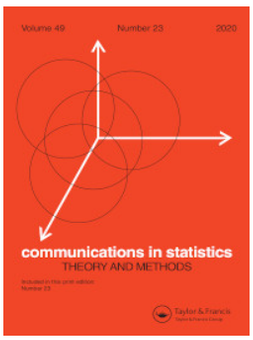

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Two generalized bivariate FGM distributions and rank reduction

Carles M. Cuadras, Walter Diaz & Sonia Salvo-Garrido

To cite this article: Carles M. Cuadras, Walter Diaz & Sonia Salvo-Garrido (2020) Two generalized bivariate FGM distributions and rank reduction, Communications in Statistics - Theory and Methods, 49:23, 5639-5665, DOI: 10.1080/03610926.2019.1620780

To link to this article: https://doi.org/10.1080/03610926.2019.1620780

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2020, VOL. 49, NO. 23, 5639–5665
https://doi.org/10.1080/03610926.2019.1620780

Two generalized bivariate FGM distributions and rank reduction

Carles M. Cuadras[a], Walter Diaz[b], and Sonia Salvo-Garrido[c]

[a]Univ. of Barcelona, Statistics, Barcelona, Spain; [b]Fac. of Economics, Univ. of Antioquia, Medellin, Colombia; [c]Dpt. of Mathematics and Statistics, Univ. La Frontera, Temuco, Chile

ABSTRACT

The Farlie-Gumbel-Morgensten (FGM) family of bivariate distributions with given marginals, is frequently used in theory and applications and has been generalized in many ways. With the help of two auxiliary distributions, we propose another generalization and study its properties. After defining the rank of a distribution as the cardinal of the set of canonical correlations, we prove that some well-known distributions have practically rank two. Consequently we introduce several extended FGM families of rank two and study how to approximate any bivariate distribution to a simpler one belonging to this family.

ARTICLE HISTORY
Received 18 January 2018
Accepted 14 May 2019

KEYWORDS
Farlie-Gumbel-Morgensten distribution; bivariate copulas; stochastic dependence; pearson contingency coefficient; rank of a distribution

AMS SUBJECT CLASSIFICATION
62H20 (primary);
60E05 (secondary)

1. Introduction

The construction and study of dependence models have interest in statistics, probability, econometrics, informatics, insurance, finance, physics, hydrology, etc. A copula function is a bivariate cdf with uniform (0, 1) marginals that captures the dependence properties of two r.v.'s defined on the same probability space. Many copulas and bivariate families of distributions have been studied in Hutchinson and Lai (1991), Joe (1997), Drouet-Mari and Kotz (2001), Kotz, Balakrishnan, and Johnson (2000), Nelsen (2006), Cuadras (2006) and Balakrishnan and Lai (2009). Among others, the so-called Farlie-Gumbel-Morgenstern (FGM) bivariate family is frequently used in theory and applications. This motivated Huang and Kotz (1999), Lai and Xie (2000), Amblard and Girard (2002, 2009), Rodríguez-Lallena and Úbeda-Flores (2004), Cuadras and Cuadras (2008) and Cuadras and Diaz (2012), to propose and study proper extensions.

Let $\mathbf{I} = [0,1]$. Recall that a bivariate copula is a function $C : \mathbf{I}^2 \to \mathbf{I}$ such that $C(u,0) = C(0,v) = 0$, $C(u,1) = u$, $C(1,v) = v$, and for $0 \le u_1 \le u_2 \le 1$ and $0 \le v_1 \le v_2 \le 1$ satisfies:

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \ge 0$$

Copulas are important because Sklar's theorem (Sklar 1959). Let H be a bivariate cdf with univariate marginals F, G. Then H can be expressed as $H(x,y) = C(F(x), G(y))$, where C is a copula related to H. Thus modeling copulas is an interesting task.

CONTACT Carles M. Cuadras ccuadras@ub.edu Univ. of Barcelona, Statistics, Barcelona, Spain.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.

This paper extends Cuadras, Fortiana, and Greenacre (2000), Cuadras (2015) and Cuadras and Diaz (2012). For the sake of clarity, we repeat some concepts, definitions and results.

Section 2 studies a new generalized family of distributions, which is related to the diagonal extension of a distribution, following Cuadras, Fortiana, and Greenacre (2000). This family has a conjugate one, with similar dependence properties. Several aspects of Section 3 appeared in Cuadras and Diaz (2012) and Cuadras (2015). Here we propose a new terminology, clarify some geometrical concepts and obtain new results. For instance, Theorem 2 is more general than the similar proposition in Cuadras (2015). Once the concept of rank reduction of a cdf has been established, Section 4 proposes a new extension of rank two. Section 5 contains another extension of rank two, already studied in Cuadras and Diaz (2012), but some new results are added. Section 6 is also new and studies the copulas associated to the families defined in Sections 4 and 5. Section 7 proposes a new distance between bivariate distributions, which is related to the diagonal expansion. Section 8 deals with the approximation of a cdf for another of lower rank. It is based on Cuadras and Diaz (2012), but contains new results concerning the quality of the approximation. Section 9 is devoted to illustrate the theoretical results with examples, following the same structure as Cuadras and Diaz (2012), but adding the bivariate normal distribution and the curved line characterizing a bivariate family in terms of canonical correlations.

