in their study of power comparisons of normality tests; offer to classify the alternatives into the following groups

Group I: Support= $(-\infty, +\infty)$, symmetric.

Group II: Support= $(-\infty, +\infty)$, asymmetric.

Group III: Support= $(0, +\infty)$.

Group IV: Support= $(0, 1)$.

It is logical that we consider the groups I and II, but we have studied all groups. In table 8 the critical values for test of normality mentioned.

Table 8: Critical values of energy statistic for $\alpha = 0.05$

n	$s = 0.3$	$s = 0.6$	$s = 0.9$
10	1.639	2.374	3.629
20	1.595	2.306	3.608
40	1.484	2.122	3.558

* Skew normal (SN) with parameters $\mu = 0$ (location), $\sigma = 1$ (scale) and $a = 2$ (shape).

**Skew double exponential (SDE) with parameters $a = 1$, $\beta = 2$ and $\mu = 0$ (location) (mixture exponential distribution with mean $\beta = 2$, and the negative of an exponential distribution with mean $a = 1$).

Table 9: Power comparison test for $n = 10$, $m = 2$ under the alternative in group I with $\alpha = 0.05$

	Entropy				Energy		
Alt	T_D	T_V	T_{Es}	T_C	$s = 0.3$	$s = 0.6$	$s = 0.9$
t(1)	0.583	0.442	0.591	0.409	0.265	0.431	0.513
t(3)	0.201	0.091	0.167	0.083	0.052	0.046	0.047
Laplace	0.163	0.065	0.140	0.057	0.058	0.047	0.033
Logistic	0.087	0.051	0.074	0.047	0.074	0.111	0.117
Average	0.258	0.216	0.243	0.149	0.112	0.159	0.177

Table 10: Power comparison test for $n = 40$, $m = 4$ under the alternative in group I with $\alpha = 0.05$

	Entropy				Energy		
Alt	T_D	T_V	T_{Es}	T_C	$s = 0.3$	$s = 0.6$	$s = 0.9$
t(1)	0.991	0.960	0.987	0.949	0.401	0.815	0.860
t(3)	0.612	0.289	0.541	0.249	0.117	0.103	0.072
Laplace	0.533	0.197	0.451	0.198	0.205	0.143	0.092
Logistic	0.210	0.053	0.160	0.048	0.066	0.170	0.166
Average	0.586	0.374	0.534	0.351	0.197	0.308	0.297

Table 11: Power comparison test for $n = 10$, $m = 2$ under the alternative in group II with $\alpha = 0.05$

	Entropy				Energy		
Alt	T_D	T_V	T_{Es}	T_C	$s = 0.3$	$s = 0.6$	$s = 0.9$
Gumbel	0.154	0.101	0.113	0.097	0.129	0.148	0.158
SN(0,1,2)*	0.071	0.058	0.062	0.053	0.767	0.720	0.612
SDE(0,1,2)**	0.216	0.117	0.178	0.111	0.261	0.396	0.436
Average	0.147	0.092	0.117	0.087	0.386	0.421	0.402

Table 12: Power comparison test for $n = 40$, $m = 4$ under the alternative in group II with $\alpha = 0.05$

	Entropy				Energy		
Alt	T_D	T_V	T_{Es}	T_C	$s = 0.3$	$s = 0.6$	$s = 0.9$
Gumbel	0.530	0.399	0.355	0.394	0.681	0.732	0.765
SN(0,1,2)	0.149	0.099	0.097	0.102	1	1	1
SDE(0,1,2)	0.693	0.420	0.586	0.385	0.673	0.880	0.922
Average	0.457	0.306	.0346	0.293	0.785	0.871	0.896

Table 13: Power comparison test for $n = 10$, $m = 2$ under the alternative in group III with $\alpha = 0.05$

	Entropy				Energy		
Alt	T_D	T_V	T_{Es}	T_C	$s = 0.3$	$s = 0.6$	$s = 0.9$
Exp(1)	0.394	0.416	0.330	0.404	0.999	1	1
Gamma(2,1)	0.222	0.179	0.158	0.173	1	1	1
Gamma(0.5,1)	0.631	0.782	0.679	0.786	1	1	0.809
LN(0,1)	0.565	0.552	0.485	0.542	1	1	1
Weibull(0.5,1)	0.813	0.931	0.876	0.926	0.998	1	0.941
Average	0.525	0.572	.0506	0.566	0.999	1	0.950

4 A goodness of fit test for Cauchy distribution

4.1 Based on energy

Cauchy distribution is a member of family of stable distributions. As an example of stable distribution, we consider the Cauchy distribution for goodness of fit test. Formula (4) by some simple algebraic operation,

Table 14: Power comparison test for $n = 10$, $m = 2$ under the alternative in group IV with $\alpha = 0.05$

	Entropy				Energy		
Alt	T_D	T_V	T_{Es}	T_C	$s = 0.3$	$s = 0.6$	$s = 0.9$
Uniform	0.028	0.167	0.061	0.170	1	1	1
Beta(2,2)	0.025	0.082	0.037	0.086	1	1	1
Beta(2,1)	0.093	0.173	0.092	0.182	1	1	1
Average	0.049	0.141	0.063	0.129	1	1	1

simplified as follows

$$Q_{n,s} = 2\sum_{j=1}^{n} \frac{\left(1+{x_j}^2\right)^{s/2} \cos\left(s\, arc \tan x_j\right)}{\cos\left(\frac{\pi s}{2}\right)} - \frac{n2^s}{\cos\left(\frac{\pi s}{2}\right)} - \frac{1}{n}\sum_{j,k=1}^{n} |x_j - x_k|^s. \tag{5}$$

4.2 Based on entropy

For this mean, we use the Kullback-Leibler discrimination function given by

$$KL\left(f, f_0\right) = \int_{-\infty}^{+\infty} f\left(x\right) \log\left(\frac{f\left(x\right)}{f_0\left(x\right)}\right) dx. \tag{6}$$

The evaluation of $KL\left(f, f_0\right)$ requires the knowledge of f and f_0, which is not operational. We can rewrite (6) to

$$KL\left(f, f_0\right) = -H\left(f\right) - \int_{-\infty}^{+\infty} f\left(x\right) \log\left(f_0\left(x\right)\right) dx. \tag{7}$$

We use Vasiceks estimator H_{mn} to estimate $H(f)$ and to estimate $\int_{-\infty}^{+\infty} f\left(x\right) \log\left(f_0\left(x\right)\right) dx$, use the below expression

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(f_0\left(x_i, \hat{\mu}, \hat{\sigma}\right)\right),$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the maximum likelihood estimators of μ and σ.
So, an estimator KL_{mn} of $KL\left(f, f_0\right)$ is obtained as follows

$$\begin{aligned} KL_{mn} &= -H_{mn} - \frac{1}{n}\sum_{i=1}^{n} \log\left(f_0\left(x_i, \hat{\mu}, \hat{\sigma}\right)\right) \\ &= -H_{mn} - \log\left(\hat{\sigma}\right) + \log\left(\pi\right) + \frac{1}{n}\sum_{i=1}^{n} \log\left(\hat{\sigma} + \left(x_i - \hat{\mu}\right)^2\right). \end{aligned}$$

Table 15: Power comparison test for $n = 20$ with $\alpha = 0.05$

	Entropy			Energy				
Alt	$m = 2$	$m = 3$	$m = 4$	$s = 0.1$	$s = 0.2$	$s = 0.3$	$s = 0.4$	AD
Stable(0.5,0)	0.101	0.022	0.009	0.645	0.719	0.766	0.791	0.401
Stable(0.8,0)	0.039	0.024	0.019	0.126	0.152	0.1839	0.1974	0.088
Stable(1.2,0)	0.089	0.113	0.114	0.035	0.030	0.024	0.017	0.036
Stable(1.5,0)	0.210	0.281	0.287	0.047	0.037	0.027	0.014	0.035
Stable(1.8,0)	0.411	0.526	0.555	0.066	0.053	0.0401	0.021	0.037
Stable(2,0)	0.595	0.727	0.760	0.078	0.064	0.045	0.029	0.038
Normal	0.595	0.728	0.754	0.234	0.211	0.168	0.096	0.045
t(2)	0.168	0.218	0.224	0.050	0.037	0.028	0.015	0.031
t(3)	0.266	0.347	0.371	0.078	0.064	0.048	0.025	0.027
t(4)	0.326	0.436	0.451	0.105	0.084	0.064	0.033	0.034
t(5)	0.376	0.491	0.510	0.116	0.102	0.073	0.43	0.035
Laplace	0.245	0.326	0.346	0.108	0.089	0.072	0.037	0.032
Gumbel	0.589	0.691	0.741	0.278	0.254	0.226	0.166	0.281

Table 16: Power comparison test for $n = 100$ with $\alpha = 0.1$

	Entropy		Energy				
Alt	$m = 6$	$m = 7$	$s = 0.1$	$s = 0.2$	$s = 0.3$	$s = 0.4$	AD
Stable(0.5,0)	0.238	0.126	0.999	0.999	0.999	0.999	0.988
Stable(0.8,0)	0.054	0.047	0.411	0.461	0.506	0.550	0.256
Stable(1.2,0)	0.271	0.284	0.154	0.156	0.143	0.137	0.108
Stable(1.5,0)	0.710	0.704	0.566	0.599	0.621	0.611	0.244
Stable(1.8,0)	0.964	0.964	0.949	0.971	0.977	0.984	0.559
Stable(2,0)	1	1	0.999	0.999	1	1	0.842
Normal	1	1	1	1	1	1	1
t(2)	0.745	0.755	0.679	0.735	0.742	0.755	0.439
t(3)	0.964	0.967	0.969	0.984	0.987	0.989	0.856
t(4)	0.944	0.995	0.997	0.999	0.999	0.999	0.968
t(5)	0.999	0.999	0.999	1	1	1	0.992
Laplace	0.999	0.999	0.994	0.997	0.999	0.999	0.950
Gumbel	1	1	1	1	1	1	1

5 Main results

- For test of normality in distributions of group I, mostly, the entropy tests are better than the test based on energy statistics.

- In group II, In most cases the test of normality based on the energy statistic is better than the entropy tests.

- In the other two groups the energy test, always is better. It that means if an irrational distribution be considered as an alternative energy test detects very well.

- In test for Cauchy distribution, in some cases entropy test has more power. However, it can be said by increasing n the power of energy test more improved to entropy test.

References

[1] Ebrahimi, N., Habibullah, M., Soofi, E. S. (1992). Testing exponentiality based on Kullback-Leibler information. *Journal of the Royal Statistical Society.* Series B (Methodological), 739-748.

[2] Esteban, M. D., Castellanos, M. E., Morales, D., Vajda, I. (2001). Monte Carlo comparison of four normality tests using different entropy estimates. *Communications in Statistics-Simulation and computation*, **30(4)**, 761-785.

[3] Nolan, J. (2003). Stable distributions: models for heavy-tailed data. Birkhauser.

[4] Rizzo, M. L. (2002). A new rotation invariant goodness-of-fit test (Doctoral dissertation, Bowling Green State University).

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

[6] Yang, G. (2012). The Energy Goodness-of-Fit Test for Univariate Stable Distributions (Doctoral dissertation, Bowling Green State University).

[7] Yousefzadeh, F., Arghami, N. R. (2008). Testing exponentiality based on type II censored data and a new cdf estimator. *Communications in StatisticsSimulation and Computation*, **37(8)**, 1479-1499.

[8] Zamanzade, E., Arghami, N. R. (2011). Goodness-of-fit test based on correcting moments of modified entropy estimator. *Journal of Statistical Computation and Simulation*, **81(12)**, 2077-2093.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

On entropy order for order statistics and their concomitants

Amiri, M., Amini-Seresht, E. and Khaledi, B.

Department of Statistics, Razi University, Kermanshah, Iran
Email: bkhaledi@hotmail.com

Abstract

Let (X, Y) and (S, T) be two continuous random vectors. It is shown that if S, $[Y|X = x]$ and $[T|S = x]$, for all x are DFR, Y is stochastically increasing in X and $(X, Y) \leq_{sst} (S, T)$, then $H(X, Y) \leq H(S, T)$, where $H(Z)$ is Shannon entropy of a random variable Z. Let (X_i, Y_i), $i = 1, \ldots, max\{m, n\}$ be a set of independent copies of (X, Y). It is also shown that if X and $[Y|X = x]$, for all x have DFR distributions and Y is stochastically increasing in X, then for $i \leq j$ and $n - i \geq m - j$, $H(X_{i:n}, Y_{[i:n]}) \leq H(X_{j:m}, Y_{[j:m]})$. Let (S_i, T_i), $i = 1, \ldots, max\{m, n\}$ be a set of independent copies of (S, T). It is observed that under certain set of mild conditions on $F_{X,Y}$ and $F_{S,T}$, for $i \leq j$ and $n - i \geq m - j$, $H(X_{i:n}, Y_{[i:n]}) \leq H(S_{j:m}, T_{[j:m]})$. Finally, we discuss some conjectures about entropy properties of vector of order statistics corresponding to a random sample of size n from a symmetric distribution which admits density.

Keywords: Dispersive order, Decreasing failure rate, Strong stochastic order, Symmetric distribution.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

Entropy of distribution and relative entropy on Polish spaces

Ghazani, Z.

Islamic Azad University, Central Tehran Branch, Iran
Email: Ghazaniz@yahoo.com

Abstract

In this paper, we assume μ is a probability measure and the conditions for maximal entropy are provided. The properties of relative entropy for probability measure on polish spaces are also discussed.

Keywords: Entropy, Uniform distribution, Geometric distribution, Product distribution.

1 Introduction

Let (Ω, A) denote a measure space and ν is a measure. The measure is defined only on a δ-subring of A since we did not assume that ν is finite. For any probability measure $\mu \in \mathcal{P}(\nu)$ define the entropy

$$H(\mu) = \int_{\Omega} -f(\omega)\log(f(\omega))d\nu(\omega).$$

Example 1. *Let ν be the counting measure on a countable set Ω, where $\mathcal{A}$ is a σ-algebra of all subsets of Ω and let the measure ν is defines on the δ-ring of all finite subsets of Ω. In this case,*

$$H(\mu) = \sum_{\omega\in\Omega} -f(\omega)\log(f(\omega)).$$

For example, for $\Omega = \mathbb{N} = \{0, 1, 2, 3, \ldots\}$ with counting measure ν, the geometric distribution $P[\{k\}] = p(1-p)^k$ has the entropy

$$\sum_{k=0}^{\infty} -(1-p)^k p\log((1-p)^k p) = \log(\frac{1-p}{p}) - \frac{\log(1-p)}{p}.$$

Given two probability measure $\mu = f\nu$ and $\tilde{\mu} = \tilde{f}\nu$ which are both absolutely continuous with respect to ν. Define the relative entropy

$$H(\tilde{\mu} \mid \mu) = \int_{\Omega} \tilde{f}(\omega)\log(\frac{\tilde{f}(\omega)}{f(\omega)})d\nu(x) \in [0, \infty].$$

It is the expectation $E_{\tilde{\mu}}[l]$ of the Likelihood coefficient $l = \log(\frac{\tilde{f}(x)}{f(x)})$. The negative relative entropy $-H(\tilde{\mu} \mid \mu)$ is also called the conditional entropy. We write also $H(f \mid \tilde{f})$ instead of $H(\tilde{\mu} \mid \mu)$.

