Title	The ordering of Shannon entropies
Manuscript ID	SS-2016-0366R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0366
Complete List of Authors	Ming-Tien Tsai
Corresponding Author	Ming-Tien Tsai
E-mail	mttsai@stat.sinica.edu.tw

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0366

THE ORDERING OF SHANNON ENTROPIES

Ming-Tien Tsai

Academia Sinica

Abstract: Via the transformation of the convex ordering of distributions to the Lorenz ordering of new distributions, the information ordering of Shannon entropies is established. The measure of the difference between two Shannon entropies enjoys some merits.

Key words and phrases: Convex ordering, Lorenz ordering.

1. Introduction

Khinchin (1957) claimed that the discrete Shannon entropy, which enjoys many nice properties, is the unique uncertainty measure for discrete distributions. However, for the continuous data models the property of Shannon entropy seems to be quite ambiguous compared to that of the discrete version. Shannon himself pointed out that the continuous version is unsatisfactory, lacks invariance under one-to-one transformation. Neither can the continous Shannon entropy be directly obtained as the limit of a sequence of discrete entropies. In this note, we focus on how to clarify some of these ambiguities. A presentation indicates that the difference between two Shannon entropies is invariant under one-to-one transformations. We may note that the difference between two Shannon entropies is essentially symmetric, while the Kullback-Leibler divergence is not. For the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0366

sake of comparison, the ordering of Shannon entropies needs some appraisal. Karlin and Rinott (1981) used the majorization ordering to compare the Shannon entropies, and they further suggested (on page 351) using the notion of totally positive of order 2 (TP_2) to get a more complete theory on Shannon entropy comparisons. We adopt their suggestion to study the ordering of Shannon entropy measures in terms of the notion of convex ordering of distribution functions. The statistical inference for the one-sided one-sample problem considered by Barlow and Doksum (1972) can then be further generalized to those of the G-sample ordered restriction problems.

2. Main Results

Let $X_1, X_2, \cdots$ be independent random variables with a common continuous distribution function $F(x)$ which has absolutely continuous density function $f(x), x \in R$ with respect to Lebesgue measure. The Shannon entropy for $f(x)$ is denoted by

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

We assume that the integrability of $-\log f(x)$ with respect to $f(x)$ exists, and denote this class of density functions by

$$\mathcal{F} = \{f : \mathrm{I}_S(f) < \infty\}. \tag{2}$$

Note that the existence of moment generation function of $-\log f(x)$ implies the existence of Shannon entropy. There are two types of distributions satisfying this class of distributions: exponential-type distribution and Cauchy-type distribution.

For any two distributions $F_g, F_{g+1} \in \mathcal{F}$, it is said that F_g is *c-ordered* (convex

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0366

ordered) with respect to F_{g+1} ($F_g \leq_c F_{g+1}$) if and only if $F_{g+1}^{-1}F_g$ is convex on the interval where $0 < F_g(x) < 1$ (van Zwet, 1964). It is known that $F_g \leq_c F_{g+1}$ implies that $f_g(x)/f_{g+1}(F_{g+1}^{-1}F_g(x))$ is nondecreasing in x for $0 < F_g(x) < 1$.

For the G-sample case,

$$\begin{aligned}
&\mathrm{I}_S(f_g) - \mathrm{I}_S(f_{g+1}) \qquad (3)\\
&= -\int_{-\infty}^{\infty} \log f_g(x) f_g(x)dx + \int_{-\infty}^{\infty} \log f_{g+1}(y) f_{g+1}(y)dy\\
&= \int_0^1 \log f_{g+1}(F_{g+1}^{-1}(u))du - \int_0^1 \log f_g(F_g^{-1}(u))du\\
&= \int_0^1 \log[\frac{f_{g+1}(F_{g+1}^{-1}(u))}{f_g(F_g^{-1}(u))}]du, \text{ for all } g = 1, \cdots, G-1.
\end{aligned}$$

