# On Copula Efficiency

Emil Stoica

# On Copula Efficiency

Emil Stoica

Dragan European University of Lugoj

**Abstract**

The connection between efficiency and copula is discussed by showing that a copula can be employed to decompose the efficiency content of a multivariate distribution into marginal and dependence components. The idea of association measures is used to show that empirical linear correlation underestimates the amplitude of the actual correlation in the case of non-Gaussian marginals. The mutual efficiency is shown to provide an upper bound for the asymptotic empirical log-likelihood of a copula.

## 1 Introduction

From the efficiency theoretic point of view dependence can be quantified by measuring the distance between a given model defined by a joint probability density $\phi(\mathbf{x})$ and a mean field model defined by $\phi_0 = \prod_{j=1}^{N} f_j(x_j)$, where $f_j(x_j)$ are marginal densities $f_j(x_j) = \int \prod_{k \neq j} dx_k \; \phi(\mathbf{x})$. The relative entropy given by

$$S\left[\phi \mid\mid \phi_0\right] = \int \prod_{j=1}^{N} dx_j \, \phi(\mathbf{x}) \log\left(\frac{\phi(\mathbf{x})}{\prod_{j=1}^{N} f_j(x_j)}\right). \tag{1}$$

defines a premetric in the space of distributions that can be employed to quantify the degree of dependence in a model, this particular measure is also known as the total correlation or, in the bivariate case, as the mutual efficiency.

The copula theory has been proposed in statistics as an approach for modeling general dependences in multivariate data. A theorem due to Sklar assures that, under very general conditions, for any joint cumulative distribution function (cdf) $F(\mathbf{x}) = \prod_{j=1}^{N} \int_{-\infty}^{x_j} dx_j \, \phi(\mathbf{x})$ there is a function $C(\mathbf{u})$ (known as the *copula function*) such that the joint cdf can be written as a function of the marginal cdfs in the form $F(\mathbf{x}) = C[F_1(x_1), \cdots F_N(x_N)]$. The converse is also true: this function couples any set of marginal cdfs to form a multivariate cdf. This provides a convenient picture of the marginals as being responsible for the idiossincratic properties of each variable and the copula function as a description of the dependence between them [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28].

A complete articulation of these two concepts is, however, curiously absent in the literature. In this paper we seek to survey the basic ideas connecting these two threads emphasizing the efficiency theoretic interpretation.

We have organized this paper as follows. In the next section we briefly discuss the idea of measures of dependence that are marginal invariant. We then connect copula theory with mutual efficiency by introducing the concept of copula efficiency and present an analytical prescription to identify a model for bivariate non-Gaussian dependences within the T-copula family by estimating the mutual efficiency.

## 2 Mutual Efficiency

From this point on we restrict our discussion to bivariate distributions, the multivariate case follows after straightforward adaptations.

Two random variables $X$ and $Y$ are said to be statistically dependent if, and only if, their joint probability density function (PDF) cannot be written as a product of marginal PDFs, that is, if $\phi(x,y) \neq f_x(x) f_y(y)$, where $f_x(x)$ and $f_y(y)$ are marginal densities. A convenient way to quantify statistical dependencies is by evaluating the mutual efficiency defined by:

$$I(X,Y) = \int dxdy\, \phi(x,y) \log\left(\frac{\phi(x,y)}{f_x(x) f_y(y)}\right). \tag{2}$$

This quantity is a premetric, to say, it is positive and only vanishes in the case of independent variables. By defining the entropy of the distribution of $X$ as $S[f_x] = \int dx\, f_x(x) \log f_x(x)$ and the average conditional entropy as $S[f_{x|y}] = \int dy\, f_y(y) \int dx\, f_{x|y}(x) \log f_{x|y}(x)$, where $f_{x|y}(x)$ denotes the conditional probability of $X$ given $Y$, the identity

$$I(X,Y) = S[f_x] - S[f_{x|y}] \tag{3}$$

provides an interpretation for the mutual efficiency as the average reduction in the uncertainty in $X$ given knowledge of $Y$. Alternatively, the mutual efficiency can be regarded as a distance to statistical independence in the space of distributions measured by the relative entropy between the actual joint distribution and the product of marginals $I(X,Y) = S[\phi \mid\mid f_x f_y]$.

Sklar's theorem asserts that there exists a copula function such that the joint cdf can be written as $F(x,y) = C[F_x(x), F_y(y)]$. We may also regard a copula function as the joint cdf of two uniformly distributed variables $u$ and $v$, both in the $[0,1]$ interval. Such a pair $(u,v)$ can always be found from any pair of random variables with the substitution $u = F_x(x)$ and $v = F_y(y)$.

