Copula based Probabilistic Measures of Uncertainty with Applications

Kumar, Pranesh
University of Northern British Columbia, Department of Mathematics and Statistics
Prince George, BC, Canada, V2N 4Z9
E-mail: kumarp@unbc.ca

1. Introduction

Probabilistic uncertainty may be viewed as one associated with a random outcome of an experiment or which is associated with the manner in which data are collected and analyzed following statistical designs. Often such type of uncertainty is summarized in terms of bias, standard error, and measures based on the statistical probability distributions. Shannon [12] laid the mathematical foundation of information theory in the context of communication theory and defined a probabilistic measure of uncertainty referred to as entropy. However earlier contributions in this direction have been due to Nyquist [11] and Hartley [4].

Let two random variables be X and Y with respective marginal probability distributions $(x_i, p_i, i = 1, .. m; \sum_i p_i = 1)$ and $(y_j, q_j, j = 1, .. n; \sum_i q_i = 1)$ and the joint probability distribution $(x_i, y_j, p_{ij}; \sum_{ij} p_{ij} = 1)$ where $p_{ij} \neq 0$ is the probability of a pair (x_i, y_j) belonging to the rectangle $R_i : [x^*_{i-1}; x^*_i] \times C_j : [y^*_{j-1}; y^*_j]$ following the partitioning of codomain of X and Y.

The measure of uncertainty associated with the variable X, called entropy, is defined as $H(X) = -\sum_i p_i \log p_i$. The uncertainty takes the maximum value when all probabilities are equal, i.e., $p_i = 1/m$. The bounds for $H(X)$ are: $0 \leq H(X) \leq \log m$. The joint entropy of X and Y is defined as

$$H(X, Y) = -\sum_{ij} p_{ij} \log p_{ij}. \quad (1.1)$$

In case X and Y are independent, $p_{ij} = p_i \; q_j$ and the entropy of the joint distribution equals the sum of respective entropies of X and Y, i.e., $H(X, Y) = H(X) + H(Y)$. The conditional entropy $H(X \mid Y)$ which is the amount of uncertainty of X remaining given advance knowledge of Y and is

$$H(X|Y) = -\sum_{ij} p_{ij} \log p_{i/j}, \quad (1.2)$$

where $p_{i/j}$ is the conditional probability of X taking a value x_i given that Y has assumed a

value y_j. The quantity $I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X,Y) - H(X|Y) - H(Y|X)$ is called the mutual information or information transmission (distance from statistical independence) between X and Y. Mutual information in terms of the Kullback-Liebler divergence between joint distribution and the two marginal distributions [8] is defined as

$$I(X,Y) = \sum_{ij} p_{ij} \log\left(p_{ij}/p_i q_j\right), \quad (1.3)$$

To transmit X, how many bits on average would it save if both ends of the line knew Y? We can answer this question by calculating information gain

$$IG(X|Y) = H(X) - H(X|Y). \quad (1.4)$$

or the relative information gain [9]:

$$r(X|Y) = I(X,Y)/H(X) = 1 - H(X|Y)/H(X) = I(X,Y)/[H(X,Y) - H(Y|X)], \quad (1.5)$$

which shows how much information about Y diminishes the uncertainty of X relative to the initial uncertainty of X. A symmetrical relative information gain measure is defined by

$$R(X,Y) = 2I(X,Y)/[H(X) + H(Y)] = 2I(X,Y)/[H(X,Y) + I(X,Y)\,, \quad (1.6)$$

which expresses the uncertainty from the joint distribution of X and Y to the uncertainty in case of independence.

2. Copula Functions

Sklar's theorem [13] states that any multivariate distribution can be expressed as the k-copula function $C(u_1,\ldots,u_i,\ldots,u_k)$ evaluated at each of the marginal distributions. Copula is not unique unless the marginal distributions are continuous. Using probability integral transform, each continuous marginal $U_i = F_i(x_i)$ has a uniform distribution on $I\in[0,1]$ where $F_i(x_i)$ is the cumulative integral of $f_i(x_i)$ for the random variable $X_i \in (-\infty,\infty)$. The k-dimensional probability distribution function F has a unique copula representation

$$F(x_1,x_2,\ldots,x_k) = C(F_1(x_1),F_2(x_2),\ldots,F_k(x_k)) = C(u_1,u_2,\ldots,u_k). \quad (2.1)$$