Barlow and Doksum (1972) considered the transformation $H_{F_g}^{-1}(u) = \int_0^{F_g^{-1}(u)} f_{g+1}[F_{g+1}^{-1}F_g(t)]dt$, $u \in [0,1]$, for all $g = 1, \cdots, G-1$. It is easy to see that $H_{F_{g+1}}^{-1}(u) = u$, $H_{F_{g+1}}(u)$ is the uniform distribution on $[0,1]$. If $F_g \leq_c F_{g+1}$, then $dH_{F_g}^{-1}(u)/du = f_{g+1}(F_{g+1}^{-1}(u))/f_g(F_g^{-1}(u))$ is nonincreasing in $u \in [0,1]$. Thus, $H_{F_g}^{-1}(u)$ is concave on $[0,1]$ and $H_{F_g}(v)$ is convex on the interval $H_{F_g}^{-1}(0) < v < H_{F_g}^{-1}(1)$. Note that $H_{F_g}(v)$ is a distribution function, where $H_{F_g}^{-1}(0) < v < H_{F_g}^{-1}(1)$, since $H_{F_g}^{-1}(u)$ (the inverse function of $H_{F_g}(v)$) is strictly increasing on $[0,1]$. This transformation depicts a Lorenz ordering of $H_{F_g}(v)$ with respect to $H_{F_{g+1}}(v)$ over the interval $[0,1]$. Let $h_{F_g}(v)$ and $h_{F_{g+1}}(v)$ be the density function of $H_{F_g}(v)$ and $H_{F_{g+1}}(v)$, respectively. Then, after some straightforward manipulations, we have that $\int_0^1 \log[f_{g+1}(F_{g+1}^{-1}(u))/f_g(F_g^{-1}(u))]du = -\int_0^1 \log h_{F_g}(v)dv$. Thus, by the information inequality we have that $-\int_0^1 \log h_{F_g}(v)dv = \int_0^1 \log[h_{F_{g+1}}(v)/h_{F_g}(v)]h_{F_{g+1}}(v)dv \geq 0$, for all $g = 1, \cdots, G-1$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0366

Theorem 1. *If $F_g \leq_c F_{g+1}$, then $I_S(f_g) \geq I_S(f_{g+1})$, for all $g = 1, \cdots, G-1$.*

The representation at equation (3) indicates that the difference between two Shannon entropies is invariant under one-to-one transformations. If $C(u) = F_{g+1}^{-1}F_g(u)$, then $[dH_{F_g}^{-1}(u)/du]^{-1} = dC(u)/du$. As such, the notion of convex ordering and the transformation of Barlow and Doksum (1972) are essentially the same, difference between $C(u)$ and $H_{F_g}(u)$ being a constant. The transformation of Barlow and Doksum (1972) is used to transform the convex ordering of distributions F_g and F_{g+1} to the Lorenz ordering of distributions $H_{F_g}(u)$ with respect to $H_{F_{g+1}}(u)$ (Gastwirth, 1971), which plays an important key for the proof of Shannon entropy ordering. This works as well for the discrete variable case, we outline it as in the following.

For any two discrete distributions F_g and F_{g+1}, let $F_g(y) = p_i^{'}, y \in [a_i, a_{i+1}), i = 1, \cdots, m$ and $F_g(a_{m+1}) = 1, a_{m+1}$ can be infinite, and let $F_{g+1}(y) = p_i, y \in [a_i, a_{i+1}), i = 1, \cdots, m$ and $F_{g+1}(a_{m+1}) = 1$. The discrete version of Barlow and Dorsum's transformation can be interpreted as follows: First, for $H_{F_{g+1}}(u)$ we start from point (0, 0) and then attach a segment with length $(p_i - p_{i-1}), i = 1, \cdots, m, p_0 = 0$, sequentially along the line with slope one to the end point (1, 1). Then, $H_{F_{g+1}}(u)$ is the continuous uniform distribution on [0, 1]. The distribution $H_{F_g}(u)$ can then be uniquely determinated by starting from the point (0, 0) and attaching a segment with length that is the increasement $(p_i^{'} - p_{i-1}^{'})$ times the corresponding change rate $(p_i^{'} - p_{i-1}^{'})/(p_i - p_{i-1})$ sequentially along the line with slope $(p_i^{'} - p_{i-1}^{'})/(p_i - p_{i-1}), i = 1, \cdots, m, p_0^{'} = 0$. Furthermore, the convex ordering is the same as the monotone likelihood ratio ordering; we have that

