7^{th} Seminar on Probability and Stochastic Processes
Isfahan University of Technology
Aug. 13-14, 2009

Information Measures via Copula Functions

G.R. Mohtashami Borzadaran
Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran.
email: gmb1334@yahoo.com

M. Amini
Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran.
email: m-amini@um.ac.ir

Abstract

In applications of differential geometry to problems of parametric inference, the notion of divergence is often used to measure the separation between two parametric densities. Among them, in this paper, we will verify measures such as Kullback-Leibler information, J-divergence, Hellinger distances, α-Divergence,... and so on. Properties and results related to distance between probability distributions derived via copula functions. Some results and inequalities obtained in view of dependence and information measures.

MSC 2000: Primary 94A15, Secondary 60E05.
Keywords: Information measures, Fisher information, Kullback-Leibler Information, Hellinger distance, α-Divergence.

1 Introduction and Preliminaries

The study of copulas and the role they play is important in probability, statistics and stochastic processes. Sklar (1959) found result due to copula. Many research papers and monograph published after 1959 that can be find lots of them in Nelsen (2006) and Mari and Kotz (2006) and their references. Frees and Valdez (1998) introduced the concept of copulas as a tool for understanding relationships among multivariate outcomes. Also, dependence and copulas have linked with each other.

The concept of entropy originated in the nineteenth century as a creation of C. E. Shannon (1948). During the last sixty years or so, a number of research papers, and monographs discussing and extending Shannon's original work have appeared. Among them Ali Ahmed et al (1989), Darbellay and Vajda (1998, 2000), Dragomir (2003), Blyth (1994), Torkkola (2003), Kapur (1989,1994), Borovkov(1998), Kagan, Linnik & Rao (1973) and Kullback (1959) are using and extending in this research.

In this paper, various measures are obtained in view of copulas for bivariate distributions. Properties of information measures and its link with copula is another direction of this research.

2 Preliminaries and Result Related to some Information Measures

Let $(\Omega, \mathcal{B}, \mu)$ be a measure space and f be a measurable function from Ω to $[0, \infty)$, such that $\int_\Omega f d\mu = 1$. The Shannon entropy (or simply the entropy) of f relative to μ, is defined by

$$H(f, \mu) = -\int_\Omega f \ln f d\mu, \text{ (with } f \ln f = 0 \text{ if } f = 0), \tag{1}$$

and assumed to be defined for which $f \ln f$ is integrable. If X is an r.v. with pdf f, then we refer to H as the entropy of X and denotes also it by the notation H_X. In the case μ is a version of counting measure, (1) leads us to a specialized version that introduced by Shannon (1948) as $H_X = -\sum_{i=1}^n p_i \ln p_i$ where $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$. One of the important issues in many applications of probability theory is finding an appropriate measure of distance between two probability distributions. A number of divergence measure for this purpose have been studied by Kullback and Leibler (1951), Renyi (1961) and a lot of references related to various type of information measures can be find in Dragomir (2003).

Assume that the set χ be the suitable support of distributions and the $\sigma-$finite measure μ are given such that $\Omega = \{f : f \to \Re, f(x, y) > 0, \int_\chi f(x, y)d\mu(x) = 1\}$. Consider F and G be two bivariate distributions which are absolutely continuous w.r.t. measure μ and $\frac{dF}{d\mu} = f$ and $\frac{dG}{d\mu} = g$ and G is a distribution function. Here, we introduced shortly the form of some familiar information measures based on bivariate distributions.

Kullback Leibler Information :

$$D_{KL}(F, G) = \int\int_{\chi\times\chi} \ln \frac{f(x, y)}{g(x, y)} f(x, y)d\mu, \tag{2}$$

χ^2- **Divergence :**

$$D_{\chi^2}(F, G) = \int\int_{\chi\times\chi} \frac{[f(x, y) - g(x, y)]^2}{f(x, y)} d\mu, \tag{3}$$

Hellinger Distance:

$$D_H(F, G) = \int\int_{\chi\times\chi} [\sqrt{f(x, y)} - \sqrt{g(x, y)}]^2 d\mu, \tag{4}$$

$\alpha-$**Divergence :**

$$D_\alpha(F, G) = \frac{1}{1 - \alpha^2} \int\int_{\chi\times\chi} \{1 - \frac{g^{\frac{1+\alpha}{2}}(x, y)}{f^{\frac{1+\alpha}{2}}(x, y)}\} f(x, y)d\mu, \tag{5}$$