The joint probability density function is written as

$$f(x_1,x_2,\ldots,x_k) = \Pi_{i=1}^{k} f_i(x_i) \times c(F_1(x_1),F_2(x_2),\ldots,F_k(x_k)), \quad (2.2)$$

where $f_i(x_i)$ is each marginal density and coupling is provided by copula density

$$c(u_1,u_2,\ldots,u_k) = \partial^k C(u_1,u_2,\ldots,u_k)/\,\partial u_1 \partial u_2 \ldots \partial u_k, \quad (2.3)$$

if it exists. In case of independent random variables, copula density $c(u_1,u_2,\ldots,u_k)$ is identically equal to one. The importance of the above equation $f(x_1,x_2,\ldots,x_k)$ is that the independent portion expressed as the product of the marginals can be separated from the function $c(u_1,u_2,\ldots,u_k)$ describing the dependence structure or shape. The dependence structure summarized by a copula is invariant under increasing and continuous transformations of the marginals.

The simplest copula is independent copula

$$\Pi: = C(u_1,u_2,\ldots,u_k) = u_1 u_2 \ldots u_k, \quad (2.4)$$

with uniform density functions for independent random variables. An empirical copula may be

estimated from the N pairs of data $\{(x_{1;t}, x_{2;t})\}_{0<t\leq N}$ by

$$C(n/N, m/N) = \sum_t 1_{\{r_{t,1}\leq n, r_{t,2}\leq m\}}, \quad (2.5)$$

where $r_{t,1}$ and $r_{t,2}$ are the rank statistics of $\{x_{1;t}\}_t$ and $\{x_{2;t}\}_t$ respectively.

Frécht [3]–Hoeffding [5] lower and upper bounds for copula respectively are

$$W(u_1, u_2, \ldots, u_k) := \max\{1 - n + \sum_i u_i, 0\} \leq C(u_1, u_2, \ldots, u_k), \quad (2.6)$$

$$C(u_1, u_2, \ldots, u_k) \leq \min_{i\in\{1,2,\ldots,k\}} u_i =: M(u_1, u_2, \ldots, u_k). \quad (2.7)$$

Relationships between copula and concordance Kendall's τ, Spearman's ρ, Gini's index γ:

$$\tau = 4\int\int_{I^2} C(u_1, u_2)dC(u_1, u_2) - 1, \quad (2.8)$$

$$\rho = 12\int\int_{I^2} u_1 u_2 dC(u_1, u_2) - 3, \quad (2.9)$$

$$\gamma = 2\int\int_{I^2} (|u_1 + u_2 - 1| - |u_1 - u_2|)dC(u_1, u_2). \quad (2.10)$$

The Pearson's linear correlation coefficient can not be expressed in terms of copula. The tail dependence index of a multivariate distribution describes the amount of dependence in the upper right tail or lower left tail of the distribution and can be used to analyze the dependence among extreme random events. Joe [6] defines the upper and lower tail dependence

$$\lambda_U := \lim_{u\to 1}[\{1 - 2u + C(u, u)\}/(1 - u)], \quad (2.11)$$

$$\lambda_L := \lim_{u\to 0}[C(u, u)/u]. \quad (2.12)$$

Copula has lower (upper) tail dependence for λ_L $(\lambda_U) \in (0,1]$ and no lower (upper) tail dependence for λ_L $(\lambda_U) = 0$. This tail dependence measure is the probability that one variable is extreme given that other is extreme. Tail dependence measures are copula-based and copula is related to the full distribution via quantile transformations, i.e., $C(u_1, u_2) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right)$.

Copulas can be simulated using univariate conditional distributions. The conditional distribution of U_i given first $i - 1$ components is

$$c(u_i|u_1, \ldots, u_{i-1}) = \frac{\partial^{i-1}C(u_1,\ldots,u_i)}{\partial u_1 \ldots \partial u_{i-1}} \Big/ \frac{\partial^{i-1}C(u_1,\ldots,u_{i-1})}{\partial u_1 \ldots \partial u_{i-1}}. \quad (2.13)$$

For $k \geq 2$, simulation procedure is: (i) Select a random number u_1 from Uniform (0,1) distribution and (ii) Simulate a value u_k from $c(u_k|u_1, \ldots, u_{k-1})$, $k = 2,3, \ldots$

