Filomat 31:13 (2017), 4225–4230
https://doi.org/10.2298/FIL1713225T

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Some New Results on Information Properties of Mixture Distributions

Abdolsaeed Toomaj[a], Reza Zarei[b,*]

[a]*Faculty of Basic Sciences and Engineering, Department of Mathematics and Statistics, Gonbad Kavous University, Gonbad Kavous, Iran*
[b]*Department of Statistics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran*

Abstract. This paper deals with information properties of mixture models in terms of the single distributions in finite model. We provide an expression for the entropy of mixture models. Also, we derive bounds as well as an approximation to approximate the behavior of entropy of mixture distributions. Moreover, some new results on the entropy of mixture distributions in terms of ordering properties of single distributions are provided. Examples are given to illustrate the results.

1. Introduction

It is known that one of the most important issues to measure of uncertainty and predictability is the Shannon entropy defined by Shannon [19]. Differential entropy as a measure of uncertainty extends the classical entropy to continuous random variables. The differential entropy unlike the discrete case gives a value within $[-\infty, \infty]$ and achieves a minimum when the random variable comprises no uncertainty and approaches a maximum as the random variable becomes uniformly distributed. Let X denote an absolutely continuous nonnegative random variable with the cumulative distribution function (cdf) $F(\cdot)$ and the probability density function (pdf) $f(\cdot)$. By definition, for a continuous random variable X with pdf $f(x)$, the differential entropy is given by

$$H(f) = H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx = -\int_0^1 \log f(F^{-1}(u)) du, \tag{1}$$

where "log" stands for the natural logarithm and $F^{-1}(u) = \inf\{x : f(x) \geq u\}$ for all $0 < u < 1$. The last equality is obtained by using the probability integral transformation $U = F(X)$. Shannon entropy (1) measures lack of uniformity under $f(\cdot)$. It is more difficult to predict an outcome with a less concentrated distribution. Another useful measure of uncertainty for measuring the distance between two distributions is the Kullback-Leibler (KL) discrimination information of random variables X and Y with $pdfs$ $f(\cdot)$ and $g(\cdot)$, respectively, which is defined by

$$K(f : g) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx = -H(f) + H(f, g), \tag{2}$$

2010 *Mathematics Subject Classification*. Primary 62B10; Secondary 94A17
Keywords. Entropy, Kullback Leibler discrimination information, Mixture distribution, Normal distribution, Stochastic orders
Received: 05 October 2015; Revised: 03 May 2016; Accepted: 25 May 2016
Communicated by Miroslav M. Ristić
*Corresponding author
Email addresses: ab.toomaj@gonbad.ac.ir, ab.toomaj@gmail.com (Abdolsaeed Toomaj), r.zarei@guilan.ac.ir, rezazareirz@yahoo.com (Reza Zarei)

where $H(f,g) = -E_f[\log g(X)]$ is known as Fraser information (see, e.g., Ebrahimi *et al.* [6]) and also is known as "inaccuracy measure" due to Kerridge [10]. The KL discrimination information is always nonnegative and is zero if and only if $f(x) = g(x)$ almost everywhere. The KL information was first introduced by Kunllback and Leibler [11] to measure of the distance between two distributions. As the pdf of g is dissimilar or farther from the pdf of f, then $K(f,g)$ is large. For more information about the other applications of Shannon entropy and KL discrimination information, see also Asadi *et. al.* [1], Ebrahimi *et. al.* [6], Ebrahimi *et. al.* [7], among others. A detailed discussion about the Shannon entropy, properties and its applications can be found in Cover and Thomas [5]. Dynamic and bivariate versions can be seen in Asadi *et al.* [1], Chamany and Baratpour [4], Jomhoori and Yousefzadeh [9] and Navarro *et al.* [14].

The monograph given by Cover and Thomas [5] often do not provide analytic calculations of differential entropy for many probability distributions specially the mixture distributions. The aim of the present paper is to investigate information properties of finite mixture model. Finite means the number of random variables is finite. Let X_i, $i = 1, \cdots, n$ be a collection of n independent random variables. Suppose $F_i(\cdot)$ be the distribution function of X_i and assume $\underline{\alpha} = (\alpha_1, \cdots, \alpha_n)$ be the mixing probabilities. Then, the distribution of finite mixture random variable X_α with cdf $F_\alpha(\cdot)$ is defined by

$$F_\alpha(x) = P(X_\alpha \leq x) = \sum_{i=1}^{n} \alpha_i F_i(x), \ x \in \mathbb{R}, \tag{3}$$

where $\alpha_i \geq 0$, $\sum_{i=1}^{n} \alpha_i = 1$. If the random variable X_i is absolutely continuous, from (3) the pdf of X_α is given by