Jeffery's Distance (J-Divergence):

$$D_J(F, G) = \int\int_{\chi\times\chi} [f(x, y) - g(x, y)] \ln \frac{f(x, y)}{g(x, y)} d\mu, \tag{6}$$

Combination of version of $\alpha-$**Divergence:**

$$D_{C\alpha}(F, G) = \frac{4}{\beta^2} \int\int_{\chi\times\chi} \frac{\{g^{\frac{\beta}{2}}(x, y) - f^{\frac{\beta}{2}}(x, y)\}^2}{g^{\beta-1}(x, y)} d\mu, \tag{7}$$

Bhattacharyya Distance:

$$D_{Bh}(F, G) = \int\int_{\chi\times\chi} \sqrt{g(x, y)f(x, y)} d\mu, \tag{8}$$

Harmonic Distance:

$$D_{Ha}(F, G) = \int\int_{\chi\times\chi} \frac{2g(x, y)f(x, y)}{g(x, y) + f(x, y)} d\mu, \tag{9}$$

Triangular Discrimination:

$$D_{\Delta}(F,G)=\int\int_{\chi\times\chi}\frac{[f(x,y)-g(x,y)]^2}{g(x,y)+f(x,y)}d\mu, \tag{10}$$

Lei and Wang Divergence:

$$D_{LW}(F,G)=\int\int_{\chi\times\chi}f(x,y)\ln\frac{2f(x,y)}{g(x,y)+f(x,y)}d\mu. \tag{11}$$

Relative Information Generating Function :

The relative information generating function of f given the reference measure g is defined by Guiasu and Reischer (1985) as,

$$R(F,G,t)=\int\int_{\chi\times\chi}[\frac{f(x,y)}{g(x,y)}]^{t-1}f(x,y)d\mu, \tag{12}$$

where $t\geq 1$ and the integral is convergent on noting that $R(F,G,1)=1$.

Power Divergence Measures :

Cressie and Read (1984) proposed the power divergence measure (PWD) which gathers most of the interesting specification. This family is indexed by

$$PWD(F,G)=\frac{1}{\lambda(\lambda+1)}\int\int_{\chi\times\chi}\{[\frac{f(x,y)}{g(x,y)}]^{\lambda}-1\}f(x,y)d\mu. \tag{13}$$

The power divergence family implies different well-known divergence measures for different values of λ. PWD for $\lambda=-2,-1,-.5,0,1$, implies Neyman Chi-square, Kullback Leibler, squared Hellinger distance, Likelihood disparity and Pearson Chi-square divergence respectively. Note that $PWD(F,G)=\frac{1}{\lambda(\lambda+1)}[R(F,G,\lambda+1)-1]$.

- It is easy to see that $D_H(F,G)=2[1-D_{Bh}(F,G)]\leq 2$. Via Taylor expansion and approximation, we can get, $D_{KL}(F,G)\approx\frac{1}{2}D_{\chi^2}(F,G)$, $D_J(F,G)\approx\frac{1}{2}[D_{\chi^2}(F,G)+D_{\chi^2}(G,F)]$, $D_{\chi^2}(F,G)\approx 4D_H(F,G)$ and $D_{\chi^2}(F,G)\geq D_H(F,G)$. The $D_{C\alpha}(F,G)$ and $D_{\alpha}(F,G)$ are linked via the following identity : $D_{C\alpha}(F,G)=16(\frac{2}{\beta}-1)A+16(1-\frac{1}{\beta})B$ where A and B are $D_{\alpha}(F,G)$ with $\alpha=1-\beta$ and $\alpha=1-2\beta$ respectively. The chi-squared divergence $D_{\chi^2}(F,G)=D_{C\alpha}(F,G)$ and $D_{\chi^2}(F,G)=2D_{\alpha}(F,G)$ on taking $\beta=2$ in (7) and $\alpha=-3$ in (5) respectively. Also, the Hellinger distance $D_H(F,G)=\frac{1}{4}D_{C\alpha}(F,G)$ and $D_H(F,G)=\frac{1}{2}D_{\alpha}(F,G)$ on taking $\beta=1$ in (7) and $\alpha=0$ in (5) respectively. The Hellinger distance is symmetric and has all properties of metric. Also, we have, $D_{Bh}(F,G), D_{Ha}(F,G), D_{\Delta}(F,G)$ are symmetric and $D_{LW}(F,G)+D_{LW}(G,F)\leq D_{\Delta}(F,G)$.