2. Generalized FGM family of distributions

Let $H(x, y)$ be the bivariate cdf of the random vector (X, Y), with univariate marginal cdf's $F(x), G(y)$ and supports $[a, b], [c, d]$, respectively. Throughout this paper, x and y in $H(x, y)$, $F(x)$, $G(y)$, as well as u and v, in $C(u, v)$, where $0 \leq u, v \leq 1$, will be suppressed, unless it is strictly necessary. We write $H \in \mathcal{F}(F, G)$, where $\mathcal{F}(F, G)$ is the so-called Fréchet-Hoeffding class of bivariate cdf's with fixed univariate marginals F, G, see Nelsen (2006, Chapter 6).

As $\mathcal{F}(F, G)$ is too general, we must work with a sub-class. The FGM family is a parametric sub-class of $\mathcal{F}(F, G)$ defined by

$$H_\theta = FG\big[1 + \theta(1-F)(1-G)\big], \quad -1 \leq \theta \leq 1$$

and the corresponding family of copulas is

$$C_\theta = uv[1 + \theta(1-u)(1-v)], \quad -1 \leq \theta \leq 1$$

2.1. Definition and properties

We propose the following generalization.

Definition 1. Let Φ, Ψ be two univariate cdf's having the same supports $[a, b], [c, d]$ as those of F, G. We define the bivariate family

$$H = FG + \lambda(F-\Phi)(G-\Psi) \tag{1}$$

This family reduces to the bivariate FGM cdf when $\Phi = F^2$ and $\Psi = G^2$, both univariate cdf's.

Consider the Fréchet-Hoeffding class $\mathcal{F}(F,G)$ of bivariate cdf's with univariate marginals F and G. It is readily proved that H has marginals F and G and reaches the stochastic independence cdf $F\times G$ when $\lambda=0$ and/or $F=\Phi$. By continuity, in general H, see (1), is also a cdf for λ near to 0. Thus $H\in\mathcal{F}(F,G)$ for some values of $\dot{\lambda}$. The range of λ is obtained in the following theorem, where $\Phi\ll F$ means that F is absolutely continuous with respect to (w.r.t.) Ψ. Then the Radon-Nikodym derivative $d\Phi/dF$ exists.

Theorem 1. *Suppose $\Phi\ll F$, $\Psi\ll G$ and that $d\Phi(x)/dF(x)\neq 1$, $d\Psi(x)/\ dG(y)\neq 1$ for some x,y. Then H is a bivariate cdf for any λ such that $\lambda_-\leq\lambda\leq\lambda_+$, where*

$$\lambda_- = \frac{-1}{\sup\{[1-d\Phi/dF][1-d\Psi/dG]\}}$$
$$\lambda_+ = \frac{-1}{\inf\{[1-d\Phi/dF][1-d\Psi/dG]\}}$$

Proof. Write $dH = dFdG+\lambda(dF-d\Phi)(dG-d\Psi)$ as

$$dH = dFdG[1+\lambda(1-d\Phi/dF)(1-d\Psi/dG)] \tag{2}$$

We should find λ such that $1+\lambda(1-d\Phi/dF)(1-d\Psi/dG)\geq 0$. This occurs when $\lambda(1-d\Phi/dF)(1-d\Psi/dG)\geq -1$. Thus $\lambda\geq -1/[(1-d\Phi/dF)(1-d\Psi/dG)]$. Hence $\lambda\geq\lambda_-$. Similarly, if the denominator is positive, $\lambda\leq 1/[(1-d\Phi/dF)(1-d\Psi/dG)]$. Hence $\lambda\leq\lambda_+$. □

Let us suppose absolute continuity of the univariate marginal distributions, i.e., there exist the probability density functions (pdf's), w.r.t. to the Lebesgue measure, $f=F',g=G',\varphi=\Phi',\psi=\Psi'$. Then the bivariate pdf is

$$\begin{aligned} h &= fg+\lambda(f-\varphi)(g-\psi)\\ &= fg\left[1+\lambda\left(1-\varphi f^{-1}\right)\left(1-\psi g^{-1}\right)\right]\end{aligned}$$

where $f^{-1}=1/f$ and $g^{-1}=1/g$. Then, if we replace $d\Phi/dF$ by φ/f we get the following result:

Corollary 1. *H is a bivariate cdf for any λ such that $\lambda_-\leq\lambda\leq\lambda_+$, where*

$$\lambda_- = \frac{-1}{\sup\left\{\left[1-\varphi(x)f^{-1}(x)\right]\left[1-\psi(y)g^{-1}(y)\right]\right\}}$$
$$\lambda_+ = \frac{-1}{\inf\left\{\left[1-\varphi(x)f^{-1}(x)\right]\left[1-\psi(y)g^{-1}(y)\right]\right\}}$$

The generalized family (1) is more flexible than FGM and has some advantages. For example, the maximum range of the correlation coefficient for the traditional FGM family is $[-1/3,1/3]$. This range can be improved.

