Journal of Statistics Applications & Probability
--- An International Journal

Statistical Dependence: Copula Functions and Mutual Information Based Measures

Pranesh Kumar

Department of Mathematics and Statistics, University of Northern British Columbia, Prince George, Canada
Email Address: kumarp@unbc.ca

Received: Dec. 11, 2011; Revised Feb. 12, 2012; Accepted Feb. 16, 2012
Published online: 1 April 2012

Abstract: Accurately and adequately modelling and analyzing relationships in real random phenomena involving several variables are prominent areas in statistical data analysis. Applications of such models are crucial and lead to severe economic and financial implications in human society. Since the beginning of developments in Statistical methodology as the formal scientific discipline, correlation based regression methods have played a central role in understanding and analyzing multivariate relationships primarily in the context of the normal distribution world and under the assumption of linear association. In this paper, we aim to focus on presenting notion of dependence of random variables in statistical sense and mathematical requirements of dependence measures. We consider copula functions and mutual information which are employed to characterize dependence. Some results on copulas and mutual information as measure of dependence are presented and illustrated using real examples. We conclude by discussing some possible research questions and by listing the important contributions in this area.

Keywords: Statistical dependence; copula function; entropy; mutual information; simulation.

1 Introduction

Understanding and modeling dependence in multivariate relationships has a pivotal role in scientific investigations. In the late nineteenth century, Sir Francis Galton [12] made a fundamental contribution to the understanding of multivariate relationships using regression analysis by which he established linkage of the distribution of heights of adult children to the distribution of their parents' heights. He showed not only that each distribution was approximately normal but also that the joint distribution could be described as a bivariate normal. Thus, the conditional distribution of adult children's height given the parents' height could also be modeled by using normal distribution. Since then regression analysis has been developed as the most widely practiced statistical technique because it permits to analyze the effects of explanatory variables on response variables. However, although widely applicable, regression analysis is limited chiefly because its basic setup requires identifying one dimension of the outcome as the primary variable of interest, dependent variable, and other dimensions as independent variables affecting dependent variable. Since this may not be of primary interest in many applications, focus should be on the more basic problem of understanding the distribution of several outcomes of a multivariate distribution. Normal distribution is most useful in describing one-dimensional data and has long dominated the studies involving multivariate distributions. Multivariate normal distributions are appealing because their marginal distributions are also normal and the association between any two random variables can be fully described knowing only the marginal distributions and an additional dependence parameter measured by the Pearson's linear correlation coefficient. However there are many situations where normal distributions fail to provide an adequate approximation to a given situation. For that reason many families of non-normal distributions have been developed mostly as immediate extensions of univariate distributions. However such a construction suffers from that a different family is

needed for each marginal distribution, extensions to more than just the bivariate case are not clear and measures of dependence often appear in the marginal distributions.

In this paper we focus on the notion of dependence of random variables in statistical sense and mathematical requirements of dependence measures. We describe copula functions and mutual information which can be alternatively used to characterize dependence. Some results on measuring dependence using copulas and mutual information are presented. We illustrate applications of these dependence measures with the help of two real data sets. Lastly we conclude by discussing some possible research questions and by listing some important contributions on this topic.

2 Statistical Dependence Measures

The notion of Pearson correlation ρ in Statistical methodology has been central in understanding dependence among random statistical variables. Although correlation is one of the omnipresent concepts, it is also one of the most misunderstood correlation concepts. The confusion may arise from the literary meaning of the word to cover any notion of dependence. From mathematicians' perspective, correlation is only one particular measure of stochastic dependence. It is the canonical measure in the world of multivariate *normal* distributions and in general for *spherical* and *elliptical* distributions. However it is well known fact that in numerous applications, distributions of the data seldom belong to this class. The correlation coefficient ρ between a pair of real-valued non-degenerate random variables X and Y with corresponding finite variances σ_x^2 and σ_y^2 is the standardized covariance σ_{xy}, i.e., $\rho = \sigma_{xy}/\,\sigma_x\,\sigma_y$, $\rho\epsilon[-1,1]$. The correlation coefficient is a measure of linear dependence only. In case of independent random variables, correlation is zero. In case of imperfect linear dependence, misinterpretations of correlation are possible [6,7,10]. Correlation is not in general an ideal dependence measure and causes problems when distributions are heavy-tailed. Some examples of commonly used heavy-tailed distributions are: One-tailed (Pareto distribution, Log-normal distribution, Lévy distribution, Weibull distribution with shape parameter less than one, Log-Cauchy distribution) and two-tailed (Cauchy distribution, family of stable distributions excepting normal distribution within that family, *t*- distribution, skew lognormal cascade distribution). Independence of two random variables implies they are uncorrelated but zero correlation does not in general imply independence. Correlation is not invariant under strictly increasing linear transformations. Invariance property is desirable for the statistical estimation and significance testing. Additionally, correlation is sensitive to outliers in the data set. The popularity of linear correlation and correlation based models is primarily because being expressed in terms of moments it is often straightforward to calculate and manipulate them under algebraic operations. For many bivariate distributions it is simple to calculate variance and covariance and hence the correlation coefficient. Another reason for the popularity of correlation is that it is a natural measure of dependence in multivariate *normal* distributions and more generally in multivariate *spherical* and *elliptical* distributions. Some examples of densities in the spherical class are those of the multivariate *t*-distribution and the logistic distribution. Another class of dependence measures is rank correlations distributions. Rank correlations are used to measure correspondence between two rankings and assess their significance. Two commonly used rank correlation measures are Kendall's τ and Spearman's ρ_s. Assuming random variables X and Y have distribution functions $F(x)$ and $F(y)$, Spearman's rank correlation $\rho_s = \rho\big(F(x), F(y)\big)$. If (X_1, Y_1) and (X_2, Y_2) are two independent pairs of random variables, then the Kendall's rank correlation is $\tau = \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0]$. The main advantage of rank correlations over linear correlation is that they are invariant under monotonic transformations. However rank correlations do not lend themselves to the same elegant variance-covariance manipulations as linear correlation does since they are not moment-based.