3 Information measures in view of Copula Distributions

The copula function $C(u,v)$ is a bivariate distribution function with uniform marginal on $[0,1]$, such that

$$F(x,y)=C_F(F_1(x),F_2(y)).$$

By Sklar's Theorem (Sklar, 1959), this copula exists and is unique if F_1 and F_2 are marginal continuous distribution functions. Thus we can construct bivariate distributions $F(x,y)=C_F(F_1(x),F_2(y))$ with given univariate marginal F_1 and F_2 by using copula C_F,(Nelsen, 2006). Then we have the following properties:

- (Nelsen, 2006) Let $F(x,y)$ be a joint distribution function with marginal $F_1(x)$ and $F_2(y)$, then (i) The copula C_F is given by

  $$C_F(u,v)=F(F_1^{-1}(u),F_2^{-1}(v)),\quad \forall u,v\in[0,1],$$

  where, F_1^{-1} and F_2^{-1} are quasi-inverses of F_1 and F_2 respectively.

  (ii) The partial derivatives $\frac{\partial C_F(u,v)}{\partial u}$ and $\frac{\partial C_F(u,v)}{\partial v}$ exist and $c(u,v)=\frac{\partial^2 C_F(u,v)}{\partial u\partial v}$ is density function

of $C_F(u,v)$.
Ma and Sun (2008) defined copula entropy as follows :
Definition Let X be a random vector with copula density $c(u)$. Copula entropy of X is

$$H_c(X) = -\int_u c(u)\ln c(u)du.$$

Kullback Leibler Information :

$$D_{KL}(F, F_1F_2) = \int_0^1\int_0^1 c(u,v)\ln c(u,v)dudv. \tag{14}$$

In this case, Kullback Leibler information is called mutual information.

Theorem 1. *Mutual information of the random variable is equal to the negative entropy of their corresponding copula function,*

$$D_{KL}(F, F_1F_2) = -H_c(X).$$

Proof : Via $f(x,y) = c(F_1(x), F_2(y))f_1(x)f_2(y)$ easily derived.

- On noting Theorem 1, difference of the information contained in joint distribution and marginal densities is equal to copula entropy. Hence,

  $$H(x) = \sum_i H(x_i) + H_c(x).$$

  Independency implies $H_c(x) = 0$.

- If we assume,

  $$\delta^* = [1 - \exp\{-2D_{KL}(F, F_1F_2)\}]^{\frac{1}{2}},$$

  when the dependence is maximal, D_{KL} tends to infinity and δ^* can be consider as a measure of dependence. As an example, let $X \sim N(\mu, \Sigma)$ where $\mu = [\mu_1, \mu_2]$ and $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ then, $\delta^* =\mid \rho \mid$ that is a suitable measure for finding correlation in this case.

- Let X and Y be identically distributed but not necessarily independent, then $\rho = \frac{D_{KL}(F,F_1F_2)}{H(X)}$ where $0 \leq \rho \leq 1$ and $\rho = 0$ implies independence. So $\rho = 1$ implies X and Y perfectly correlated.

χ^2- **Divergence :**

$$D_{\chi^2}(F, F_1F_2) = \int_0^1\int_0^1 \frac{[c(u,v)-1]^2}{c(u,v)}dudv. \tag{15}$$

Hellinger Distance:

$$D_H(F, F_1F_2) = \int_0^1\int_0^1 [\sqrt{c(u,v)} - 1]^2 dudv. \tag{16}$$

$\alpha-$**Divergence :**

$$D_\alpha(F, F_1F_2) = \frac{1}{1-\alpha^2}\int_0^1\int_0^1 [1 - (c(u,v))^{-\frac{(\alpha+1)}{2}}]c(u,v)dudv. \tag{17}$$

Jeffery's Distance (J-Divergence):

$$D_J(F, F_1F_2) = \int_0^1\int_0^1 [c(u,v) - 1]\ln c(u,v)dudv. \tag{18}$$

Combination of version of $\alpha-$Divergence:

$$D_{C\alpha}(F, F_1F_2) = \frac{4}{\beta^2}\int_0^1\int_0^1 [1 - (c(u,v))^{\frac{\beta}{2}}]^2 dudv. \tag{19}$$

Bhattacharyya Distance:

$$D_{Bh}(F, F_1F_2) = \int_0^1 \int_0^1 \sqrt{c(u,v)}dudv. \tag{20}$$

Harmonic Distance:

$$D_{Ha}(F, F_1F_2) = \int_0^1 \int_0^1 [\frac{2c(u,v)}{c(u,v)+1}]dudv. \tag{21}$$

Triangular Discrimination:

$$D_{\Delta}(F, F_1F_2) = \int_0^1 \int_0^1 [\frac{(c(u,v)-1)^2}{1+c(u,v)}]dudv. \tag{22}$$