Example 1. Consider the univariate cdf's $\Phi=F-\sin^a(\pi F)/\pi$, $\Psi=G-\sin^a(\pi G)/\pi$, where $1\leq a\leq 2$. From (1) we obtain $H=FG+(\lambda/\pi^2)\sin^a(\pi F)\ \sin^a(\pi G)$. With $a=5/4$ we have $-1.1963\leq\lambda\leq 1.1963$ (Corollary 1), and the range of the correlation coefficient with H (uniform marginals) is $[-0.5108,\ 0.5108]$. For $a=1$ this family gives

a copula appearing in Amblard and Girard (2002) and the range is $[-48/\pi^4, 48/\pi^4] = [-0.4928, 0.4928]$. Both ranges are wider than $[-1/3, 1/3]$.

Lemma 1. *Suppose* $\Phi \ll F$ *and* $\Psi \ll G$. *Let* α *and* β *be defined by*

$$\alpha = \int_a^b \left(\frac{d\Phi}{dF}\right)^2 dF, \qquad \beta = \int_c^d \left(\frac{d\Psi}{dG}\right)^2 dG$$

Then $\alpha \geq 1$ *and* $\beta \geq 1$.

Proof. The derivative $d\Phi/dF$ exists and $\int_a^b (d\Phi/dF - 1)^2 dF = \alpha - 2\int_a^b d\Phi + 1 = \alpha - 1 \geq 0$ and similarly β. Note that α and β are divergence measures in the sense of Csiszár (1975). □

Define the functions $a_1 = 1 - d\Phi/dF$, $b_1 = 1 - d\Psi/dG$. Then from (2)

$$dH = dFdG + \lambda a_1 b_1 dFdG \tag{3}$$

If the pdf's h, f, g exist, we have $h = fg + \lambda f a_1 g b_1$.

Lemma 2. $E[a_1(X)] = E[b_1(Y)] = 0$ and $E[a_1^2(X)] = \alpha - 1$, $E[b_1^2(Y)] = \beta - 1$.

Proof. $E[a_1(X)] = \int_a^b (1 - d\Phi/dF) dF = 1 - 1 = 0$. From Lemma 1 $E[a_1^2(X)] = \alpha - 1$. □

2.2. Relation with the diagonal expansion

Suppose that $H \ll FG$, so the derivative $dH/(dFdG)$ exists. A global measure of dependence is the *Pearson contingency coefficient* ${\phi_t}^2$, defined by

$${\phi_t}^2 = \int_a^b \int_c^d \left(\frac{dH}{dFdG} - 1\right)^2 dFdG \tag{4}$$

We have ${\phi_t}^2 \geq 0$ and ${\phi_t}^2 = 0$ iff there is stochastic independence between X and Y. Note that ${\phi_t}^2$ is a divergence measure in the sense of Csiszár (1975).

If ${\phi_t}^2$ is finite, then the kernel $K = dH/(dFdG) - 1$ is Hilbert-Schmidt. This means that there exists a sequence $K_N = \sum_{n=1}^N \rho_n a_n b_n$ converging to K as $N \to \infty$. Thus, if ${\phi_t}^2$ is finite, there exists the expansion (Lancaster 1958)

$$dH = dFdG + \sum_{n \geq 1} \rho_n a_n b_n dFdG \tag{5}$$

where a_n, b_n are unitary functions in $L^2([a, b])$ and $L^2([c, d])$ on F and G, respectively, in the sense that $E[a_n(X)] = E[b_n(Y)] = 0$ and $E[a_n^2(X)] = E[b_n^2(Y)] = 1$. Then $a_n(X)$ and $b_n(Y)$ are the canonical variables. The sequence of canonical correlations is $\rho_1 \geq \rho_2 \geq \cdots \geq 0$. The canonical variables are functions with maximal correlations. Thus $\rho_n = \text{cor}(a_n(X), b_n(Y))$, where cor means correlation coefficient, is maximal in the sense of canonical correlation analysis, a well-known method of multivariate analysis, see Mardia, Kent, and Bibby (1979). Hence $a_1(X)$ and $b_1(Y)$ have maximum correlation, $a_2(X)$ and $b_2(Y)$ have maximal correlation constrained to zero correlation with $a_1(X)$

and $b_1(Y)$, respectively, and so on. When H can be expanded as (5), it is said that H admits a diagonal expansion (Hutchinson and Lai 1991).

It can be proved that the Pearson contingency coefficient, see (4), can be expressed in terms of the sequence of canonical correlations:

$$\phi_t^2 = \sum_{n\geq 1} \rho_n^2$$

The first canonical correlation ρ_1 is the *maximum correlation* between a function of X and a function of Y :

$$\rho_1 = \sup_{\nu\in V_x,\ \xi\in V_y} \operatorname{cor}\big(\nu(X), \xi(Y)\big)$$

where V_x, V_y are the sets of functions with finite variance. The correlation ρ_1 is a measure of dependence, since $\rho_1 = 0$ iff the r.v.'s X, Y are stochastically independent and $\rho_1 = 1$ iff there is a functional relation between the variables, which is useful in identifying nonlinear relationships in regression. See Buja (1990).