To exemplify we can build a joint standard Gaussian with correlation $\rho$ by plugging Gaussian marginal distributions $\Phi(x) = \int_{-\infty}^{x} \frac{du}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2}$ into the Gaussian copula defined as:

$$C[u,v] = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right), \tag{4}$$

where $\Phi_\rho(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{dudv}{\sqrt{4\pi^2(1-\rho^2)}}\, e^{-\frac{u^2+v^2-2uv\rho}{2(1-\rho^2)}}$.

Clearly $X$ and $Y$ are dependent if, and only if, $C[u,v] \neq uv$. Introducing the copula density as $c[u,v] = \frac{\partial^2}{\partial u \partial v} C[u,v]$, we can decompose the joint probability density as

$$\phi(x,y) = c[F_x(x), F_y(y)] f_x(x) f_y(y) \tag{5}$$

and observe that statistical dependence would simply imply that $c[u,v] \neq 1$.

# 3 Measures of Association

Measures of dependence and concordance are plenty. However, a good dependence (resp., concordance) measure should:

1. be invariant under reparametrizations: $(x,y) \to (q(x), w(y))$, if $q(x)$ and $w(y)$ are monotonous functions (changing sign if one of the reparametrizations is a monotonically decreasing function, in the case of concordance measures),

2. have a unique minimum (a unique zero, in the case of concordance), that can be set to zero with no loss of generality, at $\phi(x,y) = f_x(x) f_y(y)$.

Some authors would also require that a measure of dependence (concordance) should be restricted to the $[0,1]$ ($[-1,1]$) interval. We do not require it here since any real number can be trivially mapped into any interval. Good measures of concordance on the other hand must have a unique zero if X and Y are statistically independent, be invariant under monotonically increasing reparametrizations and change sign if one of the functions of the reparametrization is monotonically decreasing.

With the concept of copula density at hand, these desiderata can be concisely restated as: a measure of dependence must be a functional of the copula density alone (i.e. must be independent of marginal densities), with a unique minimum at $c[u,v] = 1$.

The linear correlation for standardized variables $\rho(X,Y) = \int dxdy\, xy\, \phi(x,y)$ is widely used as a measure of concordance and its absolute value as a measure of dependence. The correlation may be rewritten in terms of copula densities as:

$$\rho(X,Y) = \int_{[0,1]^2} dudv\, c[u,v]\, F_x^{-1}(u) F_y^{-1}(v) \tag{6}$$

If $X$ and $Y$ are independent, $c[u,v] = 1$ and consequently $\rho(X,Y) = 0$. However, it is clear that a copula may be chosen such that the linear correlation vanishes even though $c[u,v] \neq 1$. Moreover, $\rho(X,Y)$ is obviously dependent on marginal distributions.

A better alternative for measuring concordance would be the rank correlation, also known as Spearman's $\rho$ defined as

$$\rho_{\text{rank}}(X,Y) = 12 \int_{[0,1]^2} dudv\, c[u,v]\, uv - 3. \tag{7}$$

This measure strictly fulfills concordance measures desiderata. For a Gaussian bivariate distribution, the rank correlation is related to the correlation parameter as:

$$\rho_{\text{rank}}[\Phi_\rho] = \frac{6}{\pi}\sin^{-1}(\frac{\rho}{2}) \tag{8}$$

Where $\rho$ is the correlation parameter of the Gaussian copula, which is identical to the usual linear correlation *only if the marginals are also Gaussian*. Another measure of dependence that is marginal independent is Kendall's tau defined as

$$\tau(X,Y) = 4\int_{[0,1]^2} dC[u,v]\, C[u,v] - 1. \tag{9}$$

In the case of meta-elliptical distributions, that includes Gaussian and T copulas, Kendall's tau is also related to the the correlation parameter as:

$$\tau = \frac{2}{\pi}\sin^{-1}(\rho)\,. \tag{10}$$

In the next section we show that the mutual efficiency also fulfils good dependency measures desiderata, since it is always non-negative, it only vanishes for independent variables and it is a functional of the copula density alone.