$$f_\alpha(x) = \sum_{i=1}^{n} \alpha_i f_i(x), \ x \in \mathbb{R}. \tag{4}$$

There are several papers about the entropy of mixture distributions in the literature; see, e.g., Tan *et al.* [20], Lu [12], Hild *et al.* [8], Rohde *et al.* [17], Poland and Shachter [16] and the references therein. Michalowicz *et al.* [13] provided an analytical expression for signal entropy in situations where the corrupting noise source is mixed-Gaussian. The rest of this paper is organized as follows: In Section 2, we provide an expression for the entropy of mixture distribution as well as bounds for it. Moreover, we derive an approximation for the given entropy. Some ordering properties of mixture distribution's entropy are discussed in Sections 3. For illustrative purposes, some examples are given. Finally, a brief conclusions are given in Section 4.

2. Entropy of Mixture Model

It is known that the entropy of mixture distributions generally cannot be calculated in the closed form due to the logarithm of a sum of mixture distribution functions, except for the special case of a single distribution which is computationally easy. Therefore, it is not easy to compute the exact value of entropy of mixture models. But, we provide an expression for the entropy of mixture model (4) by using the properties of entropy concept. It can be used to provide bounds and approximation for the entropy of mixture distributions. We have

$$\begin{aligned} H(f_\alpha) &= -\int_{-\infty}^{\infty} f_\alpha(x) \log f_\alpha(x) dx \\ &= \sum_{i=1}^{n} \alpha_i \left[-\int_{-\infty}^{\infty} f_i(x) \log f_\alpha(x) dx \right] \\ &= \sum_{i=1}^{n} \alpha_i H(f_i) + \sum_{i=1}^{n} \alpha_i K(f_i : f_\alpha). \end{aligned} \tag{5}$$

The second equality in (5) is derived from (4) and the linearity property of integration and the last equality is obtained from (2). Expression (5) can be used to compute the exact value of $H(f_\alpha)$ when the pdf (4) does

not have a complicated form. But, it is hard to find $H(f_\alpha)$ in many cases. Hence, bounds can be used in such situations. It is worth to point out that (5) provides lower and upper bounds as described in the sequel. Since $\sum_{i=1}^{n} \alpha_i K(f_i : f_\alpha) \geq 0$, then we have

$$H(f_\alpha) \geq H_L(f_\alpha) = \sum_{i=1}^{n} \alpha_i H(f_i). \tag{6}$$

It is known that the probability integral transformation $U_i = F_i(X_i),\ i = 1, \cdots, n$ are uniformly distributed in $[0, 1]$. Then the lower bound of $H(f_\alpha)$ in (6) becomes

$$H_L(f_\alpha) = \sum_{i=1}^{n} \alpha_i H(f_i) = \sum_{i=1}^{n} \alpha_i E[\log f_i(F_i^{-1}(U_i))].$$

It is worth to mention that the lower bound (6) can also be obtained by using the concavity property of entropy. One can see that the computation of the right-hand-side of (6) is easy as linear function of single distribution's entropy. To provide an upper bound, we use the convexity property of the KL discrimination information. From (5) we have

$$\begin{aligned} H(f_\alpha) &= H_L(f_\alpha) + \sum_{i=1}^{n} \alpha_i K(f_i : f_\alpha) \\ &\leq H_L(f_\alpha) + \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(f_i : f_j). \end{aligned} \tag{7}$$

As an applications of the given bounds, consider the following example.

Example 2.1. Let X_i have the Normal distribution with mean 0 and variance σ_i^2 i.e. $X_i \sim N(0, \sigma_i^2),\ i = 1, 2, \cdots, n$. It is well-known the differential entropy of Normal distribution is

$$H(f_i) = \frac{1}{2} \log(2\pi e \sigma_i^2),\ i = 1, 2, \cdots, n.$$

Moreover, it is not hard to verify that for all $i, j = 1, 2, \cdots, n$,

$$K(f_i : f_j) = \log\left(\frac{\sigma_i}{\sigma_j}\right) + \frac{\sigma_i^2}{2\sigma_j^2} - \frac{1}{2}.,$$

where σ_i and σ_j are the standard deviation. Therefore from (6), the Gaussian-mixture entropy is bounded as follows

$$\frac{1}{2} \sum_{i=1}^{n} \alpha_i \log(2\pi e \sigma_i^2) \leq H(f_\alpha) \leq \frac{1}{2} \sum_{i=1}^{n} \alpha_i \log(2\pi e \sigma_i^2) + \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \left(\log\left(\frac{\sigma_i}{\sigma_j}\right) + \frac{\sigma_i^2}{2\sigma_j^2} \right) - \frac{1}{2}. \ \square$$