A measure of dependence, like linear correlation, summarizes the dependence structure of two random variables in a single number. Another excellent discussion of dependence measures is in the paper by Embrecht, McNeil and Straumann [7]. Let $D(X, Y)$ be a measure of dependence which assigns a real Embrechts number to any real-valued pair of random variables (*X*, *Y*). Then dependence measure *D*(*X*,*Y*) is

desired to have properties: (i) Symmetry: $D(X,Y) = D(Y,X)$; (ii) Normalization:$-1 \leq D(X,Y) \leq 1$; (iii) Comonotonic or Countermonotonic: The notion of comonotonicity in probability theory is that a random vector is comonotonic if and only if all marginals are non-decreasing functions (or non-increasing functions) of the same random variable. A measure $D(X,Y)$ is comonotonic if $D(X,Y) = 1 \Leftrightarrow X,Y$ or countermonotonic if $D(X,Y) = -1 \Leftrightarrow X,Y$; (iv) For a transformation T strictly monotonic on the range of X, $D((T(X),Y) = D(X,Y)$, $T(X)$ is increasing or decreasing. Linear correlation ρ satisfies properties (i) and (ii) only. Rank correlations fulfill properties (i) - (iv) for continuous random variables X and Y. Another desirable property is: (v) $D(X,Y) = 0 \Leftrightarrow X,Y$ (Independent). However it contradicts property (iv). There are no dependence measure satisfying both properties (iv) and (v). If we desire property (v), we should measure dependence $0 \leq D^*(X,Y) \leq 1$. The disadvantage of all such dependence measures $D^*(X,Y)$ is that they cannot differentiate between positive and negative dependence [27, 49].

3 Copula Functions

Multivariate distributions where *normal* distributions fail to provide an adequate approximation can be constructed by employing the copula functions. Copula functions have emerged in mathematical finance, statistics, extreme value theory and risk management as an alternative approach for modeling multivariate dependence. Every major statistics software package like Splus, R, Mathematica, MatLab, etc. includes a module to fit copulas. The International Actuarial Association recommends using copulas for modeling dependence in insurance portfolios. Copulas are now standard tools in credit risk management.

A theorem due to Sklar [49] states that under very general conditions, for any joint cumulative probability distribution function (CDF), $F(x_1, \ldots, x_k)$, there is a function $C(.)$ known as the copula function such that the joint CDF can be partitioned as a function of the marginal CDFs, $F(x_i)$. The converse is also true that this function couples any set of marginal CDFs to form a multivariate CDF.

3.1 Copula: Definition and Properties

The k- dimensional probability distribution function F has a unique copula representation

$$F(x_1, x_2, \ldots, x_k) = C\big(F_1(x_1), F_2(x_2), \ldots, F_k(x_k)\big) = C(u_1, u_2, \ldots, u_k). \tag{3.1}$$

The joint probability density function in copula form is written as

$$f(x_1, x_2, \ldots, x_k) = \Pi_{i=1}^{k} f_i(x_i) \times c(F_1(x_1), F_2(x_2), \ldots, F_k(x_k)), \tag{3.2}$$

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

$$c(u_1, u_2, \ldots, u_k) = \partial^k C(u_1, u_2, \ldots, u_k) / \partial u_1 \partial u_2 \ \ldots \ \partial u_k, \tag{3.3}$$

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

$$\Pi := C(u_1, u_2, \ldots, u_k) = u_1 u_2 \ldots u_k, \tag{3.4}$$

with uniform density functions for independent random variables on [0,1]. The Frécht-Hoeffding bounds for copulas [10]: The lower bound for k-variate copula is

$$W(u_1, u_2, \ldots, u_k) := \max\left\{1 - n + \sum_i u_i, 0\right\} \leq C(u_1, u_2, \ldots, u_k). \tag{3.5}$$

The upper bound for k-variate copula is given by

$$C(u_1, u_2, \ldots, u_k) \leq \min_{i \in \{1,2,\ldots,k\}} u_i =: M(u_1, u_2, \ldots, u_k). \tag{3.6}$$

For all copulas, the inequality $W(u_1, \ldots, u_k) \leq C(u_1, \ldots, u_k) \leq M(u_1, \ldots, u_k)$ must be satisfied. This inequality is well known as the Frécht-Hoeffding bounds for copulas. Further, W and M are copulas themselves. It may be noted that the Frécht-Hoeffding lower bound is not a copula in dimension $k > 2$. Copulas M, W and Π have important statistical interpretations [43]. Given a pair of continuous random variables (X_1, X_2), copula of (X_1, X_2) is $M(u_1, u_2)$ if and only if each of X_1 and X_2 is almost surely increasing function of the other; copula of (X_1, X_2) is $W(u_1, u_2)$ if and only if each of X_1 and X_2 is almost surely decreasing function of the other and copula of (X_1, X_2) is $\Pi(u_1, u_2) = u_1 u_2$ if and only if X_1 and X_2 are independent.

3.2 Copula and Rank Correlations

In case of non-elliptical distributions, it is better not to use Pearson correlation. Alternatively, we use rank correlation measures like Kendall's τ, Spearman's ρ_s and Gini's index γ. Rank correlations are invariant under monotone transformations and measure concordance. Under normality, there is one-to-one relationship between these measures [29].

$$\rho = \mathrm{Sin}\frac{\pi\tau}{2}, \tag{3.7}$$

$$\rho = 2\,\mathrm{Sin}\frac{\pi\,\rho_s}{6}. \tag{3.8}$$

Kendall's τ, Spearman's ρ_s and Gini's index γ could be expressed in terms of copulas [45,50]:

$$\tau = 4\int\int_{\mathrm{I}^2} C(u_1, u_2)\, dC(u_1, u_2) - 1, \tag{3.9}$$

$$\rho_s = 12\int\int_{\mathrm{I}^2} u_1 u_2\, dC(u_1, u_2) - 3, \tag{3.10}$$

$$\gamma = 2\int\int_{\mathrm{I}^2} (|u_1 + u_2 - 1| - |u_1 - u_2|)\, dC(u_1, u_2). \tag{3.11}$$

It may be noted however that the Pearson's linear correlation coefficient can not be expressed in terms of copula.