Theorem 1.1. $0 \leq H(\tilde{\mu} \mid \mu) \leq +\infty$ *and* $H(\tilde{\mu} \mid \mu) = 0$ *if and only if* $\mu = \tilde{\mu}$.

Proof. We can assume $H(\tilde{\mu} \mid \mu) < \infty$. The function $u(x) = x\log(x)$ is convex on $\mathbb{R}^+ = [0, \infty)$ and satisfies $u(x) \geq x - 1$.

$$\begin{aligned}
H(\tilde{\mu} \mid \mu) &= \int_\Omega \tilde{f}(\omega)\log(\frac{\tilde{f}(\omega)}{f(\omega)})d\nu(\omega) \\
&= \int_\Omega \tilde{f}(\omega)\frac{\tilde{f}(\omega)}{f(\omega)}\log(\frac{\tilde{f}(\omega)}{f(\omega)})d\nu(\omega) \\
&= \int_\Omega \tilde{f}(\omega)u(\frac{f(\omega)}{\tilde{f}(\omega)})d\nu(\omega) \\
&\geq \int_\Omega \tilde{f}(\omega)((\frac{f(\omega)}{\tilde{f}(\omega)} - 1)d\nu(\omega) \\
&= \int_\Omega f(\omega) - \tilde{f}(\omega)d\nu(\omega) = 1 - 1 = 0.
\end{aligned}$$

If $\mu = \tilde{\mu}$, then $f = \tilde{f}$ almost everywhere then $\dfrac{f(\omega)}{\tilde{f}(\omega)} = 1$ almost everywhere and $H(\tilde{\mu} \mid \mu) = 0$. On the other hand, if $H(\tilde{\mu} \mid \mu) = 0$, then by the Jensen inequality

$$0 = E_{\tilde{\mu}}[u(\frac{\tilde{f}}{f})] \geq u(E_{\tilde{\mu}}[\frac{\tilde{f}}{f}]) = u(1) = 0.$$

Therefore, $E_{\tilde{\mu}}[u(\frac{\tilde{f}}{f})] = u(E_{\tilde{\mu}}[\frac{\tilde{f}}{f}])$. The strict convexity of u implies that $\frac{\tilde{f}}{f}$ must be a constant almost everywhere. Since both f and $\tilde{f}$ are densities, the equality $f = \tilde{f}$ must be true almost everywhere. □

Theorem 1.2. *The following distributions have maximal entropy,*

a) If Ω is finite with counting measure ν. The uniform distribution on Ω has maximal entropy among all distributions on Ω. It is unique with this property.

b) $\Omega = \mathbb{N}$ with counting measure ν. The geometric distribution with parameter $p = c^{-1}$ has maximal entropy among all distributions on $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ with fixed mean c. It is unique with this property.

c) $\Omega = \{0, 1\}^N$ with counting measure ν. The product distribution η^N, where $\eta(1) = p$, $\eta(0) = 1 - p$ with $p = c/N$ has maximal entropy among all distributions satisfying $E[S_N] = c$, where $S_N(\omega) = \sum_{i=1}^N \omega_i$. It is unique with this property.

d) $\Omega = [0, \infty)$ with Lebesgue measure ν. The exponential distribution with density $f(x) = \lambda e^{-\lambda x}$ with parameter λ on Ω has the maximal entropy among all distributions with fixed mean $c = 1/\lambda$. It is unique with this property.

e) $\Omega = \mathbb{R}$ with Lebesgue measure ν. The normal distribution $N(m, \sigma^2)$ has maximal entropy among all distributions with fixed mean m and fixed variance σ^2. It is unique with this property.

f) Finite measures. Let $(\Omega, \mathcal{A})$ be an arbitrary measure space for which $0 < \nu(\Omega) < \infty$. Then the measure ν with uniform distribution $f = 1/\nu(\Omega)$ has maximal entropy among all other measures on Ω. It is unique with this property.

Proof. Let $\mu = f\nu$ be the measure of the distribution from which we want to prove maximal entropy and let $\tilde{\mu} = \tilde{f}\nu$ be any other measure. The aim is to show $H(\tilde{\mu} \mid \mu) = H(\mu) - H(\tilde{\mu})$ which implies the maximality since by the Theorem 1.1 $H(\tilde{\mu} \mid \mu) \geq 0$.
In general,

$$H(\tilde{\mu} \mid \mu) = -H(\tilde{\mu}) - \int_\Omega \tilde{f}(\omega)\log(f(\omega))d\nu$$

so that in each case, we have to show

$$H(\mu) = -\int_{\Omega} \tilde{f}(\omega)\log(f(\omega))d\nu. \tag{1}$$

With

$$H(\tilde{\mu} \mid \mu) = H(\mu) - H(\tilde{\mu})$$

we also have uniqueness: if two measures $\tilde{\mu}$, μ have maximal entropy, then $H(\tilde{\mu} \mid \mu) = 0$ so that by the Theorem 1.1 $\mu = \tilde{\mu}$.

a) The density $f = 1/[\Omega]$ is constant. Therefore $H(\mu) = \log(|\Omega|)$ and equation (1) holds.

b) The geometric distribution on $\mathbb{N} = \{0, 1, 2, \ldots\}$ satisfies $P[\{k\}] = f(k) = p(1-p)^k$. We have computed the entropy before as

$$\log((1-p)/p) - (\log(1-p))/p = -\log(p) - \frac{(1-p)}{p}\log(1-p).$$

c) The discrete density is $f(\omega) = p^{S_N}(1-p)^{N-S_N}$ so that

$$\log(f(k)) = S_N \log(p) + (N - S_N)\log(1-p)$$

and

$$\sum_k \tilde{f}(k)\log(f(k)) = E[S_N]\log(p) + (N - E[S_N])\log(1-p).$$

The claim follows since we fixed $E[S_N]$.

d) The density is $f(x) = \alpha e^{-\alpha x}$, so that $\log(f(x)) = \log(\alpha) - \alpha x$. The claim follows since we fixed $E[X] = \int x d\tilde{\mu}(x)$ was assumed to be fixed for all distributions.

e) For the normal distribution $\log(f(x)) = a + b(x-m)^2$ with two real number a, b depending only on m and σ. The claim follows since we fixed $Var[X] = E[(x-m)^2]$ for all distributions.

f) The density $f = 1$ is constant. Therefore $H(\mu) = 0$ which is also on the right hand side of equation (1).

□

2 Relative entropy on Polish space

Here we collect properties of relative entropy for probability measures on Polish spaces, relative entropy is introduced as an information measure and called directed divergence. Let $\mathcal{S}$, $\mathcal{X}$, $\mathcal{Y}$ be Polish spaces. A Polish space is a separable topological space that is compatible with a complete metric. Examples of Polish spaces are

- $\mathbb{R}^d$ with the standard topology,
- any closed subset of $\mathbb{R}^d$ (or another Polish space) equipped with the induced topology,
- the space $C(\mathbb{T}, \mathcal{X})$ of continuous functions, $\mathbb{T} \subseteq (-\infty, \infty)$ an interval, $\mathcal{X}$ a complete and separable metric space, equipped with the topology of uniform convergence on compact subsets of $\mathbb{T}$,
- the space $D(\mathbb{T}, \mathcal{X})$ of cádlág functions, $\mathbb{T} \subseteq (-\infty, \infty)$ an interval, a $\mathcal{X}$ a complete and separable metric space, equipped with the Skorohod topology,
- the space $\mathcal{P}(\mathcal{X})$ of probability measures on $\mathcal{B}(\mathcal{X})$, $\mathcal{X}$ a Polish space, equipped with the weak convergence topology.

Let $\mu, \nu \in \mathcal{P}(\mathcal{S})$. The relative entropy of μ with respect to ν is given by

$$H(\mu\|\nu) = \begin{cases} \int_{\mathcal{S}} \log\left(\frac{d\mu}{d\nu}(x)\right)\mu(dx) & \text{if } \mu \ll \nu, \\ \infty & \text{else.} \end{cases}$$

Relative entropy is well-defined as a function $\mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S}) \to [0, \infty]$. Indeed, if $\mu \ll \nu$, then a density $f = \dfrac{d\mu}{d\nu}$ exists by the Radon-Nikodym theorem with f uniquely determined ν-almost surely. In this case,

$$H(\mu\|\nu) = \int_{\mathcal{S}} f(x)\log(f(x))\nu(dx).$$

Clearly, $\lim_{x\to 0^+} x\log(x) = 0$. Since $\int f d\nu = 1$ and $x\log(x) \geq x - 1$ for all $x \geq 0$ with equality if and only if $x = 1$, it follows that $H(\mu\|\nu) \geq 0$ with $H(\mu\|\nu) = 0$ if and only if $\mu = \nu$. Relative entropy can actually be defined for σ-finite measures on an arbitrary measurable space.

Lemma 2.1 (Basic properties)**.** *Properties of relative entropy $H(\cdot\|\cdot)$ for probability measures on a Polish space $\mathcal{S}$.*

- (*a*) *Relative entropy is a non-negative, convex, lower semi-continuous function $\mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{S}) \to [0, \infty]$.*
- (*b*) *For $\nu \in \mathcal{P}(\mathcal{S})$, $H(\cdot\|\nu)$ is strictly convex on $\{\mu \in \mathcal{P}(\mathcal{S}) : H(\mu\|\nu) < \infty\}$.*
- (*c*) *For $\nu \in \mathcal{P}(\mathcal{S})$, $H(\cdot\|\nu)$ has compact sublevel sets.*
- (*d*) *Let $\Pi_{\mathcal{S}}$ denote the set of finite measurable partitions of $\mathcal{S}$. Then for all $\mu, \nu \in \mathcal{P}(\mathcal{S})$,*

$$H(\mu\|\nu) = \sup_{\pi \in \Pi_{\mathcal{S}}} \sum_{A \in \pi} \mu(A)\log\left(\frac{\mu(A)}{\nu(A)}\right),$$

where $x\log(x/y) = 0$ if $x = 0$, $x\log(x/y) = \infty$ if $x > 0$ and $y = 0$.

- (*e*) *For every $A \in \mathcal{B}(\mathcal{S})$, any $\mu, \nu \in \mathcal{P}(\mathcal{S})$,*

$$H(\mu\|\nu) \geq \mu(A)\log\left(\frac{\mu(A)}{\nu(A)}\right) - 1.$$

Lemma 2.2 (Contraction property)**.** *Let $\psi : \mathcal{Y} \to \mathcal{X}$ be a Borel measurable mapping. Let $\eta \in \mathcal{P}(\mathcal{X})$, $\gamma_0 \in \mathcal{P}(\mathcal{Y})$. Then*

$$H(\eta\|\gamma_0 \circ \psi^{-1}) = \inf_{\gamma \in \mathcal{P}(\mathcal{Y}): \gamma\circ\psi^{-1} = \eta} H(\gamma\|\gamma_0), \tag{2}$$

where $\inf \varnothing = \infty$ *by convention.*

Lemma 2.3 (Chain rule)**.** *Let $\mathcal{X}$, $\mathcal{Y}$ be Polish spaces. Let $\alpha, \beta \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and denote their marginal distributions on $\mathcal{X}$ by α_1 and β_1, respectively. Let $\alpha(\cdot|\cdot)$, $\beta(\cdot|\cdot)$ be stochastic kernels on $\mathcal{Y}$ given $\mathcal{X}$ such that for all $A \in \mathcal{B}(\mathcal{X})$, $B \in \mathcal{B}(\mathcal{Y})$,*

$$\alpha(A \times B) = \int_A \alpha(B|x)\alpha_1(dx), \qquad \beta(A \times B) = \int_A \beta(B|x)\beta_1(dx).$$

Then the mapping $x \mapsto H(\alpha(\cdot|x)\|\beta(\cdot|x))$ is measurable and

$$H(\alpha\|\beta) = H(\alpha_1\|\beta_1) + \int_{\mathcal{X}} H(\alpha(\cdot|x)\|\beta(\cdot|x))\alpha_1(dx).$$

In particular, if α, β are product measures, then

$$H(\alpha_1 \otimes \alpha_2\|\beta_1 \otimes \beta_2) = H(\alpha_1\|\beta_1) + H(\alpha_2\|\beta_2).$$

The variational representation for Laplace functionals given in below is the starting point for the weak convergence approach to large deviations. Let $M_b(\mathcal{X})$ is a space of all bounded measurable function $\mathcal{X} \to R$.

Lemma 2.4. *Let $\nu \in \mathcal{P}(\mathcal{S})$. Then for all $g \in M_b(\mathcal{S})$,*

$$-\log \int_{\mathcal{S}} \exp(-g(x))\nu(dx) = \inf_{\mu \in \mathcal{P}(\mathcal{S})} \left\{ H(\mu\|\nu) + \int_{\mathcal{S}} g(x)\mu(dx) \right\},$$

Infimum in variational formula above is attained at $\mu^* \in \mathcal{P}(\mathcal{S})$ given by

$$\frac{d\mu^*}{d\nu} = \frac{\exp(-g(x))}{\int_{\mathcal{S}} \exp(-g(y))\nu(dy)}, \quad x \in \mathcal{S}.$$

Proof. Let $g \in M_b(\mathcal{S})$, and define μ^* through its density with respect to ν as above. Notice that μ^*, ν are mutually absolutely continuous. Let $\mu \in \mathcal{P}(\mathcal{S})$ be such that $H(\mu\|\nu) < \infty$. Then μ is absolutely continuous with respect to ν with density $\frac{d\mu}{d\nu}$, but also absolutely continuous with respect to μ^* with density $\frac{d\mu}{d\mu^*} = \frac{d\mu}{d\nu} \cdot \frac{d\nu}{d\mu^*}$, where $\frac{d\nu}{d\mu^*} = \frac{e^g}{\int e^g d\mu^*}$. It follows that

$$\begin{aligned} H(\mu\|\nu) + \int_{\mathcal{S}} g d\mu &= \int_{\mathcal{S}} \log\left(\frac{d\mu}{d\nu}\right) d\mu + \int_{\mathcal{S}} g d\mu \\ &= \int_{\mathcal{S}} \log\left(\frac{d\mu}{d\mu^*}\right) d\mu + \int_{\mathcal{S}} \log\left(\frac{d\mu^*}{d\nu}\right) d\mu + \int_{\mathcal{S}} g d\mu \\ &= H(\mu\|\mu^*) - \log \int_{\mathcal{S}} e^{-g} d\nu. \end{aligned}$$

This yields the assertion since $H(\mu\|\mu^*) \geq 0$ with $H(\mu\|\mu^*) = 0$ if and only if $\mu = \mu^*$. □

Theorem 2.5. *Let $\mu, \nu \in \mathcal{P}(\mathcal{S})$. Then*

$$H(\mu\|\nu) = \sup_{g \in M_b(\mathcal{S})} \left\{ \int_{\mathcal{S}} g(x)\mu(dx) - \log \int_{\mathcal{S}} \exp(g(x))\nu(dx) \right\}.$$

Proof. Let $\mu, \nu \in \mathcal{P}(\mathcal{S})$. By 2.4 for every $g \in M_b(\mathcal{S})$

$$H(\mu\|\nu) \geq -\int_{\mathcal{S}} g d\mu - \log \int_{\mathcal{S}} e^{-g} d\nu,$$

hence

$$\begin{aligned} H(\mu\|\nu) &\geq \sup_{g \in M_b(\mathcal{S})} \left\{ -\int_{\mathcal{S}} g d\mu - \log \int_{\mathcal{S}} e^{-g} d\nu \right\} \\ &= \sup_{g \in M_b(\mathcal{S})} \left\{ \int_{\mathcal{S}} g d\mu - \log \int_{\mathcal{S}} e^{g} d\nu \right\}. \end{aligned}$$

For $g \in M_b(\mathcal{S})$ set $J(g) = \int_{\mathcal{S}} g d\mu - \log \int_{\mathcal{S}} e^g d\nu$. Thus $H(\mu\|\nu) \geq \sup_{g \in M_b} J(g)$. To obtain equality, it is enough to find a sequence $(g_M)_{M \in \mathbb{N}} \subset M_b(\mathcal{S})$ such that $\limsup_{M \to \infty} J(g_M) = H(\mu\|\nu)$. We distinguish two cases.