3. Copula based Uncertainty Measures

The entropy measures associated with the joint distribution of X and Y using copula density function $c(u_1,u_2)$ can be expressed:

$$H(X,Y) = -\sum_{ij} c(u_1,u_2)\log\, c(u_1,u_2). \quad (3.1)$$

$$H(X|Y) = -\sum_{ij} c(u_1,u_2)\log\, c(u_1|u_2). \quad (3.2)$$

$$I(X,Y) = -\sum_{ij} c(u_1,u_2)\log[c(u_1,u_2)/\{c(u_1|u_2)\times c(u_2|u_1)\}]. \quad (3.3)$$

$$r(X|Y) = \frac{\sum_{ij} c(u_1,u_2)\log[c(u_1,u_2)/\{c(u_1|u_2)\times c(u_2|u_1)\}]}{\sum_{ij} c(u_1,u_2)\log[c(u_1,u_2)/c(u_2|u_1)]}, \quad (3.4)$$

$$R(X,Y) = \frac{2\sum_{ij} c(u_1,u_2)\log[c(u_1,u_2)/\{c(u_1|u_2)\times c(u_2|u_1)\}]}{\sum_{ij} c(u_1,u_2)\log[c^2(u_1,u_2)/\{c(u_1|u_2)\times c(u_2|u_1)\}]}. \quad (3.5)$$

Kovacs [9] has established

$$\int_{u_{i-1}^*}^{u_i^*}\int_{v_{j-1}^*}^{v_j^*} \frac{\partial^2 C(u,v)}{\partial u\ \partial v}\,du\,dv = C(u_i^*,v_j^*) - C(u_{i-1}^*,v_j^*) - C(u_i^*,v_{j-1}^*) + C(u_{i-1}^*,v_{j-1}^*), \quad (3.6)$$

where $u_i^* = F_1(x_i^*)$ and $v_j^* = F_2(y_j^*)$.

4. Family of the Marshall-Olkin Bivariate Copulas

Two parameters family of the Marshall-Olkin bivariate copula [9] for $u,v,\alpha,\beta \in (0,1)$ is

$$C(u,v) = \min\left(u^{1-\alpha}v, uv^{1-\beta}\right) = \begin{cases} u^{1-\alpha}v, & u^\alpha \geq v^\beta, \\ uv^{1-\beta}, & u^\alpha \leq v^\beta. \end{cases} \quad (4.1)$$

and the copula density function

$$c(u,v) = \begin{cases} u^{-\alpha}, & u^\alpha > v^\beta, \\ v^{-\beta}, & u^\alpha < v^\beta. \end{cases} \quad (4.2)$$

For this family of copula, $\tau = \frac{\alpha\beta}{\alpha-\alpha\beta+\beta}$.

When $\alpha=\beta=\theta$, two parameter $C(u,v)$ reduces to the Cuadras-Augé [2] family of copulas with one parameter. Mercier [10] has shown that the mutual information $I(X,Y)$for the one parameter copula $C(u,v)$

$$I(X,Y) = -2\frac{1-\theta}{2-\theta}\left[\log(1-\theta) + \frac{\theta}{2-\theta}\right] \quad (4.3)$$

$$I(X,Y) = -(1-\tau)\ \left[\tau + \log\left(\frac{1-\tau}{1+\tau}\right)\right] \quad (4.4)$$

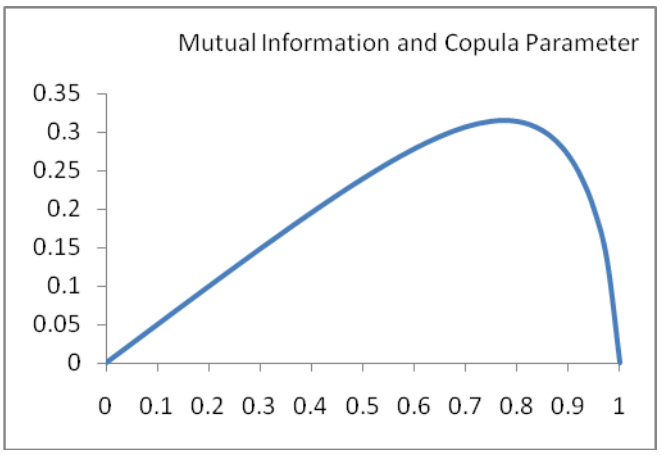

Figure 1. Mutual information and the Marshall-Olkin copula parameter.