If the pdf's exist, expansion (5) can be expressed as

$$h = fg + \sum_{n\geq 1} \rho_n f a_n g b_n \tag{6}$$

It is worth noting that the sequence of canonical correlations can be an interval rather than a countable set. This may occur when the distribution has a singular part. For instance, consider the Cuadras-Augé family

$$H = \min\{F, G\}^{\theta}(FG)^{1-\theta}, \qquad 0 \leq \theta \leq 1$$

The line $\{F(x) = G(y)\}$ has measure 0 w.r.t. $dFdG$, but positive measure w.r.t. dH. Hence dH is not absolutely continuous w.r.t. $dFdG$ and the derivative $dH/(dFdG)$ does not exist. We cannot express $dH = dFdG + \sum \rho_n a_n b_n dFdG$, i.e., the standard Lancaster theory does not apply for this family. Instead of a sequence $\rho_1 \geq \cdots \geq \rho_n \geq \cdots \geq 0$, for the Cuadras-Augé family the set of canonical correlations is obtained in a different way and described by the continuous function $\theta\rho^{1-\theta}, 0 \leq \rho \leq 1$. The range of the correlations is the interval $[0, \theta]$. Hence θ is the maximum correlation. See Cuadras (2002, 2015, 2016), Ruiz-Rivas and Cuadras (1988).

The constants α, β in the proposition below, has been defined in Lemma 1.

Proposition 1. *The first canonical correlation for the family (1) is given by*

$$\rho_1 = \lambda\sqrt{(\alpha-1)(\beta-1)}$$

and the Pearson contingency coefficient is ${\phi_t}^2 = \rho_1^2$.

Proof. Suppose $\alpha > 1, \beta > 1$. Write (3) as

$$dH = dFdG + \lambda\Big(a_1/\sqrt{\alpha-1}\ \Big)\Big(b_1/\sqrt{\beta-1}\ \Big)dFdG$$

and compare with (5). Then $A_1 = a_1/\sqrt{\alpha-1}$ and $B_1 = b_1/\sqrt{\beta-1}$ are the first canonical functions. The contingency coefficient is $\phi_t^2 = \int_a^b \int_c^d \lambda^2 a_1^2 dF b_1^2 dG = \lambda^2(\alpha-1)(\beta-1)$. □

2.3. Conjugate family

The family (1) with marginals F, G, generated by Φ, Ψ, suggests the following conjugate family

$$H_* = \Phi\Psi + \lambda(F-\Phi)(G-\Psi)$$

with marginals Φ, Ψ, here generated by F, G. Clearly $H_* \in \mathcal{F}(\Phi, \Psi)$ for suitable values of λ, and $H-FG = H_*-\Phi\Psi$, so H and its conjugate H_* should have the same dependence structure. Even though "conjugate" is used in Bayesian statistics, here this adjective means relationship between two distributions.

2.4. Dependence measures

In this section we find the covariance and two non-parametric measures of association.

If $\nu(x), \xi(y)$ are two real functions of bounded variation on $[a, b]$, $[c, d]$, Cuadras (2002) proved that

$$\operatorname{cov}\big(\nu(X), \xi(Y)\big) = \int_a^b \int_c^d \big[H(x,y)-F(x)G(y)\big]\, d\nu(x) d\xi(y) \tag{7}$$

where cov means covariance. This formula has been generalized by Diaz and Cuadras (2017).

Lemma 3. *Suppose that* $\lim_{x\to b} x[F(x)-\Phi(x)] = 0$. *Then*

$$\int_a^b (F-\Phi)dx = \mu_\Phi - \mu_F$$

where μ_F, μ_Φ *are the expectation values.*

Proof. Integrating

$$\begin{aligned} \textstyle\int_a^b (F-\Phi)dx &= [xF(x)-x\Phi(x)]_a^b - \textstyle\int_a^b x dF + \int_a^b x d\Phi \\ &= \mu_\Phi - \mu_F \end{aligned}$$

If we consider the extended real line, this difference can also be proved from

$$\mu_\Phi - \mu_F = \int_0^\infty (1-\Phi)dx - \int_{-\infty}^0 \Phi dx - \int_0^\infty (1-F)dx + \int_{-\infty}^0 F dx \qquad \square$$

If $(X, Y) \sim H$ and $(X^*, Y^*) \sim H_*$, where H and H_* are conjugate, both cdf's have the same covariance:

$$\begin{aligned} \operatorname{cov}(X, Y) &= \lambda \int_a^b (F-\Phi)dx \int_c^d (G-\Psi)dy \\ &= (\mu_\Phi - \mu_F)(\mu_\Psi - \mu_G) \\ &= \operatorname{cov}(X^*, Y^*) \end{aligned}$$

Spearman's rho coefficient is given by $\rho_S = \operatorname{cor}(F(X), G(Y)) = 12\ \operatorname{cov}(F(X),\ G(Y))$. To find this coefficient, write $F_\Phi(b) = \int_a^b \Phi dF$, $\Phi_F(b) = \int_a^b F d\Phi = 1-F_\Phi(b)$ and similarly $G_\Psi(d), \Psi_G(d)$. We have $\int_a^b (F-\Phi)dF = 1/2 - \int_a^b \Phi dF = \int_a^b (\Phi - F)d\Phi$.