First case: μ is not absolutelty continuous with respect to ν. Then $H(\mu\|\nu) = \infty$ and there exists $A \in \mathcal{B}(\mathcal{S})$ such that $\mu(A) > 0$ while $\nu(A) = 0$. Choose such a set A and set $g_M = M \cdot 1_A$. Then, for every $M \in \mathbb{N}$, $g_M = 0$ ν-almost surely, thus $\int e^{g_M} d\nu = \int e^0 d\nu = 1$, hence $\log \int e^{g_M} d\nu = 0$. It follows that

$$\limsup_{M \to \infty} J(g_M) = \limsup_{M \to \infty} \int_{\mathcal{S}} g_M d\mu = \limsup_{M \to \infty} M \cdot \mu(A) = \infty.$$

Second case: μ is absolutely continuous with respect to ν. Then we can choose a measurable function $f : \mathcal{S} \to [0, \infty)$ such that f is a density for μ with respect to ν (f a version of the Radon-Nikodym derivative $d\mu/d\nu$), and $H(\mu\|\nu) = \int f \cdot \log(f) d\nu$, where the value of the integral is in $[0, \infty]$. Set

$$g_M(x) = \log(f(x)) \cdot 1_{[1/M, M]}(f(x)) - M \cdot 1_{\{0\}}(f(x)), \qquad x \in \mathcal{S}.$$

Then

$$\begin{aligned}
&\lim_{M\to\infty}\int_{\mathcal{S}} g_M d\mu \\
&= \lim_{M\to\infty}\int_{\mathcal{S}} f\cdot\log(f)\cdot 1_{[1/M,M]}(f)d\nu \\
&= \lim_{M\to\infty}\left(\int_{\mathcal{S}}\left(f\cdot\log(f)+1_{(0,\infty)}(f)\right)\cdot 1_{[1/M,M]}(f)d\nu-\int_{\mathcal{S}}1_{[1/M,M]}(f)d\nu\right) \\
&= \int_{\mathcal{S}} f\cdot\log(f)d\nu+\nu\{f>0\}-\nu\{f>0\} \\
&= H(\mu\|\nu)
\end{aligned}$$

by dominated convergence and monotone convergence since $t\cdot\log(t)\geq -1$ for every $t\geq 0$ and $t\cdot\log(t)=0$ if $t=0$, hence, for every $x\in\mathcal{S}$, $\left(f(x)\cdot\log(f(x))+1_{(0,\infty)}(f(x))\right)\cdot 1_{[1/M,M]}(f(x))\nearrow f(x)\cdot\log(f(x))+1_{(0,\infty)}(f(x))$ as $M\to\infty$. On the other hand, again usuing dominated and monotone convergence, respectively,

$$\begin{aligned}
&\lim_{M\to\infty}\log\int_{\mathcal{S}} e^{g_M}d\nu \\
&= \log\left(\lim_{M\to\infty}\int_{\mathcal{S}}\left(f\cdot 1_{[1/M,M]}(f)+1_{(0,1/M)\cup(M,\infty)}(f)+e^{-M}\cdot 1_{\{0\}}(f)\right)d\nu\right) \\
&= \log\int_{\mathcal{S}} f\nu=\log(1)=0.
\end{aligned}$$

It follows that $\limsup_{M\to\infty}J(g_M)=\lim_{M\to\infty}\int_{\mathcal{S}} g_M d\mu=H(\mu\|\nu)$. □

References

[1] O. Kallenberg. (2001), *Foundations of Modern Probability*, Probability and Its Applications Springer, New York, 2md edition.

[2] Kullback (1959), *Information theory and statistics*, John Wiley and sons, Inc, New York.

[3] R.S. Varadhan. (1966), Asymptotic probabilities and differential equaiotns. *Comm. Pure Appl. Math*, Vol **19**, 261-286.

[4] P. Dupuis and R. S. Ellis. (1997), *A Weak Convergence Approach to the Theory of Large Deviations.* Wiley Series in Probability and Statistics. John Wiley & Sons, New York.

[5] S. R. S. Varadhan. (1966), Asymptotic probabilities and differential equations. *Comm. Pure Appl. Math*, **19**: 261-286.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

Quantile-based dynamic cumulative entropy

Hooti, F. and Ahmadi, J.

Department of Statistics
Ferdowsi University of Mashhad, Mashhad, Iran
Email: ahmadi-j@um.ac.ir

Abstract

Recently, it has been shown by many authors that quantile functions are efficient and equivalent alternatives to distribution functions in modelling and analysis of statistical data (Nair et al. 2013). In this talk, the quantile function is recalled and some reliability measures are rewritten in terms of quantile function. A quantile-based Shannon entropy function introduced by Sunoj and Sankaran (2012). Here, we consider the cumulative entropy (Rao et al. 2004, Asadi and Zohrevand, 2007) and obtain the quantile-based dynamic cumulative entropy (QDCE). Some properties of QDCE are presented.

Keywords: Entropy, Quantile, Characterization, Exponential distribution.

References

[1] Asadi, M., Zohrevand, Y., (2007), On the Dynamic Cumulative Residual Entropy, *Journal of Statistical Planning and Inference*, **137**, 1931-1941.

[2] Nair, N.U., Sankaran, P.G., Vinesh Kumar, B., (2012), Modelling Lifetimes by Quantile Functions Using Parzen's Score Function, *Statistics: A Journal of Theoretical and Applied Statistics*, **46**, 799-811.

[3] Rao, M., Chen, Y., Vemuri, B.C., Wang, F., (2004). Cumulative Residual Entropy: a New Measure of Information, *IEEE Transactions on Information Theory*, **50**, 1220-1228.

[4] Sunoj, S.M., Sankaran, P.G., (2012). Quantile based Entropy Function, *Statistics and Probability Letters*, **82**, 1049-1053.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

Entropy maximization based on generalized Gini index

Khosravi Tanak, A. and Mohtashami Borzadaran, G. R.

Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran
Email: khosravi.a_66@yahoo.com

Abstract

In economics and social sciences, inequality measures such as Gini index, Peitra index etc., are commonly used to measure the evenness of probability distributions. In this paper, we first review some entropy maximization studies under moment and inequality measures constraints. Next, we consider a generalization of Gini index and based on the principle of maximum-entropy. We find the probability distribution that maximizes the entropy among all probability distributions supported on non-negative real values with a given mean and a given generalized Gini index.

Keywords: Maximum-entropy, Inequality measures, Euler's equation, Generalized Gini index.

1 Introduction

In economic and the social sciences, approximating income distribution with regard to inequality in society is of interest. In order to measure inequality, we need a scale of inequality to evaluate it. There are various known measures of inequality among them the Gini index is a famous and well-known measure, which is calculated based on Lorenz curve. It was proposed by Gini (1936) as a measure of inequality of income. A low Gini index indicates more equal income distribution, while a high Gini index indicates more unequal distribution. The Pietra index is another inequality measure that is most useful and appropriate in the case of asymmetric and skewed probability distributions. There is a generalization of Gini index proposed by Yitzhaki (1983) attaching different weight to the lower and upper ends of the distributions which is considered in this paper.

When approximating an unknown probability distribution, the question arises, what is the best approximation? Jaynes (1957) gave a general answer to this question: the best approach is to ensure that the approximation satisfies any constraints on the unknown distribution that we are aware of, and that subject to those constraints, the distribution should have maximum entropy. This is known as the maximum-entropy principle. The problem of maximizing entropy subject to some constraint have been studied by many researchers. Recently, some works have been done in the subject of entropy maximization based on inequality measures. Eliazar and Sokolov (2010) find distribution that maximize entropy subject to given mean and Gini index also they find distribution that maximize entropy subject to given mean and Pietra index. In this research, we intend to develop their results in terms of generalized Gini index.

In this paper, we consider continuous distributions. Section 2 contains some preliminaries and the basic tools which will be used in the next sections. In section 3, we review some results on entropy maximization under some moment constraints. In section 4, the results in terms of entropy maximization subject to some inequality measures constraints have been presented. Section 5 is devoted to our result in maximization of entropy with a given generalized Gini index.

2 Preliminaries

Let X be a random variable having a continuous cumulative density function (cdf) F with probability density function (pdf) f, then the basic uncertainty measure for distribution F is defined as

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

provided the integral exists.

One of the most well-known integral functional that has been studied in variational calculus is the Lagrange functional

$$L(y) = \int_{a}^{b} G(y(x), y'(x), x)dx, \tag{2}$$

where the given function G is continuous and has continuous first partial derivatives in each of its arguments. The simplest variational problem can be stated as follow: Find the curve $y = y(x)$ for which the functional $L(y)$ has an extremum. There is a result known as *Euler's equation* for this problem.

Theorem 2.1 (Gelfand and Fomin, p. 15)**.** *Let $L(y)$ be a functional of the form*

$$L(y) = \int_{a}^{b} G(y(x), y'(x), x)dx,$$

defined on the set of functions $y(x)$ which have continuous first derivatives in $[a, b]$ and satisfy the boundary conditions $y(a) = A$, $y(b) = B$. Then a necessary condition for $L(y)$ to have an extremum for a given function $y(x)$ is that $y(x)$ satisfy Euler's equation

$$\frac{\partial G}{\partial y} - \frac{d}{dx}\frac{\partial G}{\partial y'} = 0.$$

In several optimization problems, we require the optimal function to satisfy some constraints. In fact, suppose we are looking for an extremum of the functional (2) subject to the conditions $y(a) = A$, $y(b) = B$ and

$$\int_{a}^{b} J_i(y(x), y'(x), x)dx = l_i\ ,\ i = 1, 2, ..., m,$$

where $l_1, l_2, ..., l_m$ are constants. In this case a necessary condition for an extremum is that

$$\frac{\partial}{\partial y}\left(G + \sum_{i=1}^{m}\lambda_i J_i\right) - \frac{d}{dx}\frac{\partial}{\partial y'}\left(G + \sum_{i=1}^{m}\lambda_i J_i\right) = 0,$$

where $\lambda_1, \lambda_2, ..., \lambda_m$ are Lagrange multipliers.

3 Maximum entropy

As stated by Jaynes (1957), when an inference is made on the basis of incomplete information, it should be drawn from the probability distribution that maximizes the entropy subject to the constraints on the distribution. The resulting maximum entropy probability distribution corresponds to a distribution which is consistent with the given partial information but has maximum uncertainty or entropy associated with it. For illustrating Jayne's principle, we consider a continuous random variable X with probability density function $f(x)$ about which partial information in the form of first m moments is given. For obtaining the 'most objective' probability distribution of X, we must maximize the entropy $H(f)$ which is defined in (1) subject to the constraints

$$\begin{cases} \int_{-\infty}^{\infty} f(x)dx = 1 \\ \int_{-\infty}^{\infty} g_i(x)f(x)dx = \theta_i\ , i = 1, 2, ..., m \end{cases} \tag{3}$$

From *Euler's equation* in calculus of variation, one finds maximum entropy probability density function

$$f(x) = A\exp[-c_1g_1(x) - c_2g_2(x) - ... - c_mg_m(x)],$$

Where A, c_1, c_2, ..., c_m are to be obtain by using the constraints (3). In the following, some special cases is expressed.

3.1 If the range of the random variable X is $[0, 1]$

Within the class of probability distributions supported on the unit interval $[0, 1]$, the entropy maximizer is the uniform distribution.

3.2 If the range of the random variable X is $[0, \infty)$

Within the class of probability distributions supported on non-negative real numbers, and possessing a given mean, the entropy maximizer is the exponential distribution.

3.3 If the range of the random variable X is $(-\infty, \infty)$

1. In the constraints (3), if $g_1(x) = (x-a)^2$, where a is a fixed real number, then the maximum entropy density function is given by

   $$f(x) = \frac{1}{\sqrt{2\pi\theta_1}}\exp\left(-\frac{(x-a)^2}{2\theta_1}\right), -\infty < x < \infty.$$

   In fact, when second order moment about a is prescribed to be θ_1, the maximum entropy distribution is normal with mean a and variance θ_1.

2. In the constraints (3), if $g_1(x) = x$ and $g_2(x) = (x-\theta_1)^2$, that is when the mean (θ_1) and the variance θ_2 of X is prescribed, then the maximum entropy distribution is normal with mean θ_1 and variance θ_2.

4 Maximum entropy based on inequality measures

In this section, this question is answered: what happens when maximizing entropy subject to a given mean and a given measure of inequality? The answer is provided when inequality measure be dispersion, Gini index and Pietra index.

4.1 Dispersion

One the most basic approach to gauge statistical heterogeneity is the notion of dispersion: measuring the fluctuations of the probability distribution around its mean. The dispersion is given by the functional

$$D(f) = \left(\int_{-\infty}^{\infty} |x-\mu|^p dx\right)^{\frac{1}{p}}, \; p \geq 1,$$

The greater the dispersion, the more scattered and heterogeneous the probability distribution and the smaller the dispersion, the more concentrated and homogeneous the probability distribution. In the case $p = 2$, the dispersion equals the standard deviation, and the square dispersion equals the variance of distribution. The problem is maximizing entropy of X subject to the constraints

$$\begin{cases} \int_{-\infty}^{\infty} f(x)dx = 1 \\ \int_{-\infty}^{\infty} xf(x)dx = \mu \\ D(f) = \delta \end{cases} \tag{4}$$

Eliazar and Sokolov (2010) (ref. [1]) showed that the Subbotin's distribution with following density function has maximum entropy under above constraints

$$f(x) = \frac{\phi(p)}{\sigma} \exp\left(-\frac{1}{p}\left|\frac{x-\mu}{\sigma}\right|^p\right), -\infty < x < \infty,$$

where σ is a positive scale parameter and $\phi(p) = p^{1-1/p}/2\Gamma(1/p)$.

4.2 Gini index

The most widely used tool for analyzing and visualizing income inequality is the Lorenz curve which was developed by Lorenz (1905) and is defined as follows:

$$L(u) = \frac{1}{\mu}\int_0^u F^{-1}(x)dx, \quad 0 \le u \le 1, \tag{5}$$

where $F^{-1}(x) = inf\{t : F(t) \le x\},\ 0 \le t \le 1$. In fact, $L(u)$ denotes the fraction of total income that holders of the lowest uth fraction of income possess. Several indices of income inequality are directly related to this curve, most notably the Gini index. It define as twice the area between the Lorenz curve and the equality line:

$$G(f) = 2\int_0^1 (u - L(u))dx$$

Let X be a non-negative random variable. It can be easily show that

$$G(f) = 1 - \frac{1}{\mu}\int_0^\infty \bar{F}^2(x)dx,$$

where $\bar{F}$ is survival function of X. Now, the question is what happens when maximizing entropy subject to a given mean and a given Gini index? In other words, the problem is maximizing the entropy of X subject to the constraints

$$\begin{cases} \int_{-\infty}^{\infty} f(x)dx = 1 \\ \int_{-\infty}^{\infty} xf(x)dx = \mu \\ G(f) = \gamma \end{cases}$$

Eliazar (2010) showed that the survival function of maximum entropy distribution is given by

$$\bar{F}(x) = \frac{1}{\sigma\exp(\rho x) + (1-\sigma)}, x \ge 0,$$

where σ is positive valued parameter depending on γ and $\rho = \frac{\ln\sigma}{(\sigma-1)\mu}$.