The critical part of calculation of the mixture model's entropy is the logarithm of the pdf $f_\alpha(x)$ in (4). To obtain an accurate and versatile entropy approximation that can be evaluated analytically, we use the given bounds in (6) and (7). In fact, the bounds themselves can be used for efficiently approximating the true entropy value. It is worth calculating of the average of the lower and upper bounds i.e. the center of the interval. Hence, we define

$$\widetilde{H}(f_\alpha) = H_L(f_\alpha) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(f_i : f_j).$$

Since this value is between the lower and upper bounds of the differential entropy of $H(f_\alpha)$, it is an approximation of entropy of mixture distributions as reasonable measure. One should notice that the use of approximations is clearly advantageous from the computational complexity aspect.

Example 2.2. Following Example 2.1, the approximation of $H(f_\alpha)$ is

$$\widetilde{H}(f_\alpha) = \frac{1}{2}\left[\sum_{i=1}^{n}\alpha_i \log(2\pi e\sigma_i^2) + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\left(\log\left(\frac{\sigma_i}{\sigma_j}\right) + \frac{\sigma_i^2}{2\sigma_j^2}\right) - \frac{1}{2}\right]. \square$$

3. Ordering Properties

Hereafter, we provide some results on ordering properties of entropy of mixture distributions in terms of ordering properties of single distributions. First, we need the following definitions in which X and Y denote random variables with *cdfs* $F(\cdot)$ and $G(\cdot)$, *pdfs* $f(\cdot)$ and $g(\cdot)$, and survival functions $\bar{F}(x) = 1 - F(x)$ and $\bar{G}(x) = 1 - G(x)$, respectively.

Definition 3.1. A random variable X is said to have a decreasing failure rate (*DFR*) if the hazard function $h_X(t) = f(t)/\bar{F}(t)$ is decreasing in $t \in \mathbb{R}$.

Definition 3.2. Suppose that X and Y be a two random variables with the *cdfs* $F(\cdot)$ and $G(\cdot)$, respectively.

(i) A random variable X is said to be smaller than Y in the usual stochastic order (denoted by $X \leq_{st} Y$) if $\bar{F}(t) \leq \bar{G}(t)$ for all $t \in \mathbb{R}$.

(ii) A random variable X is said to be smaller than Y in the hazard rate order (denoted by $X \leq_{hr} Y$) if $h_X(t) \geq h_Y(t)$ for all $t \in \mathbb{R}$.

(iii) A random variable X is said to be smaller than Y in the dispersive order (denoted by $X \leq_{disp} Y$) if $F^{-1}(u) - F^{-1}(v) \leq G^{-1}(u) - G^{-1}(v)$, $\forall\ 0 < v \leq u < 1$, and $F^{-1}(\cdot)$ and $G^{-1}(\cdot)$ be right continuous inverses of $F(\cdot)$ and $G(\cdot)$, respectively, or equivalently $g(G^{-1}(v)) \leq f(F^{-1}(v)),\ 0 \leq v \leq 1$.

Definition 3.3. A random variable X is said to be smaller than Y in the entropy order (denoted by $X \leq_e Y$) if $H(X) \leq H(Y)$.

It is well-known that $X \leq_{disp} Y$ implies $X \leq_e Y$; see Oja [15]. Now, we derive some results about the stochastic ordering entropy of mixture models. First, in the following theorem, we suppose the case that two mixture distributions have the same probability vectors but they constructed based on different distributions.

Theorem 3.4. *Let X_i and Y_i, $i = 1, \cdots, n$ be a collection of random variables with cdfs $F_i(\cdot)$ and $G_i(\cdot)$, respectively. Also, assume X_α and Y_α be two mixture random variables with cdfs $F_\alpha(\cdot)$ and $G_\alpha(\cdot)$, respectively. If $X_i \leq_{st} Y_i$ and Y_i is DFR for all $i = 1, \cdots, n$, then $X_\alpha \leq_e Y_\alpha$.*

Proof. Since $X_i \leq_{st} Y_i$, it is not hard to see that $X_\alpha \leq_{st} Y_\alpha$. On the other hand Y_α is *DFR* provided that Y_i is *DFR* (see Barlow and Proshcan [3]). Therefore, Theorem 2.2 of Ebrahimi *et al.* [7] concluded that $X_\alpha \leq_e Y_\alpha$. $\square$

In the next theorem, we extend the preceding results to the different probability vectors and distributions.