Proposition 2. *Spearman's rho is given by*

$$\rho_S = 12\lambda\left[\frac{1}{2}-F_\Phi(b)\right]\left[\frac{1}{2}-G_\Psi(d)\right] = 12\lambda\left[\Phi_F(b)-\frac{1}{2}\right]\left[\Psi_G(d)-\frac{1}{2}\right]$$

Hence H and its conjugate H_* have the same rho.

Proof. As var$[F(X)]$ =var$[G(Y)] = 1/12$, Spearman's rho is given by

$$12\lambda\int_a^b (F-\Phi)dF\int_c^d (G-\Psi)dG = 12\lambda\int_a^b (\Phi-F)d\Phi\int_c^d (\Psi-G)d\Psi$$ □

Kendall's tau is a measure of association given by $\tau = 4\int_a^b\int_c^d HdH-1$. To find this coefficient, from (1) we have $dH = dFdG + \lambda(dFdG + d\Phi d\Psi - d\Phi dG - dFd\Psi)$.

Proposition 3. *Kendall's tau is given by*

$$\tau = 8\lambda\left[\frac{1}{2}-F_\Phi(b)\right]\left[\frac{1}{2}-G_\Psi(d)\right] = 8\lambda\left[\Phi_F(b)-\frac{1}{2}\right]\left[\Psi_G(d)-\frac{1}{2}\right]$$

Hence H and its conjugate H_* have the same tau.

Proof. After a tedious algebra, we find

$$\int_a^b\int_c^d HdH = \int_a^b\int_c^d \left[FG+\lambda(F-\Phi)(G-\Psi)\right]dH = 1/4 + 2\lambda\left[F_\Phi(b)-\frac{1}{2}\right]\left[G_\Psi(d)-\frac{1}{2}\right] + \lambda^2\times 0$$

Thus

$$\tau = 8\lambda\left[\frac{1}{2}-F_\Phi(b)\right]\left[\frac{1}{2}-G_\Psi(d)\right]$$

For the conjugate distribution H_* we similarly find

$$\tau_* = 8\lambda\left[\Phi_F(b)-\frac{1}{2}\right]\left[\Psi_G(d)-\frac{1}{2}\right]$$

As $F_\Phi(b) = 1-\Phi_F(b), G_\Psi(d) = 1-\Psi_G(d)$, clearly $\tau = \tau_*$. □

Since $2\rho_s = 3\tau$, both coefficients satisfy the well-known inequality $-1 \le 3\tau-2\rho_s \le 1$.

3. Rank of a bivariate distribution

In this section we apply the concept of dimensionality reduction, which is quite useful in multivariate data analysis. This methodology consists in representing in low dimension (e.g., two or three), objects or individuals described by coordinates in high dimensional spaces. Principal component analysis is a well-known method.

3.1. Definition and geometrical meaning

The sequence $\rho_1 \geq \rho_2 \geq \cdots$ of canonical correlations, see (5), captures the full dependence between X and Y and the Pearson coefficient ϕ_t^2 is an overall measure of dependence, sometimes presented as the ratio $\phi_t^2/(1+\phi_t^2)$.

Definition 2. The rank of $H \in \mathcal{F}(F,G)$ such that the diagonal expansion $dH = dFdG + \sum_{n\geq 1} \rho_n A_n dF B_n dG$ exists, is the cardinal of the set $\{\rho_n\}$.

The rank of a distribution can be understood as a "geometric dimension" in regard with the so-called chi-square distance. Borrowing geometric concepts commonly used in multivariate analysis, this means that the observations can be embedded in a Euclidean (or Hilbert) space. Then, as usual in multivariate analysis, we are interested in the first principal dimensions. In other words, we seek the first canonical correlations, see below. Some examples are:

1. The independence distribution $F \times G$ has rank 0.
2. The FGM distribution has rank 1.
3. The distribution
$$H = FG + \lambda_1 F(1-F)G(1-G) + \lambda_2(2F-1)F(1-F)(2G-1)G(1-G) \tag{8}$$
has rank 2.
4. The Ali-Mikhail-Haq (AMH) distribution defined by
$$H = FG/\big[1-\theta(1-F)(1-G)\big], \quad -1 \leq \theta \leq 1$$
has infinite countable rank (see Ali, Mikhail, and Haq 1978).
5. The Cuadras-Augé distribution has uncountable rank. This property has been discussed above. See Cuadras (2015), Cuadras and Augé (1981) and Section 2.2.

Definition 3. The chi-square distance between two observations x, x' of X is

$$\delta^2(x,x') = \int_c^d \left[\frac{dH(x,y)}{dF(x)dG(y)} - \frac{dH(x',y)}{dF(x')dG(y)}\right]^2 dG(y)$$

Proposition 4. *Suppose that the diagonal expansion* $dH = dFdG + \sum_{n\geq 1} \rho_n\ A_n B_n dFdG$ *exists. Then*

$$\delta^2(x,x') = \sum_{n\geq 1} \rho_n^2 \big[A_n(x) - A_n(x')\big]^2$$

where $(\rho_1 A_1(x), \rho_2 A_2(x), \ldots)$ *are the principal coordinates of* x *w.r.t. the chi-square distance. Therefore, the embedding* $x \to (\rho_1 A_1(x), \rho_2 A_2(x), \ldots) \in \mathbf{L}$, *shows that the rank is the dimension of* $\mathbf{L}$, *where* $\boldsymbol{L}$ *is a Euclidean (or separable Hilbert) space.*

Proof. We can write the chi-square distance as

$$\delta^2(x,x') = E_Y\left\{\sum_{n\geq 1}\big[\rho_n A_n(x)B_n(Y) - \rho_n A_n(x')B_n(Y)\big]^2\right\}$$

and take step-wise expectation such that $E_Y B_i(Y)B_j(Y) = \delta_{ij}$ (Kronecker's delta). □

This result proves that the rank of a bivariate distribution makes geometric sense.