4.3 Peitra index

A second important inequality measure is the Pietra index, which is defined as the maximal vertical deviation between the Lorenz curve and the equality line

$$P(f) = \max_{0 \le u \le 1}\{u - L(u)\}.$$

Eliazar and Sokolov (2010), ref. [2], proposed an alternative formula for the Pietra index as follow:

$$P(f) = \frac{1}{\mu}\int_0^\infty \max\{0, x-\mu\}f(x)dx,$$

and showed that within the class of probability density functions possessing a given mean and a given Peitra index that is under following conditions

$$\begin{cases} \int_{-\infty}^{\infty} f(x)dx = 1 \\ \int_{-\infty}^{\infty} xf(x)dx = \mu \\ P(f) = \eta \end{cases}$$

the entropy maximizer is bi-exponential probability density function

$$f(x) = \begin{cases} c_1 \exp(\alpha x) & \text{if } 0 < x < \mu, \\ c_2 \exp(-\beta x) & \text{if } \mu < x < \infty, \end{cases}$$

where α and β are real exponents depending on μ and η; c_1 and c_2 are normalizing coefficients satisfying the relation $\ln(c_2/c_1) = (\alpha + \beta)\mu$.

5 Maximum entropy based on generalized Gini index

There are several generalizations of the Gini index proposed in the literature. An important generalization of the Gini index was proposed by Yitzhaki (1983).

$$G_\nu(f) = 1 - \int_0^1 \nu(\nu-1)(1-u)^{\nu-2} L(u)du, \ \nu > 1. \tag{6}$$

If $\nu = 2$ we obtain the Gini index. When ν increases, higher weights are attached to small incomes. We want to find the probability distribution that maximize the entropy under following constraints:

$$\begin{cases} \int_0^{\infty} f(x)dx = 1 \\ \int_0^{\infty} xf(x)dx = \mu \\ G_\nu(f) = \delta \end{cases} \tag{7}$$

Theorem 5.1. *The survival function of entropy maximizer distribution subject to a given mean and a given generalized Gini index is*

$$\bar{F}(x) = \left(\frac{1}{c_1 \exp(c_2 x) + (1 - c_1)} \right)^{\frac{1}{\nu-1}}, \ x \geq 0, \tag{8}$$

where c_1 and c_2 obtaine from constraints (7).

Proof. Using the definitions of Generalized Gini index (6) and Lorenz curve (5) we have

$$\begin{aligned} G_\nu(f) &= 1 - \frac{1}{\mu} \int_0^1 \int_0^p \nu(\nu-1)(1-p)^{\nu-2} F^{-1}(t)dtdp \\ &= 1 - \frac{1}{\mu} \int_0^1 \int_t^1 \nu(\nu-1)(1-p)^{\nu-2} F^{-1}(t)dpdt \\ &= 1 - \frac{1}{\mu} \int_0^1 \nu(1-t)^{\nu-1} F^{-1}(t)dt \\ &= 1 - \frac{1}{\mu} \int_0^{\infty} \nu x f(x) \bar{F}^{\nu-1}(x)dx \\ &= 1 - \frac{1}{\mu} \int_0^{\infty} \bar{F}^{\nu}(x)dx. \end{aligned}$$

The maximization problem is to minimize the convex functional

$$\int_0^{\infty} f(x) \ln f(x)dx,$$

subject to constraints

$$\begin{cases} \int_0^\infty f(x)dx = 1 \\ \int_0^\infty xf(x)dx = \mu \\ \int_0^\infty \bar{F}^\nu(x)dx = \eta \end{cases}$$

By theorem 2.1 and remark 2, the functional

$$\begin{aligned} L(f,\lambda) &= \int_0^\infty f(x)\ln f(x)dx + \lambda_1 \int_0^\infty f(x)dx \\ &\quad + \lambda_2 \int_0^\infty xf(x)dx + \lambda_3 \int_0^\infty \bar{F}^\nu(x)dx, \end{aligned}$$

must satisfies the Euler's equation (3). So we have

$$\begin{aligned} & \frac{f'(x)}{f(x)} + \lambda_2 + \lambda_3\nu\bar{F}^{\nu-1}(x) = 0 \\ \Leftrightarrow \quad & f'(x) + \lambda_2 f(x) + \lambda_3\nu f(x)\bar{F}^{\nu-1}(x) = 0 \\ \Leftrightarrow \quad & -\bar{F}''(x) - \lambda_2\bar{F}'(x) - \lambda_3\left(\bar{F}^\nu(x)\right)' = 0 \\ \Leftrightarrow \quad & \bar{F}'(x) + \lambda_2\bar{F}(x) + \lambda_3\bar{F}^\nu(x) = c, \end{aligned}$$

where c is a constant. Since $\bar{F}(x)$ is a survival probability function c must be equal zero. Thus, we arrive at the Bernoulli equation:

$$\bar{F}'(x) + \lambda_2\bar{F}(x) + \lambda_3\bar{F}^\nu(x) = 0.$$

The solution of this equation is survival function in (9). Since target functional is convex and constraints functionals are linear or convex, a global maximum is attained at the critical point $\bar{F}(x)$. □

References

[1] Eliazar, I. and Sokolov, M. (2010), Maximization of statistical heterogeneity: From Shannon's entropy to Gini's index, *Physica A*, **389**, 3023-3038.

[2] Eliazar, I. and Sokolov, M. (2010), Measuring statistical heterogeneity: The Pietra index, *Physica A*, **389**, 117-125.

[3] Gelfand, I. and Fomin, S. (1963), *Calculus of variations*, Prentice-Hall, Inc.

[4] Gini, C. (1936), On the measure of concentration with special reference to income and statistics, *Colorado College Publication*, **208**, 73-79.

[5] Jaynes, E.T. (1957), Information theory and statistical mechanics. *Physical Review*, **106**, 620-630.

[6] Lorenz, M. O. (1905), Methods of measuring the concentration of wealth. *Quarterly Publications of the American Statistical Association*, **9 (70)**, 209-219.

[7] Yitzhaki, S. (1983), On an extension of the Gini inequality index. *International Economic Review*, **24**, 617-628.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

On dynamic mutual informations and cumulative Kullback-Leibler discriminations

Longobardi, M.

Dipartimento di Matematica e Applicazioni "R. Caccioppoli"
Universita' degli Studi di Napoli FEDERICO II
Via Cintia - 80126 Napoli, Italy
Email: maria.longobardi@unina.it

Abstract

We consider dynamic mutual information of lifetime distributions and study this measure for bivariate past and residual lifetimes; some bounds are obtained and examples are given. We focus also on dynamic mutual information for TTE and truncated TTE models. The mutual information between the minimum and the maximum of order statistics is considered and a copula-based approach for this measure is presented. In the second part some properties of a new measure of discrimination obtained from Kullback-Leibler discrimination measure are described. A dynamic version of this measure is also proposed, and it is applied to some concepts of relative aging. Finally, we provide an application to image analysis.

Keywords: Entropy, Mutual information, Bivariate lifetimes, Time-transform exponential model.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

Parameter estimation by Kullback-Leibler divergence of Survival functions with application to censored data

Mehrali, Y. and Asadi, M.

Department of Statistics, University of Isfahan, Isfahan, 81744, Iran
Email: yasermehrali@gmail.com

Abstract

In this paper, we study estimation of parameters based on survival functions. We consider equilibrium distributions in Kullback-Leibler divergence and find a new measure of divergence. Then we use this measure in parameter estimation. Some extensions in discrete, censor and real numbers support cases also investigated.

Keywords: Censored data, Entropy, Estimation, Equilibrium distributions, Information measures.

1 Introduction

The Kullback-Leibler divergence or relative entropy is a measure of distance between two probability distribution. If X and Y have probability density functions f and g respectively, the K-L divergence of f relative to g is defined by

$$D\left(f||g\right)=\int\limits_{\mathbb{R}} f\left(x\right)\ln\frac{f\left(x\right)}{g\left(x\right)}dx.$$

$D\left(f||g\right)$ is always nonnegative and if $f=g$ $a.s.$, then $D\left(f||g\right)=0$.

Let f belongs to parametric family with k-dimensional parameter vector $\boldsymbol{\theta}\in\boldsymbol{\Theta}\subset\mathbb{R}^k$ and f_n be kernel density estimation of f. Lindsay (1994) used K-L divergence of f_n relative to f as

$$D\left(f_n||f\right)=\int f_n\left(x\right)\ln\frac{f_n\left(x\right)}{f\left(x;\boldsymbol{\theta}\right)}dx, \tag{1}$$

and defined minimum K-L divergence estimator of $\boldsymbol{\theta}$ as

$$\widehat{\boldsymbol{\theta}}=\underset{\boldsymbol{\theta}}{\arg\min}D\left(f_n\left(x\right)||f\left(x;\boldsymbol{\theta}\right)\right).$$

Many authors such as Morales et. al. (1995), Jiménz and Shao (2001), Broniatowski and Keziou (2009), Broniatowski (2011) and Cherfi, M. (2011-a, b, 2012) studied properties of minimum K-L divergence estimators.

Definition of $D\left(f_n||f\right)$ is based on f which in general may or may not exist, and f_n which even if the number of samples tends to infinity, there is no guarantee that converges to its true measure. So Liu

(2007) proposed using $\bar{F}$ instead of f and for guarantying that defined measure is nonnegative added a term and defined K-L divergence of Survival functions $\bar{F}_n$ and $\bar{F}$ by

$$\begin{aligned} KLS\left(\bar{F}_n||\bar{F}\right) &= \int_0^\infty \bar{F}_n(x)\ln\frac{\bar{F}_n(x)}{\bar{F}(x;\boldsymbol{\theta})} - \left[\bar{F}_n(x) - \bar{F}(x;\boldsymbol{\theta})\right]dx \\ &= \int_0^\infty \bar{F}_n(x)\ln\bar{F}_n(x)\,dx - \int_0^\infty \bar{F}_n(x)\ln\bar{F}(x;\boldsymbol{\theta})\,dx \\ &\quad -\left[\bar{x} - E(X)\right], \end{aligned} \tag{2}$$

where $\bar{F}_n$ is empirical estimator of $\bar{F}$. Now if consider the parts of $KLS\left(\bar{F}_n||\bar{F}\right)$ that depend on $\boldsymbol{\theta}$ and define

$$g(\boldsymbol{\theta}) = E(X) - \int_0^\infty \bar{F}_n(x)\ln\bar{F}(x;\boldsymbol{\theta})\,dx, \tag{3}$$

then minimum KLS estimator of $\boldsymbol{\theta}$ defined as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min} KLS\left(\bar{F}_n(x)\,||\bar{F}(x;\boldsymbol{\theta})\right) = \underset{\boldsymbol{\theta}}{\arg\min} g(\boldsymbol{\theta}).$$

Liu (2007) applied this estimator in uniform and exponential models and Yari and Saghafi (2012) and Yari et. al. (2013) applied for estimating parameters of weibull distribution.

As mentioned above, Liu added term $[\bar{x} - E(X)]$ to guaranty $D\left(\bar{F}_n||\bar{F}\right)$ is nonnegative. Here we consider another approach to aim this. We consider equilibrium distributions (Nair, 2013) instead of f_n and f in 1 as

$$f^*(x) = \frac{\bar{F}(x)}{E(X)}, \text{ and } f_n^*(x) = \frac{\bar{F}_n(x)}{\bar{x}}. \tag{4}$$

In next section, we define K-L divergence based on equilibrium distributions and use it to estimate parameters of distributions. Some extensions in discrete, censor and real numbers support cases also investigated in following.

2 Main results

Using equilibrium distributions instead of f_n and f , we define K-L divergence of equilibrium distributions as follow.

Definition 1. *Let $\bar{F}(x;\boldsymbol{\theta})$ be the true survival function with unknown parameters $\boldsymbol{\theta}$ and $\bar{F}_n(x)$ be the empirical survival function of a random sample of size n from $F(x;\boldsymbol{\theta})$. Define the Kullback-Leibler divergence of equilibrium distributions (KLE) by*

$$\begin{aligned} KLE(F_n||F) &= D\left(f_n^*||f^*\right) \\ &= \int_0^\infty \frac{\bar{F}_n(x)}{\bar{x}}\ln\frac{\bar{F}_n(x)/\bar{x}}{\bar{F}(x;\boldsymbol{\theta})/E(X)}dx \\ &= \frac{1}{\bar{x}}\int_0^\infty \bar{F}_n(x)\ln\frac{\bar{F}_n(x)}{\bar{F}(x;\boldsymbol{\theta})}dx \\ &\quad -\left[\ln\bar{x} - \ln E(X)\right]. \end{aligned} \tag{5}$$

Theorem 2.1. *The introduced measure is non negative and as n tends to infinity, it converges to zero.*

The KLE divergence is good enough for our purposes. Consider the parts of $KLE(F_n||F)$ that depend on $\boldsymbol{\theta}$ and define

$$g^*(\boldsymbol{\theta}) = \ln E(X) - \frac{1}{\bar{x}}\int_0^\infty \bar{F}_n(x)\ln\bar{F}(x;\boldsymbol{\theta})\,dx. \tag{6}$$

Now we define minimum KLE divergence estimator.

Let

$$\widehat{\boldsymbol{\theta}} = \arg\underset{\boldsymbol{\theta}}{\min} KLE\left(F_n\left(x\right)||F\left(x;\boldsymbol{\theta}\right)\right) = \arg\underset{\boldsymbol{\theta}}{\min} g^*\left(\boldsymbol{\theta}\right).$$

Then, $\widehat{\boldsymbol{\theta}}$ called minimum KLE divergence estimator of $\boldsymbol{\theta}$.

Yari et. al. (2013) find a simple form of 3 as

$$g\left(\boldsymbol{\theta}\right) = E\left(X\right) - \frac{1}{n}\sum_{i=1}^{n} h\left(x_i\right), \tag{7}$$

where

$$h\left(x\right) = \int_0^x \ln \bar{F}\left(y;\boldsymbol{\theta}\right) dy. \tag{8}$$

So, comparing $KLE\left(F_n||F\right)$ in 5 with $KLS\left(\bar{F}_n||\bar{F}\right)$ in 2, we can write simple form of 6 as

$$g^*\left(\boldsymbol{\theta}\right) = \ln E\left(X\right) - \frac{1}{n\bar{x}}\sum_{i=1}^{n} h\left(x_i\right) \tag{9}$$

where $h\left(x\right)$ is same as 8.

3 Some Extensions

In this section, we find some extensions minimum KLE divergence estimator. These are in discrete, censor and real numbers support cases.