3.3 Copula and Tail Dependence Measures

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. Tail dependence describes the limiting proportion that one margin exceeds a certain threshold given that the other margin has already exceeded that threshold. Upper tail dependence of a bivariate copula $C(u_1, u_2)$ is defined by [22]

$$\lambda_U := \lim_{u \to 1}\left[\frac{\{1 - 2u + C(u,u)\}}{1-u}\right]. \tag{3.12}$$

If it exists, then $C(u_1, u_2)$ has upper tail dependence for $\lambda_U \in (0,1]$ and no upper tail dependence for $\lambda_U = 0$. Similarly, lower tail dependence in terms of copula is defined

$$\lambda_L := \lim_{u \to 0}[C(u,u)/u]. \tag{3.13}$$

Copula has lower tail dependence for $\lambda_L \in (0,1]$ and no lower tail dependence for $\lambda_L = 0$. This measure is extensively used in extreme value theory. It is the probability that one variable is extreme given that other is extreme. Tail measures are copula-based and copula is related to the full distribution via quantile transformations, i.e., for all $u_1, u_2 \in (0,1]$,

$$C(u_1, u_2) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right). \tag{3.14}$$

3.4 Copula: Simulation

Simulation has a pivotal role in replicating and analyzing data. Copulas can be applied in simulation and Monte Carlo studies. Johnson [23] discusses methods to generate a sample from a given joint distribution. One such method is a recursive simulation using the univariate conditional distributions. The conditional distribution of U_i given first $i-1$ components is

$$c(u_i|u_1, \ldots, u_{i-1}) = \frac{\partial^{i-1} C(u_1,\ldots,u_i)}{\partial u_1 \ldots \partial u_{i-1}} \Big/ \frac{\partial^{i-1} C(u_1,\ldots,u_{i-1})}{\partial u_1 \ldots \partial u_{i-1}}. \tag{3.15}$$

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

3.5 Gaussian and $t(\nu)$ Copulas

Elliptical copulas are copulas for the elliptical distributions. The most commonly used elliptical distributions are the Gaussian and student - *t* distributions. The key advantage of elliptical copulas is that it is possible to specify different levels of correlation between the marginals. However elliptical copulas do not have closed form expressions and are restricted to have radial symmetry. Gaussian copula is defined by

$$C(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_2)} \int_{-\infty}^{\Phi^{-1}(u_1)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right] dxdy, \tag{3.16}$$

and the student t- copula with (ν) degrees of freedom, i.e., $t(\nu)$ copula is

$$C(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_2)} \int_{-\infty}^{\Phi^{-1}(u_1)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{\nu(1-\rho^2)}\right]^{\frac{\upsilon+2}{2}} dxdy. \tag{3.17}$$

The copula parameter ρ in terms of τ is

$$\rho = \mathrm{Sin}\left(\frac{\pi}{2}\tau\right). \tag{3.18}$$

Gaussian copulas allow any marginal distribution and any positive definite correlation matrix. Gaussian copulas consider only pairwise dependence between individual components of a random variable. However problem may be because correlation matrix can be difficult to estimate for too many parameters. Further Gaussian densities are parameterized using Pearson correlation coefficients which are not invariant under monotone transformations of original variables.

3.6 Archimedean Family of Copulas

Archimedean copulas are an important class of copulas which are easier to construct [43]. They possess nice mathematical properties and many known copula families belong to this class. Let φ be a continuous decreasing function from $[0,1]$ to $[0,\infty)$ such that $\varphi(1) = 0$ and φ^{-1} be its inverse given by

$$\varphi^{-1}(t) = \begin{cases} \varphi^{-1}(t) & 0 \leq t \leq \varphi(0) \\ 0 & \varphi(0) \leq t \leq \infty \end{cases} \tag{3.19}$$

Then the Archimedean copula is the function

$$C(u_1, u_2) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2)), \; u_1, u_2 \in [0,1]. \tag{3.20}$$

The function φ is called a generator of the copula C. Some examples of Archimedean copulas are given in Table 1. For some applications of copula based analyses in clinical, economic and engineering studies, reference is made to [19, 31-36].

3.6 Illustration: Application of the Ali-Mikhail-Haq (AMH) Copula

We consider a study in which twenty three patients were registered in a split-mouth trial for the treatment of gingivitis [41]. In these trials four sites located either on the left or right side of a patient's mouth were randomly assigned to either the treatment (chlorhexidine) or control (saline). Plaque measurements were taken pre-treatment and two weeks after baseline on four sites of the patient's upper jaw. In this illustration, we consider modeling the post-treatment proportions of sites exhibiting plaque in treatment X_1 and control X_2 groups at a two-week follow-up visit. Post-treatment proportions and summary statistics are presented in Table 2. Estimated Kendall's τ is 0.1761. The marginal distributions estimated from the *q-q* plots are: X_1~Beta (66.88, 8.16) and X_2~Beta (57.91, 17.13). The AMH copula parameter θ is estimated equal to 0. 6481. Figure 1 shows the relationship between AMH copula parameter θ and the Kendall and Spearman rank correlations. In Figure 2, we show the scatter plots of simulated bivariate data using AMH copula for n = 50 and 100. Estimated AMH copula density model from data and conditional probabilities are plotted in Figure 3. Tail dependence behavior using AMH copula is exhibited in Figure 4.

4 Mutual Information Based Measures

Dependence from the information theoretic point of view can be quantified by measuring the distance between a given joint probability density model $f(x_1, \ldots, x_k)$ and a mean field model $\prod_{i=1}^{k} f(x_i)$, where $f(.)$ denotes the density function. Information theory provides a unifying framework for ideas from areas as diverse as differential geometry, physics, statistics and telecommunications [24, 25, 26].