5. Application of the Marshall-Olkin Copula based Uncertainty Measures

For predicting the gas consumption (Y: gallons/100 miles travelled) of an automobile from its size and engine characteristics, data [1] from 38 cars were collected on weight (X_1: 1000 lb.), engine displacement (X_2: cubic inches), number of cylinders (X_3) and horsepower (X_4).

Table 1. Mutual Information using the one-paramater Marshall-Olkin copula.

	Weight	Displacement	Number of Cylinders	Horsepower
Kendall's τ	0.7869	0.6605	0.7937	0.7310
Copula parameter θ	0.8807	0.7955	0.8850	0.8446
Mutual Information	0.2855	0.3147	0.2824	0.3042

Mutual information $I(X,Y)$ in Table 1 indicates that the decrease in uncertainty of the car gas consumption caused by the knowledge of its weight and number of cyliders is about the same followed by horsepower and displacement. Higher values of copula parameters is an indication of higher degree of association between gas consumption and weight and also with number of cylinders and can be selecetd as the best predictors. The best estimated prediction relationship is $\hat{y} = -1.639 + 2.333\ x_1 - 0.008x_2 + 0.218x_3; \text{adj}\ R^2 = 89.4\%$. However the limitation of the Marshall-Olkin copula is that its parameter lies on the interval (0,1) and thus models the positive association only.

This study was supported by the author's discovery grant from the *Natural Sciences and Engineering Research Council of Canada (NSERC)*.

REFERENCES

[1] Bovas, A. And Ledolter, J. *Introduction To Regression Modeling*, Thompson Books/Cole (2006), 13.

[2] Cuadras, C.M. and Augé, J. A continuous general multivariate distribution and its properties, *Comm, Statist. A –Theory & Methods*, 10, (1981), 339-353.

[3] Frécht, M. Sue les tableaux de corrélation dont les marges son données, *Ann. Univ. Lyon*, Sect. A, 9 (1951), 53-77.

[4] Hartley, R.V.L. (1928). Transformation of information. *Bell Systems Technical Journal*, 7, 535-563.

[5] Hoeffding, W. Masstabinvariance korrelationsmasse, Schriften des Mathematischen Instituts für Angewandte Mathematik der Universität Berlin,5,3 (1940),179-233.

[6] Joe, H. (1997). *Multivariate Models and Dependent Concepts*. New York: Chapman & Hall.

[7] Kovács, E. On the using of copulas in characterizing of dependence with entropy, Pollack *Periodica-International Journal from Engineering, Information Sciences*, 2007.

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

[9] Marshall, A.W. and Olkin, I. Families of multivariate distributions, *Journal of the American Statistical Association*, 83 (1988), 834-841.

[10] Mercier, G. *Measures de Dépendance entre Images RSO*, GET/ENST Bretagne, Tech. Rep. RR-2005003-ITI, 2005, http//:perso.enst-bretagne.fr/126mercierg

[11] Nyquist, H. (1928). Certain topics in telegraph transmission theory. *Trans. AIEE*, vol. 47, pp. 617-644. Reprint as classic paper in: Proc. IEEE, Vol. 90, No. 2, Feb 2002.

[12] Shannon, C.E. *A Mathematical Theory of Communication-An Integrated Approach*, Cambridge University Press, 1948.

[13] Sklar, A. Fonctions de répartition á *n* dimensional et leurs marges, *Publ. Inst. Stat. Univ. Paris*, 8 (1959), 229-231.

RÉSUMÉ (ABSTRACT)

Uncertainty emerges when there is less information than the total information required for describing a system or environment. Uncertainty prevails in several forms and various kinds of uncertainties may arise from random fluctuations, incomplete information, imprecise perception, vagueness etc. We consider information-theoretic measures and copula functions to characterize uncertainty associated with the probabilistic systems. Copula functions join uniform marginal distributions of random variables to form their multivariate distribution functions. Copulas are useful because they separate joint distributions into two contributions- (i) marginal distributions of each variable and (ii) copula as a measure of dependence. Several families of copulas with varying shapes and simulation programs are available providing flexibility in copula based modeling. We discuss the applications of Marshall-Olkin family of copulas based information measures to analyze uncertainty.