As a generalization of the variance, the geometric variability (also called inertia and diversity coefficient) of X w.r.t. the chi-square distance, is the average:

$$V_g = \frac{1}{2}\int_a^b \int_a^b \delta^2(x,x')dF(x)dF(x')$$

Proposition 5. *If the Pearson contingency coefficient ϕ_t^2 is finite then*

$$V_g = \phi_t^2 = \sum_{n\geq 1} \rho_n^2$$

Proof. Suppose X, X' i.i.d. with cdf F. Then $E_{XX'}\{\rho_n^2[A_n(X)-A_n(X')]^2\} = 2\rho_n^2$ and $V_g = \sum_{n\geq 1}\rho_n^2$ holds. On the other hand, from (4) and assuming $X_1 \sim F$, independent of $Y_1 \sim G$, then $\phi_t^2 = E_{X_1Y_1}\{\sum_{n\geq 1}\rho_n A_n(X_1)B_n(Y_1)\}^2 = V_g$. □

3.2. Rank reduction

From the above diagonal expansion we can consider the following family

$$dH_\lambda = dFdG + \sum_{n\geq 1}\lambda_n A_n dF B_n dG, \quad |\lambda_n| \leq \rho_n \tag{9}$$

By integration we can express

$$H_\lambda(x,y) = F(x)G(y) + \sum_{n\geq 1}\lambda_n \Phi_n(x)\Psi_n(y) \tag{10}$$

where $\Phi_n(x) = \int_a^x A_n(t)dF(t)$ and similarly $\Psi_n(y)$. In general $\sum_{n\geq 1}\lambda_n\Phi_n\Psi_n$ is not an eigenexpansion of $H - FG$ (see Cuadras and Cuadras 2008). This family, as well as the function obtained after reducing the rank, is in general a signed measure (it may take negative values in a region of the support). As it is justified below, this reduction can provide a proper cdf.

Note that $FG + \lambda_n\Phi_n\Psi_n \in \mathcal{F}(F,G)$ for any λ_n such that

$$\alpha_n = \inf_{x,y}\left\{\frac{f(x)g(y)}{\Phi_n'(x)\Psi_n'(y)}\right\} \leq \lambda_n \leq \sup_{x,y}\left\{\frac{f(x)g(y)}{\Phi_n'(x)\Psi_n'(y)}\right\} = \beta_n$$

We have assumed that the pdf's exist. The following condition is necessary in order to restrict the parameters of a cdf.

Theorem 2. *Let us consider the expansions (9) and (10). Write $0 = (0,0,\ldots)$, $\boldsymbol{\lambda} = (\lambda_1,\lambda_2,\ldots)$, $\boldsymbol{\lambda}' = (\lambda_1',\lambda_2',\ldots)$, suppose all $|\lambda_n'| \leq |\lambda_n|$ and define $\mu_n = \alpha_n$ if $\lambda_n < 0$, $\mu_n = \beta_n$ if $\lambda_n > 0$. Also define $P(\mathbf{t}) = \inf_{x,y}[f(x)g(y) + \sum_{n\geq 1} t_n\Phi_n'(x)\Psi_n'(y)]$ and*

$$H_{\lambda'} = FG + \sum_{n\geq 1}\lambda_n'\Phi_n\Psi_n$$

Then $H_{\lambda'} \in \mathcal{F}(F,G)$ if a) $\alpha = \sum_{n\geq 1}(\lambda_n'/\mu_n)$ satisfies $0 \leq \alpha \leq 1$, or b) $\boldsymbol{\lambda}'$ is a point of a curve inside $\mathbb{F} = \{\mathbf{t}|\ P(\mathbf{t}) \geq 0\}$ joining 0 and $\boldsymbol{\lambda}$. In particular, $H_{\lambda'} \in \mathcal{F}(F,G)$ if $\lambda_n' = c\lambda_n$ for some constant $0 \leq c \leq 1$.

Table 1. Percentage of variability of the distribution in terms of the first and second canonical correlation for four families of copulas. This percentage is 100 times the quotient between the sum of the first squared canonical correlations and the Pearson contingency coefficient.