3.1 Discrete Case

Results is discrete case are straight forward of continuous case. Considering approach of Liu $g\left(\boldsymbol{\theta}\right)$ is same as 7 with

$$h\left(x\right) = \sum_{y=0}^{x-1} \ln \bar{F}\left(y;\boldsymbol{\theta}\right). \tag{10}$$

So, in our approach $g^*\left(\boldsymbol{\theta}\right)$ is same as 9 with $h\left(x\right)$ same as 10.

Example 2. *Let $\{X_1, \ldots, X_n\}$ be sequence of i.i.d. Geometric random variables with probability mass function*

$$P\left(X = x\right) = q^x p, \qquad x = 0, 1, \ldots .$$

It can be showed that using Liu method we have

$$\widehat{p} = \frac{\sqrt{1 + 2\left(\overline{x^2} + \overline{x}\right)} - 1}{\overline{x^2} + \overline{x}},$$

and using our method we have $\widehat{p} = 2\overline{x}/\left(\overline{x^2} + \overline{x}\right)$. So, it seems that our approach yields simpler estimator than Liu's one.

3.2 Censor Case

Let $T_1, ..., T_n$ be survival times that are i.i.d. nonnegative random variables from a c.d.f. F and survival function F_n, and $C_1, ..., C_n$ be i.i.d. nonnegative random variables independent of T_i's. In a variety of applications in biostatistics and life-time testing, we are only able to observe the smaller of T_i and C_i and an indicator of which variable is smaller:

$$X_i = \min\left\{T_i,\ C_i\right\}, \qquad \delta_i = I_{(0,C_i)}(T_i), \qquad i = 1, ..., n.$$

This is called a random censorship model and C_i′s are called censoring times. Let $x_{(1)} \leq \cdots \leq x_{(n)}$ be ordered values of X_i′s and $\delta_{(i)}$ be the δ-value associated with $x_{(i)}$. A maximum empirical likelihood estimator (MELE) of $\overline{F}$ can be written as

$$\bar{F}_n(t) = \prod_{X_{(i)} \leq t} \left(1 - \frac{\delta_{(i)}}{n-i+1}\right), \tag{11}$$

which is the well-known Kaplan-Meier (1958) product-limit estimator (see Shao (2003) for more details). If we consider $\bar{F}_n$ in 11 instead of that one in 4, then considering Liu approach, after some algebra we have

$$g(\boldsymbol{\theta}) = E(T) - \sum_{i=0}^{n} \prod_{j=1}^{i} \left(1 - \frac{\delta_{(j)}}{n-j+1}\right) \left[h\left(x_{(i+1)}\right) - h\left(x_{(i)}\right)\right],$$

and considering our approach we have

$$g^*(\boldsymbol{\theta}) = \ln E(T) - \frac{1}{\bar{x}} \sum_{i=0}^{n} \prod_{j=1}^{i} \left(1 - \frac{\delta_{(j)}}{n-j+1}\right) \left[h\left(x_{(i+1)}\right) - h\left(x_{(i)}\right)\right],$$

where $h(x)$ is same as 8.

3.3 Real number support Case

When support of random variable is $\mathbb{R}$, if $E(X) \neq 0$ and $\bar{x} \neq 0$ we define equilibrium distributions as

$$\begin{aligned} f^*(x) &= \frac{F(x) I(x<0) + \bar{F}(x) I(x \geq 0)}{E(X)}, \\ f_n^*(x) &= \frac{F_n(x) I(x<0) + \bar{F}_n(x) I(x<0)}{\bar{x}}. \end{aligned}$$

So, in our approach we find that

$$\begin{aligned} KLE(F_n||F) &= \int_{-\infty}^{\infty} f_n(x) \ln \frac{f_n^*(x)}{f^*(x;\boldsymbol{\theta})} dx \\ &= \frac{1}{\bar{x}} \int_{-\infty}^{0} F_n(x) \ln \frac{F_n(x)}{F(x;\boldsymbol{\theta})} dx \\ &\quad + \frac{1}{\bar{x}} \int_{0}^{\infty} \bar{F}_n(x) \ln \frac{\bar{F}_n(x)}{\bar{F}(x;\boldsymbol{\theta})} dx - [\ln \bar{x} - \ln E(X)]. \end{aligned}$$

Now, the parts of $KLE(F_n||F)$ that depend on $\boldsymbol{\theta}$ is

$$\begin{aligned} g^*(\boldsymbol{\theta}) &= \ln E(X) - \frac{1}{\bar{x}} \int_{-\infty}^{0} F_n(x) \ln F(x;\boldsymbol{\theta}) dx \\ &\quad - \frac{1}{\bar{x}} \int_{0}^{\infty} \bar{F}_n(x) \ln \bar{F}(x;\boldsymbol{\theta}) dx. \end{aligned} \tag{12}$$

Now let in observed sample $\{x_1, x_2, ..., x_n\}$, we observe k of them negative and $n-k$ of them nonnegative. After some algebra we can see that $g^*(\boldsymbol{\theta})$ in 12 has the simple form as

$$g^*(\boldsymbol{\theta}) = \ln E(X) - \frac{1}{n\bar{x}} \sum_{\substack{i=1 \\ x_i<0}}^{k} u(x_i) - \frac{1}{n\bar{x}} \sum_{\substack{i=k+1 \\ x_i \geq 0}}^{n} h(x_i), \tag{13}$$

where $h(x)$ is same as 8 and

$$u(x) = \int_{x}^{0} \ln \bar{F}(y;\boldsymbol{\theta}) dy. \tag{14}$$

If support be discrete, we have

$$g^*(\boldsymbol{\theta}) = \ln E(X) - \frac{1}{\bar{x}} \sum_{-\infty}^{-1} F_n(x) \ln F(x;\boldsymbol{\theta}) - \frac{1}{\bar{x}} \sum_{0}^{\infty} \bar{F}_n(x) \ln \bar{F}(x;\boldsymbol{\theta}).$$

which has the simple form same as 13 with $h(x)$ same as 10 and

$$u(x) = \sum_{y=x}^{-1} \ln \bar{F}(y;\boldsymbol{\theta}). \tag{15}$$

Similarly, considering Liu approach we have

$$g(\boldsymbol{\theta}) = E(X) - \int_{-\infty}^{0} F_n(x) \ln F(x;\boldsymbol{\theta})\, dx - \int_{0}^{\infty} \bar{F}_n(x) \ln \bar{F}(x;\boldsymbol{\theta})\, dx, \tag{16}$$

in continuous case and

$$g(\boldsymbol{\theta}) = E(X) - \sum_{-\infty}^{-1} F_n(x) \ln F(x;\boldsymbol{\theta}) - \sum_{0}^{\infty} \bar{F}_n(x) \ln \bar{F}(x;\boldsymbol{\theta}), \tag{17}$$

in discrete case, which both $g(\boldsymbol{\theta})$ in 16 and 17 have the simple form as

$$g(\boldsymbol{\theta}) = E(X) - \frac{1}{n} \sum_{\substack{i=1 \\ x_i<0}}^{k} u(x_i) - \frac{1}{n} \sum_{\substack{i=k+1 \\ x_i\geq 0}}^{n} h(x_i),$$

where depend on continuous or discrete case $h(x)$ is same as 8 or 10, and $u(x)$ is same as 14 or 15.

Example 3. *Let $\{X_1, \ldots, X_n\}$ be sequence of i.i.d. Normal random variables with probability density function*

$$\phi\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \qquad x \in \mathbb{R}.$$

In this case, we see that $h(x)$, $u(x)$, $g^*(\boldsymbol{\theta})$ and $g^*(\boldsymbol{\theta})$ don't have close form. But after derivation $g^*(\boldsymbol{\theta})$ respect to μ and setting zero we have

$$\frac{n}{\sigma}\frac{\overline{x}}{\mu} - \sum_{\substack{i=1 \\ x_i<0}}^{k} \ln \Phi\left(\frac{x_i-\mu}{\sigma}\right) + \sum_{\substack{i=k+1 \\ x_i\geq 0}}^{n} \ln \Phi\left(\frac{\mu-x_i}{\sigma}\right) + k \ln \Phi\left(-\frac{\mu}{\sigma}\right) - (n-k) \ln \Phi\left(\frac{\mu}{\sigma}\right) = 0, \tag{18}$$

and after derivation $g(\boldsymbol{\theta})$ respect to μ and setting zero we have

$$\frac{n}{\sigma} - \sum_{\substack{i=1 \\ x_i<0}}^{k} \ln \Phi\left(\frac{x_i-\mu}{\sigma}\right) + \sum_{\substack{i=k+1 \\ x_i\geq 0}}^{n} \ln \Phi\left(\frac{\mu-x_i}{\sigma}\right) + k \ln \Phi\left(-\frac{\mu}{\sigma}\right) - (n-k) \ln \Phi\left(\frac{\mu}{\sigma}\right) = 0. \tag{19}$$

Also after derivation each one of $g(\boldsymbol{\theta})$ or $g^*(\boldsymbol{\theta})$ respect to σ and setting zero we have

$$\sum_{\substack{i=1 \\ x_i<0}}^{k} \int_{\frac{x_i-\mu}{\sigma}}^{-\frac{\mu}{\sigma}} \frac{z\phi(z)}{\Phi(z)} dz - \sum_{\substack{i=k+1 \\ x_i\geq 0}}^{n} \int_{\frac{\mu-x_i}{\sigma}}^{\frac{\mu}{\sigma}} \frac{z\phi(z)}{\Phi(z)} dz = 0. \tag{20}$$

Here we see that 20 is same in both approach, and also see that 18 and 19 are just different is term $\frac{\overline{x}}{\mu}$ which tends to 1 as n tends to infinity. So in this case estimators are near to each other.

Example 4. *Let $\{X_1, \ldots, X_n\}$ be sequence of i.i.d. Pareto random variables with probability density function*

$$\phi\left(x;\alpha,\beta\right)=\frac{\alpha\beta^{\alpha}}{x^{\alpha+1}},\qquad x\geq\beta,\ \alpha>0,\ \beta>0.$$

So after some algebra,we have

$$g^{*}\left(\alpha,\beta\right)=\ln\alpha+\ln\beta-\ln\left(\alpha-1\right)+\frac{\alpha\overline{x\ln x}}{\overline{x}}-\alpha\left(\ln\beta+1\right)+\frac{\alpha\beta}{\overline{x}},$$

and

$$g\left(\alpha,\beta\right)=\frac{\alpha\beta}{\alpha-1}+\alpha\overline{x\ln x}-\alpha\overline{x}\left(\ln\beta+1\right)+\alpha\beta.$$

It can be shown that after derivation each one of $g\left(\alpha,\beta\right)$ or $g^{*}\left(\alpha,\beta\right)$ respect to α and β and setting zero we have

$$\ln\alpha-\ln\left(\alpha-1\right)-\frac{1}{\alpha-1}+\frac{\overline{x\ln x}}{\overline{x}}-\ln\overline{x}=0, \tag{21}$$

and

$$\beta=\frac{\overline{x}\left(\alpha-1\right)}{\alpha}. \tag{22}$$

Here we see that 21 and 22 are same in both approach. So in this case estimators are same as each other.

References

[1] Broniatowski, M. (2011). Minimum divergence estimators, maximum likelihood and exponential families. *arXiv preprint arXiv:*1108.0772.

[2] Broniatowski, M., & Keziou, A. (2009). Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis,* **100(1)**, 16-36.

[3] Cherfi, M. (2011a). Dual ϕ-divergences estimation in normal models. *arXiv preprint arXiv:*1108.2999.

[4] Cherfi, M. (2011b). On Bayesian Estimation via Divergences. *arXiv preprint arXiv:*1112.5854.

[5] Cherfi, M. (2012). Dual divergences estimation for censored survival data. *Journal of Statistical Planning and Inference,* **142(7)**, 1746-1756.

[6] Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory.* John Wiley & Sons.

[7] Jiménz, R., & Shao, Y. (2001). On robustness and efficiency of minimum divergence estimators. *Test*, **10(2)**, 241-248.

[8] Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.*, **53**, 457-481.

[9] Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The Annals of Statistics*, **22(2)**, 1081-1114.

[10] Liu, J. (2007). *Information theoretic content and probability.* Ph.D. Thesis, University of Florida.

[11] Morales, D., Pardo, L., & Vajda, I. (1995). Asymptotic divergence of estimates of discrete distributions. *Journal of Statistical Planning and Inference*, **48(3)**, 347-369.

[12] Nair, N. U., Sankaran, P., & Balakrishnan, N. (2013). *Quantile-based reliability analysis:* Springer.

[13] Shao, J. (2003) *Mathematical Statistics.* Springer, New York, USA.

[14] Yari, G., Mirhabibi, A., & Saghafi, A. (2013). Estimation of the Weibull parameters by Kullback-Leibler divergence of Survival functions. *Appl. Math*, **7(1)**, 187-192.

[15] Yari, G., & Saghafi, A. (2012). Unbiased Weibull Modulus Estimation Using Differential Cumulative Entropy. *Communications in Statistics-Simulation and Computation*, **41(8)**, 1372-1378.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

Entropy and information (divergence) measures

Mohtashami Borzadaran, G. R.

Department of Statistics, Faculty of Mathematical Sciences
Ferdowsi University of Mashhad, Mashhad, Iran
Email: gmb1334@yahoo.com

The extension of notion for the measure of information with application in communication theory back to the experiences of the C. E. Shannon during the second World War II. In 1948, he introduced that the entropy is a real number associated with a random variable which is equal to the expected value of the surprise that we would receive upon getting a realization of the random variable. Let, $S_a(p) = -\log_a p$ (base-2 is often used) be the measure of surprise we feel when an event with the probability p is occurring actually occurs. Then, entropy for a random variable is calculated using the probability mass (or pdf) function of the random variable via $H_a(X) = E[S_a(p(X))]$. Some of the properties and characterizations of the Shannon entropy and its extension versions are mentioned here.

Also, finding expressions for the multivariate distributions (discrete or continuous) and information measure such as mutual information with some of their properties and discussing in view of copula are reviewed.

The principle of maximum entropy provides a method to select the unknown pdf (or pmf) compatible to entropy under a specified constraint. This idea was introduced by Jaynes 1957 and obtained via a theorem by Kagan et al. (1973). Applying to maximum Renyi or Tsallis entropy and also $\phi-$entropy, as a general format subsume many special cases. Similar arguments are applicable to a multivariate set-up.

In probability theory and information theory, the Kullback Leibler divergence (also information divergence, information gain, relative entropy) is a non-symmetric measure of the difference between two probability distributions. Specifically, the Kullback Leibler divergence is typically represents the ”true” distribution of data and a theoretical model for approximation of the true distribution. Although it is often intuited as a metric or distance, the KL divergence is not a true metric. Various applications in statistics and properties of it is one of our aim in here. The link between maximum likelihood and maximum entropy and Kullback Leibler information is important for a discussion which is coming in this note. There are several types of information divergence measure that are studied in literature as extensions of the Shannon entropy and Kullback Leibler information. Some of them can be collected in Csiszar $\phi-$divergence as special cases. So, minimization of them is important and finding these optimal measures is the other direction that is discussed in this paper with the related special states such as Kullback Leibler information, χ^2-divergence, total variation, squared perimeter distance, Renyi divergence, Hellinger distance, directed divergence and so on.

Keywords: Entropy, Maximum entropy, Kullback Leibler information, Information measures, Minimization of Kullback Leibler information.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

A measure of relative entropy rate between two stochastic processes with an application in speech recognition

Nikooravesh, Z.