4.1 Entropy and Conditional Entropy

Consider a finite real valued discrete random variable X with its probability distribution $(x_i, p_i, i = 1,..m; \sum_i p_i = 1)$. The measure of uncertainty associated with the variable X is called entropy and defined as

$$H(X) = -c \sum_i p_i \log p_i, \tag{4.1}$$

where c is an arbitrary constant. Constant c is generally taken as unity and logarithm base 2 when entropy is measured in bits. The uncertainty takes the maximum value when all probabilities are equal, i.e., $p_i = 1/m$. Thus, the bounds for $H(X)$ are: $0 \leq H(X) \leq \log m$. Zero entropy implies that the process of generating X is deterministic. Closer is the value of $H(X)$ to 0, lesser is the uncertainty of X while the value of $H(X)$ being closer to $\log m$ means greater uncertainty. $H(X)$ is a monotonic increasing function of m.

For the simplicity in notations, we will denote two random variables by X and Y with respective 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 joint entropy of X and Y is then defined by

$$H(X,Y) = -\sum_{ij} p_{ij} \log p_{ij}. \tag{4.2}$$

When X and Y are independent, $p_{ij} = p_i\, q_j, \forall\, i, j$, the entropy of the joint distribution equals the sum of respective entropies of X and Y, i.e., $H(X,Y) = H(X) + H(Y)$. However when they are not independent, question is: How much uncertainty of X diminishes if we know that $Y \in C_j$. For more

properties of entropy, we refer to [25,26,38]. For general considerations, stochastic dependence of random variables X and Y results in reducing their joint entropy. In such a case, it is relevant to introduce the conditional entropy $H(X|y_j)$ which represents the amount of uncertainty of X given that y_j is observed; $H(X|y_j) = -\sum_i p_{i/j} \log p_{i/j}$ 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 conditional entropy $H(X|Y)$ is the amount of uncertainty of X remaining given advance knowledge of Y and is obtained by averaging $H(X \mid y_j)$ over all j and equals to

$$H(X|Y) = -\sum_{ij} p_{ij} \log p_{i/j}. \tag{4.3}$$

Similarly the conditional entropy $H(Y|X)$ is defined. Conditional entropy is nonnegative and nonsymmetric. It is easily verified that $H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$ and, therefore $H(X,Y) \leq H(X)$ or $H(Y)$ with equality holding if and only if X and Y are stochastically independent.

4.2 Mutual Information

For a better understanding, if we assume X and Y are input and output respectively of a stochastic system, then $H(X)$ represents the uncertainty of input X before output Y is observed while $H(X|Y)$ is the uncertainty of input X after output Y has been realized. 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 (distance from independence) between X and Y. An interesting alternative for characterizing dependence is the expression of mutual information in terms of the Kullback-Liebler distance between joint distribution and the two marginal distributions [30] defined by

$$I(X,Y) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{p_i q_j}, \tag{4.4}$$

where the Kullback-Liebler distance between two probability distributions p and q is $K(p||q) = \sum_i p_i \log (p_i/q_i)$. Mutual information can also be expressed in terms of divergence between conditional distribution and marginal distributions by

$$I(X,Y) = \sum_j q_j \sum_i p_{i/j} \log(p_{i/j}/p_i). \tag{4.5}$$

Mutual information thus measures the decrease in uncertainty of X caused by the knowledge of Y which is the same as the decrease in uncertainty of Y caused by the knowledge of X. The measure $I(X,Y)$ indicates the amount of information of X contained in Y or the amount of information of Y contained in X. Obviously $I(X,X) = H(X)$, the amount of information contained in X about itself.

To transmit X, how many bits on average would it save if both ends of the line knew Y? Information gain answers this question and is defined as

$$IG(X|Y) = H(X) - H(X|Y). \tag{4.6}$$

It is seen that $I(X,Y) = I(Y,X)$ and $I(X,Y) \geq 0$ with equality when X and Y are stochastically independent and $I(X,Y) \leq H(X)$. The relative information gain is [28]:

$$r(Y|X) = \frac{I(X,Y)}{H(X)} = I(X,Y)/[H(X,Y) - H(Y|X)], \tag{4.7}$$

which shows how much uncertainty of X diminishes given information about Y relative to the initial uncertainty of X. Other properties of the relative information gain $r(X|Y)$: $0 \leq r(X|Y) \leq 1$ and $r(X|Y)$ assume zero value if and only if X and Y are stochastically independent. In case where there is no information about which random variable influences the other or which takes values first, then a symmetrical relative information gain measure

$$R(X,Y) = 2I(X,Y)/[H(X) + H(Y)] = 2I(X,Y)/[H(X,Y) + I(X,Y), \tag{4.8}$$

expresses the uncertainty from the joint distribution of X and Y to the uncertainty in case of independence. This measure $R(X|Y)$ has the properties: $0 \leq R(X,Y) \leq 1$ and $R(X,Y) = 0$ if and only if X and Y are stochastically independent. The measures $r(X|Y)$ and $R(X,Y)$ can be used to characterize the stochastic dependence between X and Y. They are also useful in characterizing the dependence of qualitative

variables under the hypothesis that the values of the qualitative variable cover all possibilities and their common part is empty.