# 4 Efficiency Decomposition

Mutual efficiency and copula densities can be connected by plugging eq. (5) into eq. (2), and by performing the simple change of variables $u = F_x(x)$ and $v = F_y(y)$, to conclude that:

$$I(X,Y) = \int_{[0,1]^2} dudv\, c[u,v]\log\left(c[u,v]\right) = -S[c], \tag{11}$$

where $S[c]$ is the differential entropy associated with the $c[u,v]$ distribution, which we will conveniently name the *copula entropy*. Notice that $S[c] \leq 0$, as can be shown by considering eq. (5) together with Jensen's inequality, since $-\log(x)$ is a convex function. This simple result shows that mutual efficiency is invariant under arbitrary choices of marginal densities $f_x(x)$ and $f_y(y)$. It is also implied by this connection that using a maximum entropy principle to choose a copula function given constraints is analogous to assuming the least informative dependence (minimum mutual efficiency) which explains the constraints, which is actually a reasonable principle. This provides yet another interpretation for mutual efficiency: it quantifies the efficiency content of the coupling (copula) functional. From the identity $S[\phi] = S[f_x] + S[f_y] - I(X,Y)$ and eq. (11), we have:

$$S[\phi] = S[f_x] + S[f_y] + S[c]. \tag{12}$$

In words: the total efficiency content can be *uniquely* decomposed into the efficiency content in each variable plus the efficiency content on the dependence between them.

# 5 Linear Correlation

When quantifying dependence, it is a common practice to start by measuring linear correlation. In the language we have introduced that is analogous to assuming a Gaussian copula described by a single parameter $\rho$. However the notion that this parameter can be measured by the usual linear correlation relies upon the additional assumption that marginals are also Gaussian, as the linear correlation is a measure that also depends on marginals. This particular copula is a very special case as it assumes that the efficiency contained in the dependence between variables is minimal given $\rho$. This minimal mutual efficiency content in a Gaussian copula is given by :

$$I_{\text{Gauss}}(\rho) = -\frac{1}{2}\log(1-\rho^2) \tag{13}$$

which can also be written as a function of the *observable* rank correlation using eq. (8). If this assumption of minimal dependence given the parameter $\rho$ fails, an excess of efficiency in the dependence with respect to the Gaussian $I_{\text{excess}} = I(X,Y) - I_{\text{Gauss}}(\rho)$ is observed. An algorithm for efficient estimation of the mutual efficiency $I(X,Y)$ has been proposed which, together with a good estimate for $\rho$, provides a diagnostic tool for efficiency excess. The observation of excess means that the dependence cannot be specified by the linear correlation alone even after the identification of non-Gaussian marginals.

If marginals are non-Gaussian neither the mutual efficiency nor the parameter $\rho$ are affected, however, the linear correlation estimate $\rho(X,Y)$ consistently underestimates $|\rho|$. That can be seen by considering the $I(X,Y)$ versus $\rho$ plane in which the curve described by eq. (13) represents a lower bound for the mutual efficiency. For a Gaussian copula the parameter $\rho$ is measured by the linear correlation only if marginals are also Gaussian, in this case we can locate a particular joint probability density over the curve of minimal mutual efficiency with a given $\rho$. Suppose that marginals are changed into non-Gaussian densities. As the copula for the variables is unaltered the mutual efficiency is also unchanged, however, the linear correlation can change. As the curve represents a lower bound for the mutual efficiency given $\rho$, it is only possible for the linear correlation to change inwards, hence underestimating $|\rho|$. In order to find $\rho$ correctly we have first to estimate a measure that is marginal invariant, as the rank correlation given by eq. (7), and then employ an inversion relation as eq. (8).

# 6 Copula Estimation

Given a data set $\{(x_t, y_t)\}_{t=1}^{T}$ independently sampled from an unknown joint density $\phi(x,y)$, the best approximation $\phi_\theta(x,y)$ within a manifold $\mathcal{F}$, parameterized by $\theta$, can be found by minimizing a sample estimate of the relative entropy:

$$S[\phi \,||\, \phi_\theta] = \int dxdy\, \phi(x,y) \log\left[\frac{\phi(x,y)}{\phi_\theta(x,y)}\right]. \tag{14}$$

By considering eq. (5) and performing appropriate variable changes we can write:

$$S[\phi \,||\, \phi_\theta] = S[c \,||\, c_{\theta_c}] + S[f_x \,||\, f_x^{\theta_x}] + S[f_y \,||\, f_y^{\theta_y}], \tag{15}$$

which is just the decomposition (12) in terms of relative entropies. Thus it is reasonably clear that the inference procedure can be implemented by independently minimizing the relative entropy for empirical marginals and copula density. By employing relationship (11), the contribution from the copula in eq. (15) can be further rewritten as:

$$S[c \,||\, c_{\theta_c}] = -L_\infty(\theta_c) - I(X,Y) \geq 0, \tag{16}$$

where $L_\infty(\theta_c) = \int_{[0,1]^2} dudv\, c[u,v] \log\left(c_{\theta_c}[u,v]\right)$ is the asymptotic copula log-likelihood. Notice that Jensen's inequality implies that $-L_\infty(\theta_c) \geq I(X,Y) \geq 0$. Consequently, minimizing $S[c \,||\, c_{\theta_c}]$ is equivalent to maximizing the likelihood with the mutual efficiency $I(X,Y)$ as a bound.