Lei and Wang Divergence:

$$D_{LW}(F, F_1F_2) = \int_0^1 \int_0^1 c(u,v)\ln[\frac{(2c(u,v)}{1+c(u,v)}]dudv. \tag{23}$$

Relative Information Generating Function :
The relative information generating function of f given the reference measure g is defined by Guiasu and Reischer (1985) as,

$$R(F, F_1F_2, t) = \int_0^1 \int_0^1 [c(u,v)]^t dudv. \tag{24}$$

where $t \geq 1$ and the integral is convergent on noting that $R(F, G, 1) = 1$.
Power Divergence Measures :
Cressie and Read (1984) proposed the power divergence measure (PWD) which gathers most of the interesting specification. This family is indexed by

$$PWD(F, F_1F_2) = \frac{1}{\lambda(\lambda+1)} \int_0^1 \int_0^1 [(c(u,v))^{\lambda} - 1]c(u,v)dudv. \tag{25}$$

The power divergence family implies different well-known divergence measures for different values of λ. PWD for $\lambda = -2, -1, -.5, 0, 1$, implies Neyman Chi-square, Kullback Leibler, squared Hellinger distance, Likelihood disparity and Pearson Chi-square divergence respectively. Note that $PWD(F, G) = \frac{1}{\lambda(\lambda+1)}[R(F, G, \lambda+1) - 1]$.

4 Inequalities of Information Measures for weakly negative Dependence

Ranjbar et. al.(2008) presented a new definition of dependence which is discussed in this section.
Definition The random variables X and Y are said Weakly Negatively Dependent (WND) if there exist a $\gamma > 1$ such that, $f(x_1, x_2) \leq \gamma.f_1(x_1).f_2(x_2)$ where $f(x_1, x_2)$, $f_1(x_1)$ and $f_2(x_2)$ are joint density and marginal densities of X and Y, respectively.

The class of WND random variables is well defined and a large class of these random variables can be find. The following examples are evidence of WND random variables:
Example (i) Suppose that X_1 and X_2 have half-normal distribution, then

$$f_{X_1,X_2}(x_1, x_2) = \frac{2}{\pi\sqrt{1-\rho^2}} exp\left[-\frac{1}{2(1-\rho^2)}\{x_1^2 + x_2^2 - 2\rho x_1 x_2\}\right]; x_1, x_2 > 0,$$

$$f_{X_i}(x_i) = \sqrt{\frac{1}{\pi}} exp\{-\frac{1}{2}x_i^2\}; i = 1, 2.$$

If $-1 < \rho \leq 0$, then X_1 and X_2 are Negative Quadrant Dependence (NQD) random variables. Moreover,

$$\frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1(x_1)}f_{X_2(x_2)}} = \frac{1}{\sqrt{1-\rho^2}}\exp\left[\frac{-\rho^2}{2(1-\rho^2)}(x_1^2+x_2^2)+\frac{\rho}{1-\rho^2}x_1x_2\right] \leq \frac{1}{\sqrt{1-\rho^2}}.$$

Then $f(x_1,x_2) \leq \gamma.f_1(x_1).f_2(x_2)$, where $\gamma = 1/\sqrt{1-\rho^2} \geq 1$. So, X_1 and X_2 are WND.

(ii) Let X and Y be two random variables with joint Farlie-Gumbel-Morgenstern (FGM) distribution, we have

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)\left[1+\alpha(1-2F_X(x))(1-2F_Y(y))\right].$$

On the other hand, it's obvious that

$$|1+\alpha(1-2F_X(x))(1-2F_Y(y)| \leq 1+|\alpha|,$$

and

$$f_{X,Y}(x,y) \leq [1+|\alpha|]f_X(x)f_Y(y).$$

Therefore, the random variables X and Y are WND with $\gamma = 1+|\alpha| \geq 1$. In addition, we know if $-1 < \alpha \leq 0$, then X and Y are Negative Quadrant Dependence(NQD)
So, we have the following inequalities for the information measures :

$$D_{KL}(F,F_1F_2) \leq \ln\gamma,$$

$$D_H(F,F_1F_2) \leq (\sqrt{\gamma}-1)^2, D_\alpha(F,F_1F_2) \leq \frac{1}{1-\alpha^2}[1-\gamma^{-\frac{\alpha+1}{2}}],$$

$$D_J(F,F_1F_2) \leq (\gamma-1)\ln\gamma, D_{C\alpha}(F,F_1F_2) \leq \frac{4}{\beta^2}[1-\gamma^{\frac{\beta}{2}}]^2,$$