4.3 Copula Based Information -Theoretic Measures

The joint entropy $H(X,Y)$ associated with the joint distribution of X and Y using copula density function $c(u_1,u_2)$ from (3.3) can be expressed

$$H(X,Y) = -\sum_{ij} c(u_1,u_2)\log\, c(u_1,u_2). \tag{4.9}$$

The conditional entropy $H(X|Y)$ expressed in terms of conditional copula density function $c(u_1|u_2)$ from (3.17) is

$$H(X|Y) = -\sum_{ij} c(u_1,u_2)\log\, c(u_1|u_2). \tag{4.10}$$

The mutual information (distance from independence) $I(X,Y)$ between X and Y using copula functions is expressed by

$$I(X,Y) = -\sum_{ij} c(u_1,u_2)\log[c(u_1,u_2)/\{c(u_1|u_2)\times c(u_2|u_1)\}]. \tag{4.11}$$

The relative information gain $r(X|Y)$ in terms of copula functions

$$r(X|Y) = \frac{\sum_{ij} c(u_1,u_2)\log[c(u_1,u_2)/\{c(u_1|u_2)\times c(u_2|u_1)\}]}{\sum_{ij} c(u_1,u_2)\log[c(u_1,u_2)/\{c(u_2|u_1)\}]}, \tag{4.12}$$

and the symmetrical relative information gain measure

$$R(X,Y) = \frac{2\sum_{ij} c(u_1,u_2)\log\left[\dfrac{c(u_1,u_2)}{\{c(u_1|u_2)\times c(u_2|u_1)\}}\right]}{\sum_{ij} c(u_1,u_2)\log\left[\dfrac{\{c(u_1,u_2)\}^2}{\{c(u_1|u_2)\times c(u_2|u_1)\}}\right]}. \tag{4.13}$$

Evaluation of these expressions become cumbersome depending upon the copula functions and the marginal probability distributions. An alternative computational method [28] is by expressing probabilities of a pair (x_i,y_j) belonging to the rectangle $R_i: [x_{i-1}^*; x_i^*]\times C_j: [y_{j-1}^*; y_j^*]$ in terms of associated copula function $C(u_1,u_2)$ as

$$p_{ij} = \int_{x_{i-1}^*}^{x_i^*}\int_{y_{j-1}^*}^{y_j^*} f(x,y)dxdy = \int_{u_{1i-1}^*}^{u_{1i}^*}\int_{u_{2j-1}^*}^{u_{2j}^*} c(u_1,u_2)du_2du_1, \tag{4.14}$$

$$p_{i/j} = \int_{u_{1i-1}^*}^{u_{1i}^*}\int_{u_{2j-1}^*}^{u_{2j}^*} c(u_1,u_2)du_2du_1/(u_{2j}-u_{2j-1}), \tag{4.15}$$

where $u_{1i}^* = F_1(x_i^*)$ and $u_{2j}^* = F_2(y_j^*)$.
The integrals appearing in (4.14) and (4.15) can be expressed in terms of copula

$$\int_{u_{1i-1}^*}^{u_{1i}^*}\int_{u_{2j-1}^*}^{u_{2j}^*} \frac{\partial^2 C(u_1,u_2)}{\partial u_1\ \partial\, u_2}du_1du_2 = C(u_{1i}^*,u_{2j}^*) - C(u_{1i-1}^*,u_{2j}^*) - C(u_{1i}^*,u_{2j-1}^*) + C(u_{1i-1}^*,u_{2j-1}^*). \tag{4.16}$$

It is easy to calculate information measures by using (4.16) because $C(u_1,u_2)$ is evaluated at the points of partition only.

4.4 Mutual Information using Marshall-Olkin Copula

One parameter Marshall-Olkin copula [37] is defined by

$$C(u_1,u_2) = \min\left(u_1^{1-\theta}\,u_2, u_1u_2^{1-\theta}\right), \quad u_1,u_2,\theta \in (0,1]. \tag{4.17}$$

The copula density function $c(u,v)$ is

$$c(u_1,u_2) = \begin{cases} (1-\theta){u_1}^{-\theta}, & \text{if } u_1 > u_2, \\ (1-\theta){u_2}^{-\theta}, & \text{if } u_1 < u_2, \\ 0, & \text{if } u_1 = u_2. \end{cases} \quad (4.18)$$

The copula parameter θ in terms of Kendall's τ has a simple expression

$$\theta = \frac{2\tau}{1+\tau}. \quad (4.19)$$

The mutual information $I(X,Y)$ is the entropy of the copula $C(u,v)$ itself whatever the marginal distributions may be [39]. Using one parameter Marshall-Olkin copula $C(u,v)$,

$$I(X,Y) = -2\frac{1-\theta}{2-\theta}\left[\log(1-\theta) + \frac{\theta}{2-\theta}\right], \quad (4.20)$$

or in terms of Kendall's τ

$$I(X,Y) = -(1-\tau)\left[\tau + \log\left(\frac{1-\tau}{1+\tau}\right)\right]. \quad (4.21)$$

In Figure 5, we depict the behaviour of the mutual information $I(X,Y)$ versus the copula parameter $\theta \in (0,1]$. This parametrization of mutual information based on one parameter Marshall-Olkin copula is much more accurate than based on the correlation parameter while keeping the same level of computational complexity.

4.5 Illustration

We consider two examples to illustrate applications of the information based measures. These examples represent situations which refer to the univariate and bivariate distributions.

Example 1. Benford's Law is a powerful and relatively simple tool for pointing suspicion at frauds, embezzlers, tax evaders, sloppy accountants and even computer bugs. The income tax and accounting agencies often use detection software based on Benford's Law. Dr. Frank Benford, a physicist at the General Electric Company, noticed that pages of logarithms corresponding to numbers starting with the numeral 1 were much dirtier and more worn than other pages. He thought that it was unlikely that physicists and engineers had some special preference for logarithms starting with 1. He therefore embarked on a mathematical analysis of 20,229 sets of numbers from different applications. All these seemingly unrelated sets of numbers followed the same first-digit probability pattern as the worn pages of logarithm tables suggested. In all cases, the number 1 turned up as the first digit about 30 percent of the time, more often than any other. He derived a formula to explain this phenomenon. If absolute certainty is considered as 1 and absolute impossibility as 0, then the leading digit $d \in [1, b-1]$ in base b occurs with probability $P(d) = \log_b\left[1 + \frac{1}{d}\right]$. This quantity is exactly the space between d and $d+1$ in a logarithmic scale. Probability distribution is given in Table 3. The entropy as a measure of equality of digits 1 through 9 is $H(d) = -\sum_i p_i \log_{10} p_i = 0.87$ dits/digit and maximum entropy $H_{\max}(d) = \log_{10} 9 = 0.95$ dits/digit. Thus, the uncertainty in the distribution of digits in the table is less than the maximum possible uncertainty. This reduction in uncertainty is due to the information available that all digits in the table do not occur in the same proportion.