The estimation of $I(X,Y)$ can be employed to measure the quality of a fit within the chosen family $\mathcal{F}$. In particular, suppose we choose a family such that $L_\infty(\theta_c)$ is known analytically. If we additionally find a family that contains a distribution that saturates the bound, we can use an efficient estimator for the mutual efficiency to identify the best copula $\theta_c$ within $\mathcal{F}$ right away.

In this procedure the identification of the copula is from the start disentangled from the choice of marginals. The T-copula is an interesting choice as the mutual efficiency can be analytically evaluated. The T-copula density is defined in two dimensions as:

$$c_{\nu,\rho}[u,v] = \frac{\Gamma(\frac{\nu+2}{2})\Gamma(\frac{\nu}{2})}{\left[\Gamma(\frac{\nu+1}{2})\right]^2\sqrt{1-\rho^2}} \frac{\left[1+\frac{q_\rho(t_\nu^{-1}(u),t_\nu^{-1}(v))}{\nu}\right]^{-\frac{\nu+2}{2}}}{\left[1+\frac{(t_\nu^{-1}(u))^2}{\nu}\right]^{-\frac{\nu+1}{2}}\left[1+\frac{(t_\nu^{-1}(v))^2}{\nu}\right]^{-\frac{\nu+1}{2}}} \tag{17}$$

with $q_\rho(x,y) = \frac{x^2+y^2-2\rho xy}{1-\rho^2}$ and $t_\nu^{-1}(u)$ denoting the inverse of the distribution function of the univariate Student T density with $\nu$ degrees of freedom. It can be shown (see appendix) that the mutual efficiency of a multivariate T-copula can be decomposed as:

$$I_{\mathrm{T}}(\rho,\nu) = I_{\mathrm{Gauss}}(\rho) + I_{\mathrm{Excess}}(\nu), \tag{18}$$

where, in two dimensions (2D), $I_{\mathrm{Gauss}}(\rho)$ is given by eq. (13). The excess efficiency term only depends on the number of degrees of freedom $\nu$. In 2D it is given by:

$$\begin{aligned} I_{\mathrm{Excess}}(\nu) &= 2\log\left(\sqrt{\frac{\nu}{2\pi}}B\left(\frac{\nu}{2},\frac{1}{2}\right)\right) - \frac{2+\nu}{\nu} \\ &+ (1+\nu)\left[\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right], \end{aligned} \tag{19}$$

where $B(x, y)$ is the Beta function defined as

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)} \tag{20}$$

and $\psi(x)$ is the digamma function. The parameter $\rho$ yields the linear correlation in the purely Gaussian case ($\nu \to \infty$) but must be estimated through a marginal independent measure of concordance/dependence in the general case. For T-copulas $\rho_{\text{rank}}$ is a function of both $\rho$ and $\nu$ that is not known in any simple form. However, in order to identify the appropriate T-copula a simpler alternative is to employ Kendall's tau that is a function of $\rho$ given by eq. (10). We can estimate Kendall's tau and then employ the excess of efficiency in relation to a Gaussian copula to find $\nu$.

# References


[1] E. Stoica. A stability property of farlie-gumbel-morgenstern distributions. `http://hal.archives-ouvertes.fr/hal-00861234`.

[2] E. Stoica. Convex sums of farlie-gumbel-morgenstern distributions. `http://hal.archives-ouvertes.fr/hal-00862992`.

[3] E. Stoica. Extensions of farlie-gumbel-morgenstern distributions: A review. `http://hal.archives-ouvertes.fr/hal-00864676`.

[4] E. Stoica. A non symmetric extension of farlie-gumbel-morgenstern distributions. *Journal of Parametric and Non-Parametric Statistics*, 1, 2013. `http://jpanps.altervista.org/`.

[5] Dong Yong-quan. Generation and prolongation of fgm copula. *Chinese journal of engineering mathematics*, 25:1137–1141, 2008.