$$D_{Bh}(F,F_1F_2) \leq \sqrt{\gamma}, D_{Ha}(F,F_1F_2) \leq \frac{2\gamma}{1+\gamma},$$

$$D_\Delta(F,F_1F_2) \leq (\gamma-1)^2, D_{LW}(F,F_1F_2) \leq \gamma\ln[\frac{2\gamma}{1+\gamma}],$$

$$R(F,F_1F_2,t) \leq \gamma^{t-1}, PWD(F,G) \leq \frac{1}{\lambda(\lambda+1)}\{\gamma^\lambda-1\}.$$

- The power divergence family implies different well-known divergence measures for different values of λ. PWD for $\lambda = -2,-1,-.5,0,1$, implies Neyman Chi-square, Kullback Leibler, squared Hellinger distance, Likelihood disparity and Pearson Chi-square divergence respectively. Note that $PWD(F,F_1F_2) = \frac{1}{\lambda(\lambda+1)}[R(F,G,\lambda+1)-1]$.

- $\gamma = 1$ implies that this two random variables are independent.

5 Conclusion

The mutual information is actually negative copula entropy. We derived forms of some information measures based on copula functions. Also, result and characterization obtained via various information measures on using copula. For various bivariate distributions finding characterizations and results with novelty is the direction of continuing our research.

References

[1] Ali Ahmed, N. & Gokhale, D. V. *Entropy expressions and their estimators for multivariate distributions.* IEEE Tran. Inf. Theory, **Vol. 35, No. 3**, 688-692.1989.

[2] Blyth, S. *Local divergence and association.* Biometrika, **81**, 579-84.1994.

[3] Borovkov, A. A. *Mathematical Statistics.* Gordon and Breach Science Publishers.1998.

[4] Cressie, N. & Read, T. R. C. *Multinomial goodness-of-fit tests.* J. Roy. Statist. Soc., **Ser. B 46**, 440-464. 1984.

[5] Darbellay, G. A. & Vajda, I. *Estimation of the mutual information with data-dependent partitions.* Tech. Report, , **No. 1921**, Inst. of Inf. Theory and Automation Academy of Sciences of the Czech Republic.1998.

[6] Darbellay, G. A. & Vajda, I. *Entropy expressions for multivariate continuous distributions.* IEEE Inf. Theory , **Vol. 46, No. 2**, 709-712.2000.

[7] Dragomir, S. S. *On the $p-$ logarithmic and $\alpha-$ power divergence measures in information theory.* arXiv:math.PR/0304240 v1.2003.

[8] Frees, E. W. *Understanding relationship using copulas.* Presented at the 23nd Actuarial Research Conference, **6-8 August,** 1997.

[9] Guiasu, S. & Reischer, C. *The relative information generating function.* Information Sciences, **35**, 235-241. 1985.

[10] Kagan, A. M., Linnik, Yu. V. & Rao, C. R. *Characterization Problems in Mathematical Statistics.* Wiley, New York.1973.

[11] Kapur, J. N. *Maximum Entropy Models in Sciences and Engineering.* New York : John Wiley.1989.

[12] Kapur, J. N. *Measure of Information and their Applications.* New York : John Wiley.1994.

[13] Kullback, S. & Leibler, R. A. *On information and sufficiency. Ann. Math. Statist.*, **22**, 79-86.1951.

[14] Kullback, S. *Information and Statistics.* Wiley, New York.1959.

[15] Mari, D.D. and Kotz, S. *Correlation and dependence.* Imperical College Press.2001.

[16] Nelsen, R.B. *An Introduction to Copulas.* Springer, New York.2006.

[17] Ranjbar, V., Amini,M. and Bozorgnia, A.*Asymptotic behavior of weighted sums of dependent random variables with heavy tailed distributions.* J.Science, Islamic Republic of Iran. 19(4)(2008), 357-363.

[18] Renyi, A.*On measures of entropy and information.* Proc. 4^{th} Berkeley Symposium. Statist. Probability. **1**, 547-561.1961.

[19] Shannon, C. E. *A mathematical theory of communication.* Bell System Technical Journal, **27**, 379-423.1948.

[20] Sklar, A. *Fonctions de répartition à n dimensions et leurs marges.* Publ. Inst. Statist. Univ. Paris. **8,** 229–231.1959.

[21] Torkkola, K.*Feature extraction by Non-parametric mutual information maximization.* J. of Machine Learning Research, **3**, 1415-1438.2003.