Theorem 3.5. *Let X_i and Y_i, $i = 1, \cdots, n$ be a collection of random variables with cdfs $F_i(\cdot)$ and $G_i(\cdot)$, respectively. Also, assume $\alpha = (\alpha_1, \cdots, \alpha_n)$ and $\beta = (\beta_1, \cdots, \beta_n)$ be two mixing probability vectors and X_α and Y_β be two mixture random variables with cdfs $F_\alpha(\cdot)$ and $G_\beta(\cdot)$, respectively. If for all $i = 1, \cdots, n$,*

(i) *$X_i \leq_{st} Y_i$ and $X_i \leq_{st} X_{i+1}$,*

(ii) *$\alpha \leq_{st} \beta$,*

(iii) *Y_i is DFR,*

then $X_\alpha \leq_e Y_\beta$.

Proof. Let X_β denote the random mixtures of $X_1, \cdots, X_n$ with mixing probability vector $\beta = (\beta_1, \cdots, \beta_n)$. The condition $X_i \leq_{st} X_{i+1}$ means that $\bar{F}_i(x)$ is an increasing function of i. Let us suppose that $\bar{F}_\alpha(x)$ and $\bar{G}_\alpha(x)$ be the survival functions of X_α and X_α, respectively. We have

$$\bar{F}_\alpha(x) = \sum_{i=1}^{n} \alpha_i \bar{F}_i(x) \leq \sum_{i=1}^{n} \beta_i \bar{F}_i(x) \leq \sum_{i=1}^{n} \beta_i \bar{G}_i(x) = \bar{G}_\beta(x), \ x \in R.$$

The first inequality is obtained from Theorem 1.A. 3 of Shaked and Shanthikumar [18] by noting that $\bar{F}_i(x)$ is an increasing function of i and $\alpha \leq_{st} \beta$. The last inequality is obtained from the assumption $X_i \leq_{st} Y_i$. On the other hand Y_β is *DFR* provided that Y_i is *DFR*. Therefore, Theorem 2.2 of Ebrahimi *et al.* [7] completes the proof. □

As a special case of Theorem 3.5, we suppose that $X_i \stackrel{d}{=} Y_i$, $i = 1, \cdots, n$, where notation $\stackrel{d}{=}$ means equality in distribution.

Corollary 3.6. *Let X_i be a collection of random variables with cdfs $F_i(\cdot)$. Assume $\underline{\alpha} = (\alpha_1, \cdots, \alpha_n)$ and $\underline{\beta} = (\beta_1, \cdots, \beta_n)$ be two mixing probability vectors and X_α and X_β be two mixture random variables with cdfs $F_\alpha(\cdot)$ and $F_\beta(\cdot)$, respectively. If for all $i = 1, \cdots, n$,*

(i) $X_i \leq_{st} X_{i+1}$,

(ii) $\underline{\alpha} \leq_{st} \underline{\beta}$,

(iii) X_i *is DFR,*

then $X_\alpha \leq_e X_\beta$.

As an application of the preceding corollary, consider the following example.

Example 3.7. Let X_i have the *exponential* distribution with mean λ_i for all $i = 1, \cdots, n$ which is *DFR*. Assume that $\lambda_1 \geq \cdots \geq \lambda_n$, then $X_1 \leq_{st} \cdots \leq_{st} X_n$. As an application of Corollary 3.6 immediately yields $X_\alpha \leq_e X_\beta$ for every two probability vectors $\underline{\alpha}$ and $\underline{\beta}$ such that $\underline{\alpha} \leq_{st} \underline{\beta}$. □

In the forthcoming theorem, we provide the same result of Corollary 3.6 for the hazard rate which is a stronger stochastic order. First of all, we need the following theorem due to Bagai and Kochar [2].

Theorem 3.8. *Let X and Y be two random variables with the cdfs $F(\cdot)$ and $G(\cdot)$, respectively. If $X \leq_{hr} Y$ and either X or Y is DFR, then $X \leq_{disp} Y$.*

Theorem 3.9. *Suppose that X_i be a collection of random variables with cdfs $F_i(\cdot)$ in which it is DFR ($i = 1, \cdots, n,$). Assume $\underline{\alpha} = (\alpha_1, \cdots, \alpha_n)$ and $\underline{\beta} = (\beta_1, \cdots, \beta_n)$ be two probability vectors. Let X_α and X_β be two mixture random variables with cumulative distribution functions $F_\alpha(\cdot)$ and $F_\beta(\cdot)$, respectively. If $X_i \leq_{hr} X_{i+1}$ for all $i = 1, \cdots, n$, and $\underline{\alpha} \leq_{hr} \underline{\beta}$, then $X_\alpha \leq_e X_\beta$.*