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0366

$F_g \leq_c F_{g+1}$ implies that $(F_g(a_i) - F_g(a_{i-1}))/(F_{g+1}(a_i) - F_{g+1}(a_{i-1}))$ is nondecreasing in a_i. Namely, $F_g \leq_c F_{g+1}$ implies that the slope $(p_i^{'} - p_{i-1}^{'})/(p_i - p_{i-1})$ is nondecreasing in a_i. Moreover, it is easy to see that a_i is nondecreasing in i. Thus, we may conclude that the $H_{F_g}(u)$ is a piecewise linear, continuous, slope nondecreasing and convex function. Hence, it depicts the Lorenz ordering of $H_{F_g}(u)$ with respect to $H_{F_{g+1}}(u)$ over the interval $[0, 1]$. Thus, the proof in Theorem 1 can go through for the discrete variable case as well.

3. Examples

The notions of totally positive of order 2, the convex ordering of two distributions, and the monotone likelihood ratio are essentially equivalent. In the literature, there is a wide class of distributions with monotone likelihood ratio (for the details see Lehmann, 1986), which ensures the ordering of Shannon entropies by Theorem 1.

In biostatistics, many studies fall under the setup of the well-known Cox proportional hazard (Cox, 1972); its basic model is that $\bar{F}_{g+1}(x) = (\bar{F}_g(x))^{\lambda_g}, \lambda_g > 0$, for all $g = 1, \cdots, G-1$. When $\lambda_g < 1$ for all $g = 1, \cdots, G-1$, it is easy to see that $F_g(x) \geq_c F_{g+1}(x)$, and hence by Theorem 1 we have $\mathrm{I}_S(f_g) \leq \mathrm{I}_S(f_{g+1})$ for distributions either continuous or discrete, for all $g = 1, \cdots, G-1$. Similarly, when $\lambda_g > 1$, $F_g(x) \leq_c F_{g+1}(x)$, and hence $\mathrm{I}_S(f_g) \geq \mathrm{I}_S(f_{g+1})$ for all $g = 1, \cdots, G-1$.

Asymptotic relative efficiency is essentially defined via large deviation asymptotics. It is well known that the Sanov theorem and its generalizations reduce the problem of large deviations to a minimization problem of Kullback-Leibler divergence on the corre-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0366

sponding set of distributions. Quadratic form statistics, which often have (asymptotic) noncentral χ^2 distributions, play an important role in statistics; the noncentral χ^2 distributions have monotone likelihood ratios, thus Theorem 1 is applicable. For procedures having the (asymptotic) noncentral χ^2 density functions with the same noncentrality, but different degrees of freedom, the method based on the asymptotic relative efficiency fails to compare them. However, the method based on the difference between two Shannon entropies can overcome the disadvantage of the method via large deviation asymptotics, such as the Kullback-Leibler divergence.

References

Barlow, R. E. and Doksum, K. A. (1972). Isotonic tests for convex orderings. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I*, 293-323. Univ. California Press, Berkeley, Calif.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. B* **34**, 187-220.

Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica* **39**, 1037-1039.

Karlin, S. and Rinott, Y. (1981). Entropy inequalities for classes of probability distributions II. The multivariate case. *Adv. Appl. Prob.* **13**, 325-351.

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory.* Translated by R. A. Silverman and M. D. Friedman. Dover Publications, Inc., New York.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0366

van Zwet, W. R. (1964). *Convex Transformations of Random Variables.* Math. Centrum, Amsterdam.

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan 11529

E-mail: mttsai@stat.sinica.edu.tw