Department of Basic Science, Birjand University of Technology, Birjand, 97175-569, Iran
Email: nikooravesh@birjandut.ac.ir

Abstract

In this paper, we study the relative entropy rate between a homogeneous Markov chain and a hidden Markov chain defined by observing the output of a discrete stochastic channel whose input is the finite state space homogeneous stationary Markov chain. For this purpose, we obtain the relative entropy between two finite subsequences of above mentioned chains and define the relative entropy rate between these stochastic processes, then calculate the maximum of the relative entropy rate by the concept of convexity and study the convergence of it.

Keywords: Relative entropy rate, Mutual information, Stochastic channel, Markov chain, Hidden Markov chain.

Introduction

Suppose $\{X_n\}_{n\in\mathbf{N}}$ is a homogeneous stationary Markov chain with finite state space $S = \{0, 1, 2, ..., N-1\}$ and $\{Y_n\}_{n\in\mathbf{N}}$ is a hidden Markov chain (HMC) which is observed through a discrete stochastic channel where the input of channel is the Markov chain. The output state space of channel is characterized by channel's statistical properties. From now on we study the channels state spaces which have been equal to the state spaces of input chains.

Let $\mathbf{P} = \{p_{ab}\}$ be the one-step transition probability matrix of the Markov chain such that $p_{ab} = Pr\{X_n = b|X_{n-1} = a\}$ for $a, b \in S$ and $\mathbf{Q} = \{q_{ab}\}$ be the noisy matrix of channel where $q_{ab} = Pr\{Y_n = b|X_n = a\}$ for $a, b \in S$. Also the initial distribution of the Markov chain is denoted by the vector $\mathbf{\Pi}_0$ such that $\mathbf{\Pi}_0(i) = Pr\{X_0 = i\}$ for $i \in S$.

At the rest of this paper we try to obtain the relative entropy and mutual information between two finite subsequences $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$ and to define the relative entropy rate and mutual information rate between a Markov chain and its corresponding hidden Markov chain. From now on X_1^n denotes the subsequence $X_1, X_2, ..., X_n$ for simplicity.

Relative entropy was first defined by Kullback and Leibler [8]. It is known under a variety of names, including the Kullback-Leibler distance, cross entropy, information divergence, and information for discrimination, and it has been studied in detail by Csiszar [6] and Amari [1]. The relative entropy between two random variables is developed to two sequences of variables and it is used for comparing two stochastic processes. Kesidis and Walrand derived the relative entropy between two Markov transition rate matrices [7]. Chazottes, Giardina and Redig [4] applied it for comparing two Markov chains.

Hidden Markov processes (HMP)s were introduced in full generality in 1966 by Baum and Petrie [2] who referred to them as probabilistic functions of Markov chains. Indeed, the observation sequence depends probabilistically on the Markov chain. During 1966-1969, Baum and Petrie studied statistical properties of stationary ergodic finite-state space HMPs. They developed an ergodic theorem for almost-sure convergence of the relative entropy density of one HMP with respect to another. In 1970, Baum, Petrie, Soules and Weiss developed forward-backward recursions for calculating the conditional probability of a state given an observation sequence from a general HMP [3].

In this paper the relative entropy rate between a Markov chain and a HMC is studied. It is possible by considering properties of channel to have many hidden Markov chains respecting a Markov chain. So conditions of the system whose works are based on the hidden Markov models will be controlled by noting the relative entropy rate. Section 1 includes some required preliminaries and definitions. Section 2 discusses the existance and convergence of the relative entropy rate, by maximum of it wich obtain by the mutual information rate. At last in section 3 we show that relative entropy rate has an application in speech recognition.

1 Preliminaries

In probability theory, entropy and mutual information are introduced by Shannon [9]. The entropy of a random variable X by distribution P_X taking values from a finite set E is defined by him as

$$H(X) = -E_X \log P(X) = -\sum_{i\in E} P_X(i)\log P_X(i), \tag{1}$$

with the convention $0\log 0 = 0$. Consider two random variables X and Y with joint distribution $P_{X,Y}(x,y)$. The entropy of these variables is

$$\begin{aligned} H(X,Y) &= -E_{X,Y}\log P_{X,Y}(X,Y) \\ &= -\sum_{i\in E}\sum_{j\in E} P_{X,Y}(i,j)\log P_{X,Y}(i,j). \end{aligned} \tag{2}$$

Also the conditional entropy could be defined as

$$\begin{aligned} H(X|Y) &= -E_{X,Y}\log P_{X|Y}(X|Y) \\ &= -\sum_{i\in E}\sum_{j\in E} P_{X,Y}(i,j)\log P_{X|Y}(i|j). \end{aligned} \tag{3}$$

In statistics, the relative entropy arises as an expected logarithm of the likelihood ratio of the distribution probability functions of these variables i.e.

$$\begin{aligned} D(P_X||P_Y) &= E_X \log\frac{P_X(X)}{P_Y(X)} \\ &= \sum_{i\in E} P_X(i)\log\frac{P_X(i)}{P_Y(i)}. \end{aligned} \tag{4}$$

Corollary 1. *Two variables X and Y are identical distribution if and only if $D(P_X||P_Y) = 0$.*

For relative entropy, one can write

$$D(P_{X_1,X_2}||P_{Y_1,Y_2}) = D(P_{X_1}||P_{Y_1}) + D(P_{X_2|X_1}||P_{Y_2|Y_1}). \tag{5}$$

Proof. [5] pages 24-25.

This equation is known as the chain rule for relative entropy. The mutual information $I(X;Y)$ is the relative entropy between the joint distribution $P_{X,Y}(x,y)$ and the product distribution $P_X(x)P_Y(y)$, i.e.,

$$\begin{aligned} I(X;Y) &= D(P_{X,Y}||P_XP_Y) \\ &= \sum_{i\in E}\sum_{j\in E} P_{X,Y}(i,j)\log\frac{P_{X,Y}(i,j)}{P_X(i)P_Y(j)}. \end{aligned} \tag{6}$$

We can rewrite the definition of mutual information as

$$I(X;Y) = H(X) - H(X|Y), \tag{7}$$

and the chain rule for mutual information is

$$I(X_1^n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1^{i-1}). \tag{8}$$

Lemma 1.1. *For any two random variables, X and Y,*

$$I(X;Y) \geq 0, \tag{9}$$

whit equality if and only if X and Y are independent.

Proof. [5], Page 30.

2 Existence of the relative entropy rate

The relative entropy rate between two stochastic processes $\{X_n\}_{n\in\mathbf{N}}$ and $\{Y_n\}_{n\in\mathbf{N}}$ defined in [10] as

$$D(\mathcal{X}||\mathcal{Y}) := \lim_{n\to\infty}\frac{1}{N^n}D(P_{X_1^n}||P_{Y_1^n}), \tag{10}$$

where $S = \{0,1,2,...,N-1\}$.

Consider $q_{ij} = \dfrac{1}{N}$ for any $0 \leq i,j \leq N-1$. The Mutual informatin X_i and Y_i is,

$$\begin{aligned} I(X_i;Y_i) &= H(Y_i) - H(Y_i|X_i) \\ &= \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} P_{X,Y}(x,y)\log\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \\ &= 0. \end{aligned} \tag{11}$$

Because

$$P_{X,Y}(x,y) = P_{Y|X}(y|x)P_X(x) = \frac{1}{N}P_X(x) \tag{12}$$

and

$$P_Y(y) = \sum_{x=0}^{N-1} P_{X,Y}(x,y) = \sum_{x=0}^{N-1}\frac{1}{N}P_X(x) = \frac{1}{N}, \tag{13}$$

so $P_{X,Y}(x,y) = P_X(x)P_Y(y)$. By lemma 2.1 X_i and Y_i is independent for any $0 \leq i \leq N-1$.

Lemma 2.1. *Let $(X,Y) \sim P(x,y) = p(x)p(y|x)$. The mutual information $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$. $D(p||q)$ is convex in the pair (p,q).*

Proof. [5], Pages 32 and 33.

$p(y|x)$ is constant and equals to $\frac{1}{N}$, so $I(X;Y)$ is concave with minimum value and $D(p||q)$ is convex. Decreasing of mutual information will result increasing of relative entropy. For obtaining the maximum value of the relative entropy rate, we have

$$\begin{aligned}
P_{Y_i|Y_1^{i-1}}(s_i|s_1^{i-1}) &= \sum_{x_1^i\in S^i} P_{Y_n|Y_1^{i-1},X_1^i}(s_i|s_1^{i-1},x_1^i)P_{X_1^i}(x_1^i)\\
&= \sum_{x_1^i\in S^i} P_{Y_i|X_i}(s_i|x_i)P_{X_1^i}(x_1^i)\\
&= \sum_{x_1^i\in S^i} \frac{1}{N}P_{X_1^i}(x_1^i)=\frac{1}{N}.
\end{aligned}\tag{14}$$

Where entries of Q are equal to $\frac{1}{N}$, the relative entropy rate of X and Y is maximum. We replace $P_{Y_n|Y_1^{n-1}}$ in definition of $D(P_{X_1^n}||P_{Y_1^n})$ by $\frac{1}{N}$, so we can obtain $\bar{D}(P_{X_1^n}||P_{Y_1^n})$ as,

$$\begin{aligned}
\bar{D}(P_{X_1^n}||P_{Y_1^n}) &= \bar{D}(P_{X_1^{n-1}}||P_{Y_1^{n-1}})+\sum_{s_1^n\in S^n} P_{X_n|X_{n-1}}(s_n|s_{n-1})\log N\\
&\quad+\sum_{s_1^n\in S^n} P_{X_n|X_{n-1}}(s_n|s_{n-1})\log P_{X_n|X_{n-1}}(s_n|s_{n-1})\\
&= \bar{D}(P_{X_1^{n-1}}||P_{Y_1^{n-1}})+N^{n-1}\log N\\
&\quad+N^{n-2}\sum_{s_1^2\in S^2} P_{X_2|X_1}(s_2|s_1)\log P_{X_2|X_1}(s_2|s_1).
\end{aligned}\tag{15}$$

Know that $\bar{D}(P_{X_1^n}||P_{Y_1^n})$ is the maximum of $D(P_{X_1^n}||P_{Y_1^n})$. Let $\alpha = N\log N+\sum_{s_1^2\in S^2} P_{X_2|X_1}(s_2|s_1)\log P_{X_2|X_1}(s_2|s_1)$ for simplicity. So

$$\begin{aligned}
\bar{D}(P_{X_1^n}||P_{Y_1^n}) &= \bar{D}(P_{X_1^{n-1}}||P_{Y_1^{n-1}})+N^{n-2}\alpha\\
&= \bar{D}(P_{X_1^{n-2}}||P_{Y_1^{n-2}})+(N^{n-2}+N^{n-3})\alpha\\
&\vdots\\
&= \bar{D}(P_{X_1}||P_{Y_1})+\sum_{i=2}^{n} N^{n-i}\alpha\\
&= \bar{D}(P_{X_1}||P_{Y_1})+\frac{N^{n-1}-1}{N-1}\alpha.
\end{aligned}\tag{16}$$

By noting the definition of the relative entropy rate in (10), we can write

$$\frac{1}{N^n}\bar{D}(P_{X_1^n}||P_{Y_1^n})=\frac{1}{N^n}\bar{D}(P_{X_1}||P_{Y_1})+\frac{N^{n-1}-1}{N^n(N-1)}\alpha.\tag{17}$$

We know $\frac{1}{N^n}\bar{D}(P_{X_1^n}||P_{Y_1^n})$ is increasing with respect to n, for every arbitrary matrices $\mathbf{P}$ and $\mathbf{Q}$. Also for every arbitrary matrix $\mathbf{P}$, the relation $\frac{1}{N^n}D(P_{X_1^n}||P_{Y_1^n})$ is maximum for matrix $\mathbf{Q}$ with entries $q_{x,y}=N^{-1}$. The amount of this maximum is $\frac{1}{N^n}\bar{D}(P_{X_1^n}||P_{Y_1^n})$ in (17). So $D(\mathcal{X}||\mathcal{Y})$ is well defined. α is depend on matrix $\mathbf{P}$ so we can get

$$\sup_{\mathbf{Q}}\{D(\mathcal{X}||\mathcal{Y})\}=\bar{D}(\mathcal{X}||\mathcal{Y}),\tag{18}$$

where

$$\bar{D}(\mathcal{X}||\mathcal{Y})=\lim_{n\to\infty}\frac{1}{|S|^n}\bar{D}(P_{X_1^n}||P_{Y_1^n})=\frac{\alpha}{N(N-1)}.\tag{19}$$

3 Application

In visual speech recognition, the deaf viewers do not hear the speaker but just see the lip movements. So they should try to guess the speaker's words based on their past experience. Here, uttering the words randomly ends in a Markov chain and seeing the speaker's lip movements seeing is like a stochastic channel, the output of which is a sequence of words guessed by the deaf with the risk of wrongly recognizing the words. This output sequence is a hidden Markov chain.

Consider the following test: six Persian words "one, two, three, four, five and six" are expected to be recognized. 20 persons were asked to repeat each of them 10 times. In other words, each word would be repeated 200 times. Their lip movements were filmed, which makes it possible to identify the uttered words by different methods.

For any hidden Markov chain model, we need the following elements:
i) The initial distribution vector
ii)Propapility transition matrix
iii)noisy matrix

Suppose that Table17 contains the number of diagnoses of words.

Table 17: The number of observations, true or false

	one	two	three	four	five	six
one	154	1	38	1	4	0
two	0	194	0	4	0	2
three	39	0	155	0	1	1
four	2	3	3	185	4	3
five	4	0	1	3	188	2
six	1	2	3	7	3	192

Now due to the hidden Markov model elements to Table17, we have:

$$P = \frac{1}{6}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \quad , \quad Q = \frac{1}{200}\begin{bmatrix} 154 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 192 \end{bmatrix} \tag{20}$$

and $\pi_0 = \frac{1}{6}[1\ 1\ 1\ 1\ 1\ 1]$. Similarly, these matrices can be obtained for each table, which contains the results of experiments with different methods.

The relative entropy between a Markov chain and its corresponding hidden Markov chain in this model for some sequence with size n, is calculated and presented in Table18.

Table 18: The relative entropy for n=3 to 10

n	$D(X_1^n \| Y_1^n)$	n	$D(X_1^n \| Y_1^n)$
3	0.3685	7	0.3872
4	0.3793	8	0.3879
5	0.3831	9	0.3884
6	0.3856	10	0.3886

The sequence of numbers in Table 2 is Convergent. The value of the limit of this sequence is equal to the relative entropy. (It should be noted that the speed of convergence of this sequence is so high, that up to 3 decimal places is sufficient to get its value of the limit n = 10.)

We know that the relative entropy is the measure of the distance between two distributions. In statistics, it arises as an expected logarithm of the likelihood ratio. The relative entropy $D(p_X||p_Y)$ is the measure of the inefficiency of assuming that the distribution is p_Y when the true distribution is p_X. So it can be used to compare different ways of recognizing the visual speech. Suppose that, the sequential speaker's words of the speaker which form a stochastic Markov process can be recognized by two or more different methods in terms of the speaker's lip movements.