Proof. Since $X_i \leq_{hr} X_{i+1}$ and $\underline{\alpha} \leq_{hr} \underline{\beta}$, by Theorem 1.B.14 of Shaked and Shanthikumar [18], we have $X_\alpha \leq_{hr} X_\beta$. On the other hand X_α and X_β are *DFR* provided that X_i is *DFR*. Hence Theorem 3.8 implies $X_\alpha \leq_{disp} X_\beta$ which this means that $X_\alpha \leq_e X_\beta$ and thus the desired result follows. □

4. Conclusion

In this paper, we provided some new results on information properties of mixture distributions. Moreover, we derived bounds for the entropy of mixture models as well as the approximation on it. In the sequel, we provided some stochastic ordering properties of entropy of mixture distributions in terms of single models.

Acknowledgment

The authors would like to thank the anonymous referee for careful reading of the paper and many helpful suggestions which improved the earlier version of this manuscript.

References

[1] Asadi, M., Ebrahimi, N., Hamedani, G. G. and Soofi, E. S. (2006). Information measures for pareto distributions and order statistics, in *Advances on Distribution Theory and Order Statistics*, N. Balakrishnan, E. Castillo, and J.M. Sarabia, eds. Boston: Birkhauser, pp. 207-223.
[2] Bagai, l. and Kochar, S.C. (1986) On tail ordering and comparison of failure rates. *Communications in Statistics: Theory and Methods.* **15**, pp. 1377-1388.
[3] Barlow, R. E. and Proschan, F. (1981) *Statistical Theory of Reliability and Life testing*. Silver Spring, MD: To Begin With.
[4] Chamany, A. and Baratpour, S. (2014) A dynamic discrimination information based on cumulative residual entropy and its properties. *Communications in Statistics: Theory and Methods*. **43**(6), pp. 1041-1049.
[5] Cover, T. A. and Thomas, J. A. (2006) *Elements of Information Theory*. New Jersey: Wiley and Sons, Inc.
[6] Ebrahimi, N., Soofi, E. S. and Soyer, R. (2010) Information measures in perspective. *International Statistical Review.* **78**, pp. 383-412.
[7] 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.
[8] Hild, K. E., Pinto, D., Erdogmus, D. and Principe, J. C. (2005) Convolutive blind source separation by minimizing mutual information between segments of signals. *IEEE Transactions on Circuits and Systems I.* **52**, pp. 2188-2196.
[9] Jomhoori, S. and Yousefzadeh, F. (2014) On estimating the residual Renyi entropy under progressive censoring. *Communications in Statistics: Theory and Methods.* **43** (10-12), pp. 2395-2405.
[10] Kerridge, D.F. (1961) Inaccuracy and inference. *Journal of the Royal Statistical Society, Series B.* **23**, pp. 184-194.
[11] Kullback, S. and Leibler, R.A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics.* **22**, pp. 79-86.
[12] Lu, Z. W. (2007) An iterative algorithm for entropy regularized likelihood learning on gaussian mixture with automatic model selection. *Neurocomputing.* **69**, pp. 1674-1677.
[13] Michalowicz, J. V., Nichols, J. M. and Bucholtz, F. (2008) Calculation of differential entropy for a mixed gaussian distribution. *Entropy.* **10**, pp. 200-206.
[14] Navarro, J., Sunoj, S. M. and Linu, M. N. (2014) Characterizations of bivariate models using some dynamic conditional information divergence measures. *Communications in Statistics: Theory and Methods.* **43**(9), pp. 1939-1948.
[15] Oja, H. (1981) On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics.* **8**, pp. 154-168.
[16] Poland, W. B. and Shachter, R. D. (1993) Mixtures of gaussians and minimum relative entropy techniques for modeling continuous uncertainties. In Uncertainty in Artificial Intelligence: Proceedings of the Ninth Conference: 183-190. San Mateo, CA: Morgan Kaufmann.
[17] Rohde, G. K., Nichols, J. M., Bucholtz, F., and Michalowicz, J. V. (2007) Signal estimation based on mutual information maximization. In Forty-First Asilomar Conference on Signals, Systems, and Computers, IEEE.
[18] Shaked, M. and Shanthikumar, J. G. (2007) *Stochastic Orders*. Springer Verlag, New York.
[19] Shannon, C. E. (1948) A mathematical theory of communication. *The Bell System Technical Journal.* **27**, pp. 379-423 and 623-656.
[20] Tan, Y., Tantum, S. L., and Collins, L. M. (2004) Cramer-Rao lower bound for estimating quadrupole resonance signals in non-gaussian noise. *IEEE Signal Processing Letters.* **11**, pp. 490-493.