	$\theta = 1$		$\theta = 0.5$	
$P_k = (\sum_1^k \rho_i^2)/\phi_t^2$	$100P_1$	$100P_2$	$100P_1$	$100P_2$
Ali−Mikhail−Haq	61.2	89.6	98.9	99.9
Gumbel−Barnett	86.8	98.3	92.0	99.3
Celebioglu−Cuadras	99.3	99.9	99.8	99.9
New family	99.8	99.9	99.9	100

Proof. Clearly $FG + \mu_n \Phi_n \Psi_n \in \mathcal{F}(F, G)$ and $0 < \lambda'_n/\mu_n < 1$ for all n. Then we have that

$$H_{\lambda'} = (1-\alpha)FG + \sum_{n\geq 1} (\lambda'_n/\mu_n)(FG + \mu_n \Phi_n \Psi_n)$$

with $|\lambda'_n| \leq |\mu_n|$ is a mixture of the cdf's FG, $FG + \mu_1 \Phi_1 \Psi_1$, $FG + \mu_2 \Phi_2 \Psi_2$, etc. Hence $H_{\lambda'}$ is also a cdf belonging to $\mathcal{F}(F, G)$.

On the other hand, if $\mathbf{t} \in \mathbb{F}$ then $FG + \sum_{n\geq 1} t_n \Phi_n \Psi_n$ belongs to $\mathcal{F}(F, G)$. Of course $0 \in \mathbb{F}$ and $\boldsymbol{\lambda} \in \mathbb{F}$. As $\mathcal{F}(F, G)$ is closed under mixtures, $\mathbb{F}$ is a convex set. Then any regular curve joining 0 and $\boldsymbol{\lambda}$ inside $\mathbb{F}$ provides cdf's of $\mathcal{F}(F, G)$. In particular a straight line. □

Example 2. For the copula

$$C = uv + \lambda_1 u(1-u)v(1-v) + \lambda_2(2u^3-3u^2+u)(2v^3-3v^2+v)$$

we have $\alpha_1 = -1$, $\beta_1 = 1$, $\alpha_2 = -1$, $\beta_2 = 2$. Then, if λ_1, λ_2 are positive, we get a copula for λ'_1, λ'_2 positive such that $\lambda'_1 + \lambda'_2/2 \leq 1$, provided that λ_1, λ_2 give a proper copula C and $\lambda'_1 \leq \lambda_1, \lambda'_2 \leq \lambda_2$.

As a consequence of Theorem 2, the cdf H with finite rank N, or infinite countable rank, can be approximated by H_D with smaller rank D (i.e., $D < N$ or $D < \infty$), defined by

$$dH_D = dFdG + \sum_{n=1}^{D} \rho_n A_n B_n dFdG$$

In general, H_D is a signed measure. H_D is a proper cdf if each $dH_{(n)} = dFdG + \rho_n A_n B_n dFdG$ is the differential of a cdf $H_{(n)}$. A simple example is the FGM approximation of rank $D = 1$ to the cdf (8) whose rank is $N = 2$.

If the densities h, f, g exist, we have $h = fg + \sum_{n\geq 1} \rho_n f A_n g B_n$, which can be approximated by $h_D = fg + \sum_{n=1}^{D} \rho_n f A_n g B_n$

The proportion of geometric variability of H accounted for by H_D is:

$$P_D = \frac{\sum_{n=1}^{D} \rho_n^2}{\phi_t^2} \tag{11}$$

This proportion is employed in some methods of multivariate data analysis. For instance, it is used in correspondence analysis to measure the quality of the graphical representation of a contingency table w.r.t. the chi-square distance, see Greenacre

(1984), Cuadras and Cuadras (2006). As it has been shown above, see Section 3.1, this distance can be extended to r.v.'s.

In general, the rank of a cdf is greater than 2. We are interested in distributions of rank 2 as possible approximations to a distribution.

For example, for the following families of copulas (see Ali, Mikhail, and Haq 1978; Celebioglu 1997; Cuadras 2009, 2017; Nelsen 2006):

$$\begin{array}{ll} \text{Ali}-\text{Mikhail}-\text{Haq}: & uv/[1-\theta(1-u)(1-v)],\ -1\le\theta\le 1 \\ \text{Gumbel}-\text{Barnett}: & uv\exp(-\theta\ln u\ln v),\ 0\le\theta\le 1 \\ \text{Celebioglu}-\text{Cuadras}: & uv\exp[\theta(1-u)(1-v)],\ -1\le\theta\le 1 \\ \text{New family}: & uv\exp\{\sin[\theta(1-u)(1-v)]\},\ -1\le\theta\le 1 \end{array}$$

it turns out that the FGM family is the first order approximation in a Taylor's expansion. For example,

$$\begin{aligned} uv\exp(-\theta\ln u\ln v) &\simeq uv(1-\theta\ln u\ln v)+\cdots && (\text{expanding } e^{-x}) \\ &\simeq uv[1-\theta(1-u)(1-v)]+\cdots && (\text{expanding } \ln x) \end{aligned}$$

and Gumbel-Barnett (parameter θ) can be approximated by FGM (parameter $-\theta$).

However, the full rank of these four copulas is countable (i.e., $\aleph_0$), whereas the approximation of rank 2 gives an average proportion P_2 greater than 0.94, see Table 1. This approximation could be a signed measure rather than a proper cdf. However, also in correspondence analysis, the two-dimensional representation of a contingency table $\mathbf{N}$, could exhibit a table $\mathbf{N}'$ containing negative frequencies.

The two extensions of the FGM family next proposed, have rank 2, i.e., are two-dimensional in the above geometrical sense.