In each of these methods, the recognized sequence of words is a hidden Markov process. The rate of relative entropy between these two processes is a measure of distance between them. So a reduction in relative entropy indicates how close the uttered word and speakers identification are. Usually, the amount of efficiency is calculated by dividing the number of correct diagnoses on the number of total occurrences of words. This value is obtained 0.89 by the data of Table 1. Since, these values are directly associated with the word and the movement of the speaker's lips, it will be changed by the change of words. On the other hand, if the efficiency of these two methods is equal, how should the methods be selected? Due to the dependency of the relative entropy only to the probability (values) and not to the words, it would be better to use this value for determining the effectiveness of the method. The lower the relative entropy rate is, the more efficient the method is.

References

[1] Amari, S. (1985), *Differential geometrical methods in statistics.* Springer-Verlag New York.

[2] Baum, L.E., Petrie, T. (1966), Statistical inference for probabilistic functions of finite state Markov chains.*J. Ann. Math. Statist.*, **37** 1554-1563.

[3] Baum, L.E., Petrie, T., Soules, G., Weiss, N. (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains.*J. Ann. Math. Statist.*, **41** 164-171.

[4] Chazottes, J. R., Giardina, C., Redig, F. (2006), Relative entropy and waiting times for continuous time Markov processes. *Electronic J. Prob.*, **11** 1049-1068.

[5] Cover, T.M., Thomas, J.A. (2006), *Elements of information theory,* Jone Wiley and Sons, Inc., New York.

[6] Csiszar, I. (1967), Information type measures of difference of probability distributions and indirect observations.*J. Stud. Sci. Math. Hung.*, **2** 299-318.

[7] Kesidis, G., Walrand. J. (1993), Relative entropy between Markov transition rate matrices.*J. IEEE Trans. Inform. Theo.*, **39** 1056-1057.

[8] Kullback, S., Leibler, R. (1948), On information and sufficiency.*J. Ann. Math. Statist.*, (1951), **22** 79-86.

[9] Shannon, E.C. (1948), A mathematical theory of communication.*J. Bell Syst. Tech.*, **27** 379-423 ,623-656.

[10] Yari, G.H., Nikooravesh, Z. (2011), Relative entropy rate between a Markov chain and its corresponding hidden Markov chain.*J. Statist. Res. Iran*, **8** 97-109.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

Expression and bounds for the entropy of coherent system

Toomaj, A. [1]**and Doostparast, M.**[2]

[1] Department of Statistics, Gonbad Kavous University, Gonbad Kavous, Iran
[2] Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran
Email: ab.toomaj@gmail.com

Abstract

Entropy of engineering systems, as a measure of uncertainty, has been studied in statistical and reliability literature. In this paper, we provide an expression for the entropy of a coherent system lifetime by using the concept of minimal signature when lifetimes of components are independent and identically distributed. We also obtain bounds for the entropy of system lifetime in terms of the entropy of component lifetimes. It is shown that bounds are very useful when the system has a large number of components or the configuration of the system is complicated. Some examples are also given.

Keywords: Bridge system, Coherent system, KL information, k-out-of-n system, Minimal signature.

1 Introduction

Coherent systems have widely been used in various areas of engineering reliability. For the definition and basic properties of coherent systems, we refer the reader to Barlow and Proschan [1]. The performance of a system's design and the lifetime of the system are of interest and may be measured in a variety of ways. In the recent decades, a very useful measure is the notion of system signature successfully applied to compute the system characteristics. The system signature is the n-dimensional probability vector $\mathbf{s} = (s_1, \cdots, s_n)$ whose the ith element is $s_i = P(T = X_{i:n})$, where T denotes the system lifetime and $X_{i:n}$ is the order statistics of n independent and identically distributed (i.i.d.) component lifetimes, see, e.g., Samaniego [9]. Hereafter, we consider a coherent system consisting of n i.i.d. components with lifetimes $X_1, \cdots, X_n$ having the common cumulative distribution function (cdf) F which is absolutely continuous with a probability density function (pdf) f. We denote the system lifetime by T and the pdfs of the order statistics associated to component lifetimes i.e. $X_{1:n}, \cdots, X_{n:n}$, by $f_{1:n}, \cdots, f_{n:n}$, respectively. It follows that (see, e.g., Samaniego [10])

$$\bar{F}_T(t) = P(T > t) = \sum_{i=1}^{n} s_i \bar{F}_{i:n}(t), \; t > 0, \tag{1}$$

where $\bar{F}_{i:n}(t) = 1 - F_{i:n}(t)$ is the survival function of $X_{i:n}$. From (1), the pdf of T can be written as

$$f_T(t) = \sum_{i=1}^{n} s_i f_{i:n}(t), \tag{2}$$

where

$$f_{i:n}(t) = i\binom{n}{i}[F(t)]^{i-1}[\bar{F}(t)]^{n-i}f(t),\ t > 0. \tag{3}$$

The probability vector $\mathbf{s} = (s_1, \cdots, s_n)$ is called the *signature* of the coherent system. Equation (2) can be found as a weighted combination of the k-out-of-n systems which fails due to failure of the k-th component. Another useful representation for (1) when the components are are exchangeable, i.e. $F(x_1, \cdots, x_n) = P(X_1 \leq x_1, \cdots, X_n \leq x_n) = F(x_{\pi_1}, \cdots, x_{\pi_n})$ for any permutation $\pi = (\pi_1, \cdots, \pi_n)$ of the indices $\{1, \cdots, n\}$, is

$$\bar{F}_T(t) = \sum_{i=1}^{n} a_i \bar{F}_{1:i}(t),\ t > 0. \tag{4}$$

The vector $\mathbf{a} = (a_1, \cdots, a_n)$ is called *minimal signature* (see, e.g., Navvaro *et al.*, [6]). Specifically, when lifetimes of components are i.i.d., Expression (4) yields

$$\bar{F}_T(t) = \sum_{i=1}^{n} a_i [\bar{F}(t)]^i,\ t > 0, \tag{5}$$

and hence

$$f_T(t) \quad = \quad \sum_{i=1}^{n} i a_i [\bar{F}(t)]^{i-1} f(t). \tag{6}$$

In this paper, we provide an expression for the entropy of coherent system lifetime by using the concept of minimal signature. The Shannon [11] entropy is a measure of uncertainty and predictability of the random variable X. If X is an absolutely continuous nonnegative random variable with the pdf f, then X may be viewed as the random lifetime of a system or a component or a living organism. The Shannon entropy of X is defined by

$$H(X) = H(f) = -\int_0^{\infty} f(x)\log(f(x))dx. \tag{7}$$

Throughout this paper, "log" will denote the natural logarithm. The index $H(f)$ measures the uniformity of a distribution function. Another measure of uncertainty of two distributions is well-known Kullback-Leibler (KL) discrimination information of random variables X and Y with pdfs f and g, respectively, defined by

$$\begin{aligned} K(f:g) &= \int_0^{\infty} f(x)\log\frac{f(x)}{g(x)}dx \\ &= -H(f) + H(f,g), \end{aligned} \tag{8}$$

where $H(f,g) = -E_f[\log g(X)]$ is known as *Fraser information* (see, e.g., Ebrahimi *et al.* [2]). The KL discrimination information is always nonnegative. Note that $K(f:g) = 0$ if and only if $f(x) = g(x)$ almost everywhere. The KL information was first introduced by Kunllback and Leibler [4] to measure of the distance between two distributions.

Studying of the entropy of engineering systems has been studied in statistical and reliability literature, see, e.g., Ebrahimi *et al.* [5], Wong and Chen [14], Park [8] and Toomaj and Doostparast [12, 13]. Toomaj and Doostparat [12, 13] studied the entropy of the lifetime of the coherent system as well as the KL discrimination of coherent systems lifetime. They obtained several results about the information properties of coherent systems by using the concept of system signature. Specially, they proposed an order to choose a preferable system among two coherent systems.

2 Expression and bounds for $H(T)$

Consider a coherent system with lifetime T and the minimal signature vector $\mathbf{a} = (a_1, \cdots, a_n)$ consists n i.i.d. components with lifetimes $X_1, \cdots, X_n$ having the common cdf F. It is known that the corresponding transformations of component lifetimes are i.i.d. random variables $U_i = F(X_i)$ which are uniformly distributed on $[0, 1]$ i.e. $U_i \sim U[0, 1]$. Also the random variables $W_i = F(X_{1:i})$ has the beta distribution with parameters 1 and i, i.e. $W_i \sim B(1, i)$. It is well known that the pdf of W_i is given by

$$g_i(w) = i(1-w)^{i-1}, \quad 0 < w < 1, \ i = 1, \cdots, n. \tag{9}$$

The entropy of coherent system lifetime is found by using the transformation $V = F(T)$. The PDF of $V = F(T)$ is $g_V(v) = \sum_{i=1}^{n} s_i g_i(v)$ so that the Jacobian of the transformation for $T = F^{-1}(V)$ is $1/f(F^{-1}(v))$. Applying $T = F^{-1}(V)$ the transformation formula for the system's entropy, the following useful representation for the coherent system lifetime can be derived

$$\begin{aligned} H(T) &= H(V) - E[\log f(F^{-1}(V))], \\ &= H(V) - \sum_{i=1}^{n} a_i E[\log f(F^{-1}(V))], \end{aligned} \tag{10}$$

where $H(V)$ is the entropy of a coherent system with lifetime V and minimal signature $\mathbf{a}$ consisting of n i.i.d. component lifetimes follow the standard uniform distribution. Expression (10) is useful to develop various results about the entropy of coherent system's lifetime. When the number of components is large or the structure of the system is complicated, evaluating of $H(V)$ is not easy in such situations. Note that this is a common situations in practice. The bounds are very useful in these cases. In the forthcoming theorem, we use the mentioned earlier results to provide bounds for $H(T)$ in terms of the entropy of the component lifetime. First, we have the following lemma.

Lemma 2.1. *If V denotes the lifetime of the coherent system consisting of n possibly i.i.d. component lifetimes having the common marginal standard uniform distribution, then*

$$-\log \sup_{0<v<1} g_V(v) \leq H(V) \leq 0. \tag{11}$$

Proof. Let T be the lifetime of the coherent system consisting of n i.i.d. component lifetimes having the common marginal pdf f and cdf F. By applying transformations $U = F(X)$ and $V = F(T)$, we have

$$\begin{aligned} K(f_T : f) &= K(g_V : U) = \int_0^1 g_V(v) \log g_V(v) dv, \\ &= -H(V) \geq 0. \end{aligned}$$

The first equality follows from the invariant property of the KL discrimination information under one-to-one transformations $U = F(X)$ and $V = F(T)$ and this derive the upper bound. To obtain the lower bound, since $g_V(v) \leq \sup_{0<v<1} g_V(v)$, we have

$$H(V) \geq -\log \sup_{0<v<1} g_V(v) \int_0^1 g_V(v) dv = -\log \sup_{0<v<1} g_V(v),$$

and the desired result follows. □ □

Theorem 2.2. *Let T denote the lifetime of the coherent system with minimal signature* $\boldsymbol{a}$ *consisting of n i.i.d. component lifetimes having the common cdf F.*
(a) *By noting that $H(X) < \infty$, we have*

$$H(T) \geq -\log \sup_{0<v<1} g_V(v) + \sup_{v \in (0,1)} g_V(v)[H(X) + I(A)], \tag{12}$$

$$H(T) \leq \sup_{v \in (0,1)} g_V(v)[H(X) + I(\bar{A})], \tag{13}$$

where $A = \{x : f(x) \leq 1\}$, $\bar{A} = \{x : f(x) > 1\}$ *and*

$$I(A) = \int_A f(x) \log f(x) dx.$$

(b) *Suppose* $M = f(m) < \infty$, *where* $m = \sup\{x : f(x) \leq M\}$ *is the mode of the distribution and* $H(X) < \infty$. *Then*

$$\begin{aligned} H(T) &\geq -[\log \sup_{0<v<1} g_V(v) + \log M], & (14)\\ H(T) &\leq -\log M + \sup_{v\in(0,1)} g_V(v)[H(X) + \log M]. & (15) \end{aligned}$$

Proof. (a) From (10), we have

$$\begin{aligned} -E[\log f(F^{-1}(V))] &= -\int_{A_1} g_V(v) \log f(F^{-1}(v)) dv & (16)\\ &\quad -\int_{\bar{A}_1} g_V(v) \log f(F^{-1}(v)) dv & (17)\\ &\quad \leq -\int_{A_1} g_V(v) \log f(F^{-1}(v)) dv\\ &\quad \leq \sup_{v\in(0,1)} g_V(v) \left[-\int_{A_1} \log f(F^{-1}(v)) dv\right]\\ &= \sup_{v\in(0,1)} g_V(v) \left[-\int_A f(x) \log f(x) dx\right]\\ &= \sup_{v\in(0,1)} g_V(v) \left[H(X) + \int_{\bar{A}} f(x) \log f(x) dx\right], & (18) \end{aligned}$$

where

$$A_1 = \{v : f(F^{-1}(v)) \leq 1\}, \ \bar{A}_1 = \{v : f(F^{-1}(v)) > 1\}.$$

The first inequality in (18) is obtained by the fact that the integrate in (17) is nonnegative and the second inequality obtained by noting that $g_V(v) \leq \sup_{v\in(0,1)} g_V(v)$, $\forall\ v \in (0, 1)$. By using equations (10), (11) and (18), the desired result in (13) follows. The lower bound in (12) can be obtained by using equations (10) and (11) and the fact that the integrate in (16) is nonpositive.

(b) Let us consider a coherent system with lifetime $Y = MT$ consisting of n i.i.d. components with lifetimes $Z_1, \cdots, Z_n$ where $Z_i = MX_i$, $i = 1, \cdots, n$. Since $f_Z(z) = \frac{1}{M} f(\frac{z}{m}) \leq 1$, for all $z > 0$, hence $I(A) = -H(Z)$ and $I(\bar{A}) = 0$. From Part (a), we have

$$-\log \sup_{0<v<1} g_V(v) \leq H(Y) \leq \sup_{v\in(0,1)} g_V(v) H(Z).$$

By noting that $H(Z) = H(X) + \log M$ and $H(Y) = H(T) + \log M$, the desired result in Part (b) follows. □ □

As an application of the obtained bounds in Equations (12)-(14), consider the following example.

Example 5. Let us consider the bridge system with lifetime T and the minimal signature $\mathbf{a} = (0, 2, 2, -5, 2)$ consisting of $n = 5$ i.i.d. component lifetimes having the common CDF F. It is not hard to verify that

$$\sup_{v\in(0,1)} g_V(v) = g_V(\frac{1}{2}) = 1.625.$$

(i) Let X follow a uniform distribution on [0,b]. It is known that $M = b^{-1}$ and $H(X) = \log b$, hence Part (b) implies

$$-0.4855078 + \log b \leq H(T) \leq \log b.$$

(ii) If X has the exponential distribution with mean $1/\lambda$, then we have $M = \lambda$ and $H(X) = 1 - \log\lambda$. The entropy of the system's lifetime is bounded as follows:

$$-[0.4855078 + \log\lambda] \leq H(T) \leq 1.625 - \log\lambda.$$

(iii) Suppose that X has the Parto distribution type II with the following pdf

$$f(x) = \alpha(1+x)^{-(\alpha+1)},\ x \geq 0,\ \alpha > 0.$$

It is easy to see that $M = \alpha$ and $H(X) = \alpha^{-1} - \log\alpha + 1$. Therefore Part `(b)` yields

$$-[0.4855078 + \log\alpha] \leq H(T) \leq -\log\alpha + 1.625(1 + \alpha^{-1}).$$

□

The bounds in Theorem 2.2 are useful when the probability distribution does not have a closed form, and hence the density function $g_V(v)$ cannot be easily evaluated. Explicit expressions for many well-known distributions are available, and then the evaluation of proposed bounds in Theorem 2.2 are numerically simple.