Example 2. A mobile ad hoc computer network consists of several computers (nodes) that move within a network area. When the recieving node is out of range, message must be sent to a nearby node which then forwards it along a routing path towards its destination. Data overhead is the number of bytes of information that must be transmitted along with the messages to get them to the right places. A successful protocol will generally have a low data overhead. Data on average node speed (Speed), length of time that

nodes pause at each destination (Pause Time), link change rate (LCR) and data overhead (Overhead) for the 25 simulated mobile ad hoc networks is taken from [2] and summary statistics are given in Table 4.

From the summary statistics in Table 4, dependence measures using Kendall's τ between Overhead and Speed is 0.0.467, Overhead and Pause Time is 0.577 and Overhead and LCR is -0.239. Correlations of Overhead with Speed and Pause Time both are positive and highly significant however with LCR, it is negative and not significant. Therefore, Marshall -Olkin copula being contrained for positive correlations can not be applied to measure dependence using LCR. We calcualted Marshall-Olkin copula parameter θ for Overhead and Speed as 0.6367 and for Overhead and Pause Time as 0.7318.

Since Marshall-Olkin copula parameter $\theta \in (0,1]$, $\theta = 0.6367$ and 0.7318 indicate a higher degree of dependence. Uncertainty in this example is bounded between 0 and 3.2189. Mutual information using natural log for Overhead and Speed is 0.2907 and for Overhead and Pause Time is 0.3126. Thus, uncertainty in Overhaead caused by the knowledge of Speed is higher compared to Pause Time. Alternatively we can say that the amount of information of Overhead contained in Pause Time is more than the information of Overhead contained in Speed. Pause Time and Speed are important variables in modeling the dependence of Overhead.

5 Conclusions

Pearson's linear correlation based statistical methods have dominated statistical modeling and inference literature until recent. However researchers have now realized problems with uses of correlation and accepted the fact that such methods are useful only when considering multivariate *normal* populations. Multivariate *normal* distributions are appealing because their marginal distributions are also *normal* and the association between any two random variables can be fully described knowing only the marginal distributions and the dependence parameter measured by the Pearson's linear correlation coefficient. There are often situations in the *non-normal* world wherein *normal* distributions fail to provide an adequate approximation. Therefore dependence metrics like information measures and copulas which seem to be appropriate alternative to the correlation need special considerations and investigations in the context of statistical inference. Copula functions and copula parameters are applied to model the dependence and simulate multivariate populations. There exist several families of copulas from which the best copula can be selected for a particular application. Mutual information in terms of Kullback-Leibler divergence is often studied however there are other several generalized divergence measures which may be investigated. Mutual information is expressible in terms of copula functions and thus copulas can play an important role in analyzing mutual information. Marginal distributions and copula of a multivariate distribution are inextricably linked. Copula separates the dependence from the marginal distributions. Various families of copulas like Archimedean, Gaussian, $t(\nu)$, elliptical, extreme value are available and may be preferred because of mathematical tractability. Copulas are considered as an alternative to Gaussian models in a non-Gaussian world. There is almost no or very little statistical theory, estimation and significance testing, developed based on copulas. Sensitivity studies of estimation procedures and goodness-of-tests for copulas are unknown. It is unclear whether a good fit of the copula to the data yields a good fit to the distribution of the population data.

Table 1. Archimedean copulas, Generator Functions and Kendall's τ.

Copula	Generator $\varphi(t)$	$C(u,v)$	Kendall's τ
Clayton	$(t^{-\theta}-1)/\theta, \theta \in [-1,\infty)\setminus\{0\}$	$(u_1^{-\theta}+u_2^{-\theta}-1)^{-1/\theta}$	$\theta/(\theta+2)$
Gumbel	$(-\ln t)^{\theta}, \theta \in [1,\infty)$	$\exp[-\{(-\ln u_1)^{\theta}+(-\ln u_2)^{\theta}\}^{1/\theta}]$	$(\theta-1)/\theta$
Frank	$-\ln[(e^{-t\theta}-1)/(e^{-\theta}-1)]$, $\theta \in (-\infty,\infty)\setminus\{0\}$	$-\frac{1}{\theta}\ln[1+\frac{(e^{-u_1\theta}-1)(e^{-u_2\theta}-1)}{e^{-\theta}-1}]$	$\frac{1-4\{1-D_1(\theta)\}}{\theta}$

Ali-Mikhail-Haq	$\ln\left[\frac{1-\theta(1-t)}{t}\right], \theta \in [-1,1)$	$\frac{u_1 u_2}{1-\theta(1-u_1)(1-u_2)}$	$\frac{3\theta-2}{3\theta} - \frac{2(1-\theta)^2\ln(1-\theta)}{3\theta^2}$
Frank-Joe	$-\ln[1-(1-t)^\theta], \theta \in [1,\infty)$	$1-[(1-u_1)^\theta + (1-u_2)^\theta - (1-u_1)^\theta(1-u_2)^\theta]$	No closed form

* $D_k(x)$ is the Debye function for any positive integer k : $D_k(x) = \frac{k}{x^k}\int_0^x \frac{t^k}{e^t-1}dt$.

Table 2. Post-treatment proportions of sites exhibiting plaque in treatment and control groups ($n = 23$).

	Treatment	Control		
Mean	0.8913	0.7717	Pearson correlation	0.1351
Standard deviation	0.1656	0.2373	Kendall's τ	0.1761
Skewness	-1.2882	-0.5346	Spearman's ρ_s	0.2604
Marginal distribution	Beta(66.88,8.16)	Beta(57.91,17.13)	AMH copula θ	0.6481

Table 3. Probability distribution of digit $d \in [1, b-1]$ in base $b = 10$.

d:	1	2	3	4	5	6	7	8	9
p:	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Table 4. Average node speed, pause time that nodes pause at each destination, link change rate (LCR) and data overhead for simulated mobile ad hoc networks ($n = 25$).