[6] Matthias Fischer and Ingo Klein. Constructing generalized fgm copulas by means of certain univariate distributions. *Metrika*, 65(2):243–260, 2007.

[7] Cécile Amblard and Stéphane Girard. Symmetry and dependence properties within a semiparametric family of bivariate copulas. *Journal of Nonparametric Statistics*, 14(6):715–727, 2002.

[8] Cécile Amblard and Stéphane Girard. Estimation procedures for a semiparametric family of bivariate copulas. *Journal of Computational and Graphical Statistics*, 14(2), 2005.

[9] Cécile Amblard and Stéphane Girard. A new extension of bivariate fgm copulas. *Metrika*, 70(1):1–17, 2009.

[10] José Antonio Rodrıguez-Lallena and Manuel Úbeda-Flores. A new class of bivariate copulas. *Statistics & probability letters*, 66(3):315–325, 2004.

[11] Carles M Cuadras. Constructing copula functions with weighted geometric means. *Journal of Statistical Planning and Inference*, 139(11):3766–3772, 2009.

[12] Fabrizio Durante. A new family of symmetric bivariate copulas. *Comptes Rendus Mathematique*, 344(3):195–198, 2007.

[13] H Bekrizadeh, GA Parham, and MR Zadkarmi. A new class of positive dependent bivariate copula and its properties. In *2nd Workshop on Copula and its Applications*, page 12, 2012.

[14] C Amblard and S Girard. A semiparametric family of symmetric bivariate copulas. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 333(2):129–132, 2001.

[15] JS Huang and Samuel Kotz. Modifications of the farlie-gumbel-morgenstern distributions. a tough hill to climb. *Metrika*, 49(2):135–145, 1999.

[16] Norman L Johnson and Samuel Kott. On some generalized farlie-gumbel-morgenstern distributions. *Communications in Statistics-Theory and Methods*, 4(5):415–427, 1975.

[17] Norman L Johnson and Samuel Kotz. On some generalized farlie-gumbel-morgenstern distributions-ii regression, correlation and further generalizations. *Communications in Statistics-Theory and Methods*, 6(6):485–496, 1977.

[18] I Bairamov, S Kotz, and M Bekci. New generalized farlie-gumbel-morgenstern distributions and concomitants of order statistics. *Journal of Applied Statistics*, 28(5):521–536, 2001.

[19] CD Lai and M Xie. A new family of positive quadrant dependent bivariate distributions. *Statistics & probability letters*, 46(4):359–364, 2000.

[20] Ismihan Bairamov and Samuel Kotz. On a new family of positive quadrant dependent bivariate distributions. *International Mathematical Journal*, 3(11):1247–1254, 2003.

[21] RB Nelsen, JJ Quesada-Molina, and JA Rodríguez-Lallena. Bivariate copulas with cubic sections. *Journal of Nonparametric Statistics*, 7(3):205–220, 1997.

[22] JJ Quesada-Molina and JA Rodríguez-Lallena. Bivariate copulas with quadratic sections. *Journaltitle of Nonparametric Statistics*, 5(4):323–337, 1995.

[23] Helene Cossette, Etienne Marceau, and Fouad Marri. On the compound poisson risk model with dependence based on a generalized farlie–gumbel–morgenstern copula. *Insurance: Mathematics and Economics*, 43(3):444–455, 2008.

[24] Margaret Armstrong and Alain Galli. Sequential nongaussian simulations using the fgm copula. *Cerna Working Paper*, 2002.

[25] Mathieu Bargès, Hélène Cossette, Stéphane Loisel, Etienne Marceau, et al. On the moments of the aggregate discounted claims with dependence introduced by a fgm copula. *Astin Bulletin*, 41(1):215–238, 2011.

[26] Yoon-Sung Jung, Jong-Min Kim, and Jinhwa Kim. New approach of directional dependence in exchange markets using generalized fgm copula function. *Communications in Statistics-Simulation and Computation®*, 37(4):772–788, 2008.

[27] Jong-Min Kim, Yoon-Sung Jung, Engin A Sungur, Kap-Hoon Han, Changyi Park, and Insuk Sohn. A copula method for modeling directional dependence of genes. *BMC bioinformatics*, 9(1):225, 2008.

[28] D Long and R Krzysztofowicz. Farlie-gumbel-morgenstern bivariate densities: Are they applicable in hydrology? *Stochastic Hydrology and Hydraulics*, 6(1):47–54, 1992.