References

[1] Barlow, R. E. and Proschan, F. (1981), *Statistical Theory of Reliability and Life testing.* Silver Spring, MD: To Begin With.

[2] Ebrahimi, N., Soofi, E. S. and Soyer, R. (2010), Information measures in perspective. *International Statistical Review*, **78**, pp. 383-412.

[3] Ebrahimi, N., Soofi, E. S. and Zahedi, H. (2004), Information properties of order statistics and spacings. *IEEE Transactions on Information Theory*, **46**, pp. 209-220.

[4] Kullback, S. and Leibler, R.A. (1951), On information and sufficiency. *The annals of mathematical statistics*, **22**, pp. 79-86.

[5] Kochar, S., Mukerjee, H. and Samaniego, F. J. (1999), The signature of a coherent system and its application to comparisons among systems. *Naval Research Logistics*, **46**, pp. 507-523.

[6] Navarro, J., Ruiz, J. M. and Sandoval, C. J. (2007), Properties of coherent systems with dependent components. *Communications in Statistics: Theory and Methods*, **36**, pp.175-191.

[7] Navarro, J., Samaniego, F. J., Balakrishnan, N. and Bhattacharya, D. (2008), On the application and extension of system signatures to problems in engineering reliability. *Naval Research Logistics*, **55**, pp. 313-327.

[8] Park, S. (1995), The entropy of consecutive order statistics. *IEEE Transactions on Information Theory*, **41**, pp. 2003-2007.

[9] Samaniego, F. J. (1985), On closure of the IFR class under formation of coherent systems. *IEEE Transactions on Reliability*, **R-34**, pp. 69-72.

[10] Samaniego, F. J. (2007), *System Signatures and Their Applications in Engineering Reliability.* New York: Springer, International series in operations research and management science, **110**.

[11] Shannon, C. E. (1948), A mathematical theory of communication. *Bell System Technical Journal*, **27**, pp. 379-423 and 623-656.

[12] Toomaj, A. and Doostparast, M. (2014), A note on signature based expressions for the entropy of mixed r-out-of-n systems. *Naval Research Logistics*, **61**, pp. 202-206.

[13] Toomaj, A. and Doostparast, M. (2014), On the Kullback-Leibler information for mixed systems. *International Journal of Systems Science*, DOI: 10.1080/00207721.2014.998744.

[14] Wong, K. M. and Chen, SH. (1990), The entropy of ordered sequences and order statistics. *IEEE Transactions on Information Theory*, **36**, pp. 276-284.

The 2nd Workshop on
Information Measures and Their Applications
28-29 January, 2015, Ordered and Spatial Data Center of Excellence
Ferdowsi University of Mashhad, Iran

A short note on the cumulative residual entropy

Zohrevand, Y. [1], Hashemi, R.[1] and Asadi, M.[2]

[1] Department of Statistics, Razi University , Kermanshah, Iran
[2] Department of Statistics, University of Isfahan, Isfahan, Iran
Email: y.zohrevand@gmail.com

Abstract

In this paper, we study maximum entropy approach in terms of Shannon entropy and entropy of equilibrium distribution, known as cumulative residual entropy (CRE). In classical maximum entropy approach the model whose uncertainty is maximum in a set of distributions, under some constraints (usually moments constraints), selected as the maximum entropy model. We present maximum entropy of equilibrium distribution model, under some partial ordering in well known family of lifetime distributions.

Keywords: Uncertainty, Cumulative Residual Entropy, Equilibrium Distribution, Partial Ordering, Maximum Entropy.

1 Introduction

Let X be a nonnegative absolutely continuous random variable with distribution function F and probability density function f, respectively. Shannon (1948) introduced a measure of uncertainty based on probability mass function for discrete random variables. Differential entropy as a well-known measure of uncertainty in the continuous case is defined as

$$H(f) = -\int_0^\infty f(x) log f(x) dx,$$

Rao et al. (2004) introduced CRE for nonnegative random variables based on cumulative distribution function as

$$\varepsilon(F) = -\int_0^\infty \overline{F}(x) log \overline{F}(x) dx,$$

where $\overline{F} = 1 - F$ is survival function.CRE has many interesting applications in different branches of sciences such as reliability theory, survival analysis, computer vision, image processing and etc.

Asadi and Zohrevand (2007) showed that CRE is equal to expectation of mean residual life function and introduced the dynamic version of CRE (DCRE).

$$\varepsilon(X) = -\int_0^\infty \overline{F}(x) log \overline{F}(x) dx = E[m_F(X)],$$

where $m_F(x)$ is mean residual life function of distribution F.

A relationship between $H(f)$ and $\varepsilon(F)$ can be drown via Shannon entropy of equilibrium distribution (ED). Let X be a *r.v.* with survival function $\bar{F}(x)$ and $\mu = E(X) < \infty$. Then the PDF of ED is

$$f_e(x) = \frac{\bar{F}(x)}{\mu},$$

hence easily can see

$$H(f_e) = \frac{\varepsilon(F)}{\mu} + log\mu.$$

Several applications and properties of CRE and DCRE have studied by many researchers such as Rao (2005), Zografos and Nadarajah (2005), Di Crescenzo and Longobardi(2009),(2011), Navarro et al. (2010), Longobardi (2014) and Zardasht et al. (2014).

In the rest of paper, we study CRE ordering under some well known partial ordering and present the concept of maximum entropy of ED of distributions.

2 Ordering of CRE and Maximum Equilibrium Distributions Entropy

Definition 2. *Let Ω_F be a set of nonnegative absolutely continuous distributions with some constraints on F's. If $F^* \in \Omega_F$ be such that for all $F \in \Omega_F$*

$$\varepsilon(F) \leq \varepsilon(F^*),$$

then F^* is maximum ED entropy (MEDE) in Ω_F.

Usually in procedure of generating maximum entropy (ME) models, in main approach, it is set some moment constraints on distributions. In lifetime studies, there exist situations in which hazard rate (HR) function or mean residual life (MRL) function satisfy some conditions. Ebrahimi (2000) studied the ME models under these conditions. Asadi et al. (2004) investigated ME models for residual lifetime in terms of HR order with monotone PDF. They stated constraints based on differential equations on HR and MRL functions. Asadi et al. (2014) presented a general exponential form of MEDE models under some moment constraints which is similar to ME models.

In present paper we investigate MEDE and ME models under some partial ordering as the constraints on distributions.

Definition 3. *Let X and Y be two nonnegative absolutely continuous r.v. with survival functions $\bar{F}$ and $\bar{G}$, mean residual lives m_F and m_G and hazard functions r_F and r_G, respectively. There is several ordering of distributions based on these functions.*

(a) X is said to be smaller than Y in stochastic order, denoted by $X \leq_{st} Y$, if $\bar{F}(t) \leq \bar{G}(t)$ for all $t > 0$.

(b) X is said to be smaller than Y in mean residual life order, denoted by $X \leq_{mrl} Y$, if $m_F(t) \leq m_G(t)$ for all $t > 0$.

(c) X is said to be smaller than Y in hzard order, denoted by $X \leq_{hr} Y$, if $r_F(t) \geq r_G(t)$ for all $t > 0$.

(d) X is said to be smaller than Y in likelihood ratio order, denoted by $X \leq_{lr} Y$, if for all $t > 0$, $\frac{g(t)}{f(t)}$ be increasing in t.

(e) X is said to be smaller than Y in convex order, denoted by $X \leq_{cx} Y$, if for all convex function ϕ, $E(\phi(X)) \leq E(\phi(Y))$.

In following theorems we present MEDE and ME models under some conditions. Before this aim, we have to state an useful lemma of Ebrahimi et al. [5].

Lemma 2.1. *Suppose F and G be two probability distribution functions with survival functions $\bar{F}$ and $\bar{G}$ respectively. Let F is absolutely continuous relative to G ($F \prec\prec G$), and $\bar{F}(t) \leq \bar{G}(t)$ for all $t > 0$, if the pdf of g is decreasing (increasing), then $H(f) \leq (\geq) H(g)$.*

Theorem 2.2. *Let Ω_F={$F : r_F(t) \geq r_G(t)$ for all $t > 0$ and $F \prec\prec G$ } be a set of nonnegative absolutely continuous distributions. Then $F^* = G$ is the MEDE in Ω_F.*

Proof: Suppose $X \leq_{hr} Y$, from Theorem 1.C.13 of [6] we have $X_e \leq_{lr} Y_e$, where X_e, Y_e have decreasing equilibrium density function related to X and Y respectively. So $X_e \leq_{lr} Y_e$ implies $X_e \leq_{hr} Y_e$ and $X_e \leq_{st} Y_e$. Also, one can see if X is absolutely continuous with respect to Y ($F \prec\prec G$) then for the related equilibrium distributions we have, X_e is absolutely continuous with respect to Y_e (also $F_e \prec\prec F$, $G_e \prec\prec G$). Finally from Lemma 2.1, we have $H(f_e) \leq H(g_e)$, in other word G is MEDE model in Ω_F.

Corollary 1. *In Theorem 2.2, if the pdf g is decreasing then G also is ME model in Ω_F. For example, in decreasing failure rate (DFR) family of distributions this property hold.*

Example 6. *A useful and applicable family of distributions in survival analysis and reliability theory is the family of proportional hazard models. Let Ω_F={$F : r_F(t) = c.r_G(t)$ for all $c > 1$ and $t > 0$ }. If $F \prec\prec G$, then the conditions of Theorem 2.2 can hold in this set of distributions and G is MEDE in Ω_F. Note that if g, the PDF of G, is decreasing then G also is ME in Ω_F.*

Theorem 2.3. *Let Ω_F={$F : m_F(t) \leq m_G(t)$ for all $t > 0$ and $F \prec\prec G$ } be a set of nonnegative absolutely continuous distributions. Then $F^* = G$ is the MEDE in Ω_F if the expectations of F and G are exist and finite.*

Proof: From Theorem 2.A.4 of [6], we have $X \leq_{mrl} Y$ if, and only if $X_e \leq_{hr} Y_e$, and this implies $X_e \leq_{st} Y_e$. As in Ω_F, $F \prec\prec G$, then $F_e \prec\prec G_e$ and from Lemma 2.1 and the fact that $g_e(t)$ is decreasing, one can conclude that G is MEDE model in Ω_F.

Corollary 2. *In Theorem 2.3, if the pdf g be decreasing and log-concave, then G also is MDE in Ω_F.*

Proof: See the Corollary 1 of Asadi et al.[2].

Example 7. *Another well-known family of distributions in survival analysis and reliability theory is the proportional mean residual life distributions. Let Ω_F={$F : m_F(t) = c.m_G(t)$ for all $0 < c < 1$, $t > 0$ and $\mu_G < \infty$ }. Then from Theorem 2.3 for all $0 < c < 1$, G is MEDE in Ω_F.*

If the distributions F and reference distribution G in Lemma 2.1 are selected suitably, especially in mixtures family and weighted distributions family, then one can generate ME and MEDE models in these classes of distributions using above theorems.

Corollary 3. *Suppose G as a reference distribution, be a nonnegative absolutely continuous distribution and $\Omega_F = \{F : F \leq_{hr} G$ and $F \prec\prec G\}$. Let $H(x) = pF(x) + (1-p)G(x)$ for some $p \in (0,1)$, be a mixture of F's and G, then G is MEDE in Ω_F. Also if the pdf h, be decreasing then G is ME for all $F, H \in \Omega_F$.*

Proof: If $X \leq_{hr} Y$ and W is a r.v. with distribution H. Then from Theorem 1.B.22 of [6], we have $X \leq_{hr} W \leq_{hr} Y$, and this implies $X \leq_{st} W \leq_{st} Y$. Also, if $F \prec\prec G$, it is easy to show that $H \prec\prec G$. As the conditions of Lemma 2.1 hold for all $F, H \in \Omega_F$, hence using the fact that g_e and h_e are decreasing, G is MEDE in Ω_F.

Corollary 4. *Let X and Y are two nonnegative absolutely continuous random variables with distributions F and G respectively and $F \prec\prec G$. Let $f_w(x) = \frac{w(x)}{E(w(X))}f(x)$ and $g_w(x) = \frac{w(x)}{E(w(Y))}g(x)$ be the weighted PDF of r.v. X and Y with finite $E[w(Y)]$ and $E[w(X)]$ respectively. If $w(x)$ is increasing and $X \leq_{hr} Y$ then $X^w \leq_{hr} Y^w$. Also easily can see $r_{Y^w}(x) = \frac{w(x)}{E(w(Y)|Y>x)}.r_Y(x)$, hence as $w(t)$ is increasing we have $w(t) \leq E(w(Y)|Y > t)$, for all $t > 0$ then $Y^w \leq_{hr} Y$ (see example 1.B.23 of [6]). From Theorem 2.2, G is MEDE in the set of distributions f^w's and g^w's which satisfy above conditions.*

Theorem 2.4. *Suppose Ω_F={$F : X \leq_{cx} Y$ and $\mu = E(Y) < \infty$, $F \prec\prec G$ }, then $G = F^*$ is MEDE in Ω_F.*

Proof: Let $X \leq_{cx} Y$ then $X_e \leq_{st} Y_e$ (see Theorem 3.A.65 of [6]). As $g_e(x)$ is decreasing, using lemma 2.1 we have $H(X_e) \leq H(Y_e)$ this implies $H(f_e) \leq H(g_e)$. Hence G is MEDE in Ω_F.

In following we study the conditions under which MEDE exists in DMRL family.

Theorem 2.5. *Let* Ω_F=$\{F$ *: F is DMRL and* $\mu_F < \infty$ $\}$*. If* $k = \sup_{F\in\Omega_F} \mu_F$ *exist, then MEDE is exponential distribution with mean* k *in* Ω_F*.*
Proof: Let F be DMRL then from theorem 4.2 of [Asadi and Zohrevand (2007)] we have

$$\varepsilon(F;t) \leq m_F(t).$$

As the exponential distribution is IMRL then for all $t > 0$, $\varepsilon(F^*;t) \geq m_{F^*}(t) = k$. Then we have

$$\varepsilon(F;t) \leq m_F(t) \leq m_{F^*}(t) \leq \varepsilon(F^*;t).$$

Now in $t = 0$, proof is complete.

References

[1] Asadi, M., Ebrahimi,N., Soofi, E. and Zarezadeh, S. (2014), New Maximum Entropy Methods for Modeling Lifetime Distributions, *Naval Research Logistic* **16**, 427-434.

[2] Asadi,M., Ebrahimi, N., Hamedani, G.G. and Soofi, E. (2004), Maximum Dynamic Entropy Models, *J.Appl. Prob.* **41**, 379-390.

[3] Asadi,M., Zohrevand, Y. (2007), On the dynamic cumulative residual entropy, *Journal of Statistical Planning and Inference* **137**, 1931-1941.

[4] Ebrahimi,N. (2000), The maximum entropy method for lifetime distributions, *Sankhya: The Indian Journal of Statistics* **62 (2)**, 236-243.

[5] Ebrahimi,N., Soofi, E.S. and Soyer,R. (2013), When are observed Failures More Informative than observed Survival, *Navel Research Logistics* **60**, 102-110.

[6] Shaked ,M. and Shanthikumar, J.G. (2007), Stochastic Orders, Netherlands: Springer.