	Overhead (kB)	Speed (m/s)	Pause Time (s)	LCR (100/s)	Overhead vs.	Speed	Pause Time	LCR (100/s)
Mean	481.773	21	30	15.227	Correlation	0.526*	0.738*	-0.239
Standard deviation	28.957	13.070	14.434	8.088	Kendall's τ	0.467*	0.577*	-0.040
Skewness	-1.840	0.219	0	1.401	Spearman's ρ_s	0.565*	0.722*	-0.007

* 1% significance level.

Figure 1. Rank correlations and AMH Copula parameter.

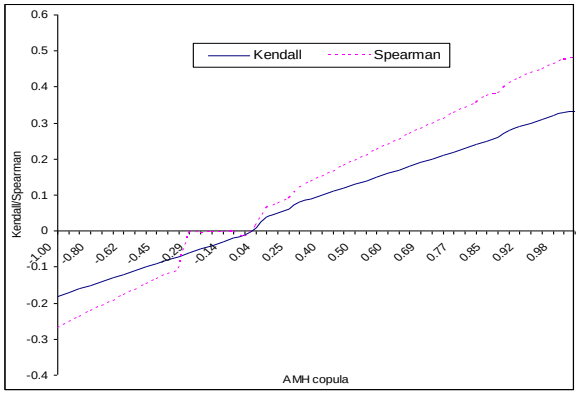


Figure 2. Scatter plots of AMH copula simulated data for n = 50 and 100.

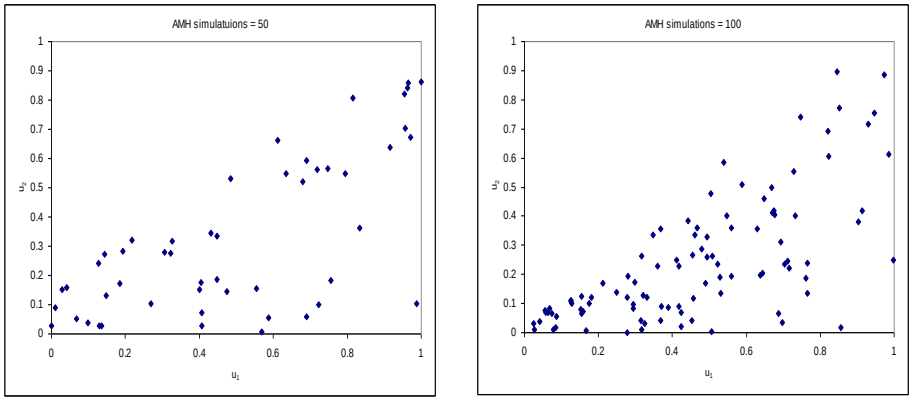


Figure 3. Joint probability model and conditional probability model.

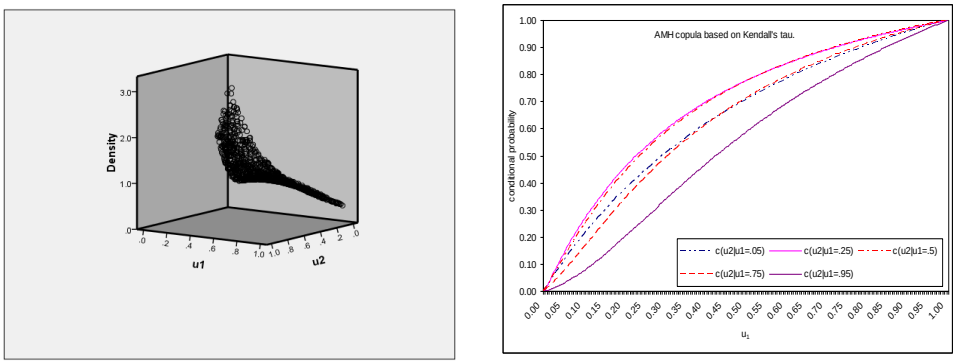


Figure 4. Tail dependence indices for AMH copula parameter θ = -1, -0.5, 0, 0.5, 1.

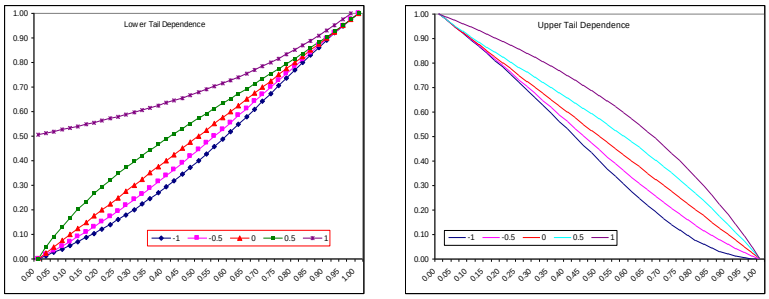
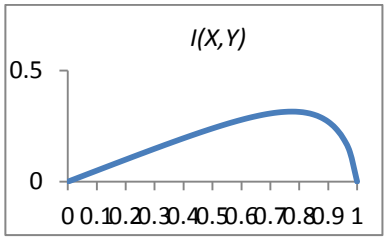


Figure 5. Mutual information and dependence parameter.



Acknowledgments

This work was supported by author's Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] Akaike, H. Information theory and an extension of the maximum likelihood principle. *Proc. Second Int. Symp. Information Theory* (1972), 267-281.

[2] Boleng, J., Navidi, W. and Camp, T. *Proceedings of the International Conference on Wireless Networks* (2002), 293-298.

[3] Calsaverini, R.S. and Vicente, R. An information theoretic approach to statistical dependence: Copula information. *Europhysics Letters* (2009), 88, 68003, doi:10.1209/0295-5075/88/68003.

[4] Clayton, D.G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65 (1978),141-151.

[5] Cuadras, C.M., Fortiana, J. and Rodrguez Lallena, J.A. *Distributions with Given Marginals and Statistical Modelling* (2002). Dodrechts: Kluwer Academic Publishers.

[6] Embrechts, P., McNeil, A. and Straumann, D. Correlation and dependence in risk management: Properties and Pitfalls. *Risk*,12,5 (1997), 69-71.

[7] Embrechts, P., McNeil, A. and Straumann, D. Correlation and dependence in risk management: properties and pitfalls. *Risk Management: Value at Risk and Beyond*, ed. M.A.H. Dempster, Cambridge University Press, Cambridge (2002), 176-223.

[8] Fang, K.-T., Kotz, S. and Ng, K.-W. *Symmetric Multivariate and Related Distributions* (1987). London: Chapman & Hall.

[9] Frank, M.J. On the simultaneous associativity of $F(x,y)$ and $x+y-F(x,y)$. *Aequationes Mathematicae* 19 (1979) ,194-226.

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

[11] Frees E. W., and E. Valdez. Understanding relationships using copulas. *North American Actuarial Journal*, 2,1 (1998),1-25.

[12] Galton, F. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15 (1886), 246–263.

[13] Genest, C. Franks family of bivariate distributions. *Biometrika*, 74 (1987), 549-555.

[14] Genest, C. and Mackay, J. The joy of copulas: Bivariate distributions with uniform marginals. *American Statistician*, 40 (1986) ,280-283.

[15] Genest, C., and Rivest, L. Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, 88 (1993),1034-1043.

[16] Genest, C., Ghoudi, K. and Rivest, L. A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82 (1995) ,543-552.

[17] Gumbel, E.J. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55 (1960), 698-707.

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

[19] Herath, H. and Kumar, Pranesh. New research directions in engineering economics – modeling dependencies with copulas. *Engineering Economist*, 52:4 (2007), 305-331.

[20] Hougaard, P. A class of multivariate failure time distributions. *Biometrika*,73 (1986),671-678.

[21] Hutchinson, T.P. and Lai, C.D. *Continuous Bivariate Distributions Emphasizing Applications* (1990). Adelaide, South Australia: Rumsby Scientific Publishing.

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

[23] Johnson, M.E. *Multivariate Statistical Simulation* (1987). New York: John Wiley & Sons.

[24] Kapur, J.N. *Maximum Entropy Models in Science and Engineering* (1989) . Wiley Eastern, Delhi.

[25] Kapur, J.N. and Kesavan, H.K. *Entropy Maximization Principles with Applications* (1992). Academic Press.

[26] Karmeshu and Pal, N.R. Uncertainty, entropy and maximum entropy principles- An overview. In *Entropy Measures, Maximum Entropy Principles and Engineering Applications* (2002), Karmeshu (Ed), Springer.

[27] Kimeldorf, G. and Sampson, A. R. Monotone dependence. *Annals of Statistics*, 6 (1978), 895-903.

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

[29] Kruskal, W.H. Ordinal Measures of Association. *Journal of American Statistical Association*, 53 (1958), 284, 814-861.

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

[31] Kumar, Pranesh. Copulas as an alternative dependence measure and copula based simulation with applications to clinical data. *Bulletin Int. Statist. Inst.*, LXII (2007), 2674-2677.

[32] Kumar, Pranesh. Applications of the Farlie-Gumbel-Morgenstern copulas in predicting the properties of the Titanium welds. *International Journal of Mathematics*, 1, 1 (2009),13-22.

[33] Kumar, Pranesh. Copula functions as a tool in statistical modelling and simulation.*Proceedings of the International Conference on Methods and Models in Computer Science* (ICM2CS09). IEEE Xplore (2009).

[34] Kumar, Pranesh and Shoukri, M. M. Evaluating aortic stenosis using the Archimedean copula methodology. *Journal of Data Science*,6 (2008), 173-187.

[35] Kumar, Pranesh and Shoukri, M. M. (2007). Copula Functions for Modelling Dependence Structure with Applications in the Analysis of Clinical Data. *Journal of Indian Soc. Agric. Statist*., 61(2), 179-191.

[36] Kumar, P. (2011). Copulas: Distribution functions and simulation. In Lovric, Miodrag (Ed), *International Encyclopedia of Statistical Science*. Heidelberg: Springer Science+Business Media, LLC.

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

[38] Mathai, A.M. and Rathie, P.N. (1975). *Basic Concepts in Information Theory and Statistics*. John Wiley & Sons.

[39] Mercier, G. (2005). *Measures de Dépendance entre Images RSO*. GET/ENST Bretagne, Tech. report RR-2005003-TI. http//:perso.enst-bretagne.fr/126mercierg.

[40] Montgomery, D.C. (2009). *Design and Analysis of Experiments*. 7th edition, John Wiley.

[41] Morrow, D., Wood, D.P. and Speechley, M. (1992). Clinical effect of subgingival chlorhexidine irrigation on gingivitis in adolescent orthodontic patients. *American Journal of Orthodontics and Dentofacial Orthopedics*, 101, 408-413.

[42] Nelsen, R. (2006). *An Introduction to Copulas*. New York: Springer.

[43] Nelsen, R.B., Quesada Molina, J.J., Rodrguez Lallena, J.A. and Úbeda Flores, M. (2001). Bounds on bivariate distribution functions with given margins and measures of association. *Commun. Statist.-Theory Meth*., 30, 1155-1162.

[44] 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.

[45] Scarsini, M. (1984). On measures of concordance. *Stochastica*,8,201-219.

[46] Schweizer, B. and Sklar, A. (1983). *Probabilistic Metric Spaces*. New York: North Holland.

[47] Schweizer, B. and Wolff, E. (1981). On nonparametric measures of dependence for random variables. *Annals of Statistics*, 9, 879-885.

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

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

[50] Tjøstheim, D. (1996). Measures of dependence and tests of independence. *Statistics*, 28, 249-284.

[51] Yao, Y.Y. (2002). Information-theoretic measures for knowledge discovery. In *Entropy Measures, Maximum Entropy Principles and Engineering Applications*, Karmeshu (Ed), Springer.