For a better understanding of some aspects of canonical correlation analysis, Hilbert space, Hilbert-Schmidt kernel, singular value decomposition, Mercer's theorem, signed measure and other concepts and results on functional analysis, used here and in the next section, see Hannan (1961), Ash (1965, 1972), Eagleson (1979), Buja (1990) and Letac (2008).

4. First extension

A singular value decomposition (SVD) of a kernel $K(x,y)$, with $x\in[a,b], y\in[c,d]$ is

$$K(x,y)=\sum_{i\ge 1}\lambda_i\xi_i(x)\bar{\xi}_i(y)$$

where $\lambda_1\ge\lambda_2\ge\cdots$ is the decreasing sequence of singular values and $\{\xi_i\},\{\bar{\xi}_i\}$ are the corresponding unitary and orthogonal functions. In particular, if K is symmetric, then $\{\xi_i\}=\{\bar{\xi}_i\}$ and the above SVD is an eigendecomposition (Mercer's theorem).

The FGM family $H_\theta=FG[1+\theta(1-F)(1-G)]$ can be interpreted as the singular value decomposition $K=H_\theta-FG=\theta F(1-F)G(1-G)$. Clearly, if $F=G$ and H is symmetric, we have an eigendecomposition, where $F(1-F)=G(1-G)$ is the only eigenfunction.

A generalization of the FGM family, obtained as a SVD, is $H_1=FG+\lambda_1\xi_1\bar{\xi}_1$. This distribution appeared in Farlie (1960) and has been rediscovered by Rodríguez-Lallena

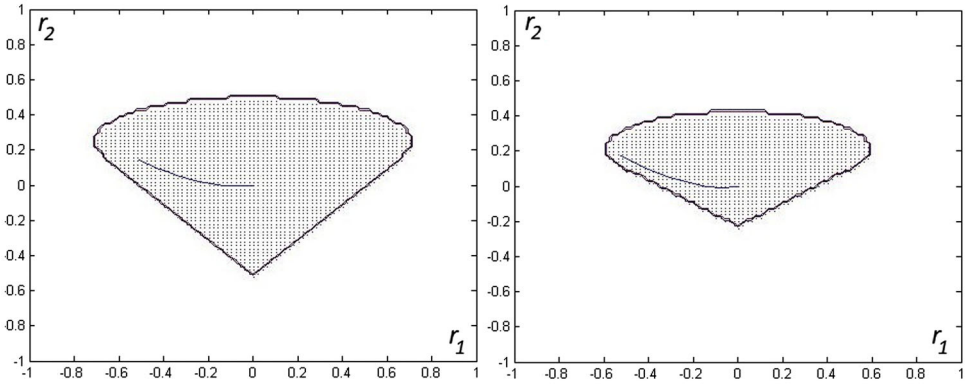

Figure 1. Region of the correlations (parameters) for which the density is positive, for the trigonometric cdf (left), see (21), and the polynomial cdf of degree two (right), see (17). The line indicates the possible correlations under the Gumbel-Barnett copula.

and Úbeda-Flores (2004). The family (1) is a particular case: $H-FG = \lambda(F-\Phi)(G-\Psi)$. In general H_1 gives rise to the following extension of rank 2.

Theorem 3. *Let* $K_1 = \lambda_1\xi_1\bar{\xi}_1$ *a SVD of* $K_1 = H_1-FG$, *where* $H_1 \in \mathcal{F}(F,G)$. *Then* $\xi_1(a) = \xi_1(b) = \bar{\xi}_1(c) = \bar{\xi}_1(d) = 0$, *and*

$$H_2 = FG + \lambda_1\xi_1\bar{\xi}_1 + \lambda_2\xi_2\bar{\xi}_2$$

is also a cdf of $\mathcal{F}(F,G)$ *for suitable values of* λ_1, λ_2, *where* $\xi_2 = \xi_1\xi_1'$, $\bar{\xi}_2 = \bar{\xi}_1\bar{\xi}_1'$. *If moreover* ξ_1, $\bar{\xi}_1$ *are increasing and* $\xi_1 < F$, $\bar{\xi}_1 < G$, *then* H_1 *belongs to the family* (1), *and* H_2 *is an extension of this family.*

Proof. $\bar{\xi}_1(y) \neq 0$ for some $y \in (c,d)$. Then $K_1(a,y) = 0 = \lambda_1\xi_1(a)\bar{\xi}_1(y)$, hence $\xi_1(a) = 0$, and similarly $\xi_1(b) = \bar{\xi}_1(c) = \bar{\xi}_1(d) = 0$. Integration by parts gives

$$\int_a^b \xi_1^2(x)\xi_1'(x)dx = \xi_1^3(x)|_a^b - 2\int_a^b \xi_1^2(x)\xi_1'(x)dx$$
$$= 0 - 2\int_a^b \xi_1^2(x)\xi_1'(x)dx$$

Thus $\int_a^b \xi_1^2(x)\xi_1'(x)dx = 0$ and $H_2-FG = \lambda_1\xi_1\bar{\xi}_1 + \lambda_2\xi_2\bar{\xi}_2$ is a SVD of $K_2 = H_2-FG$. As also $\xi_2(a) = \xi_2(b) = \bar{\xi}_2(c) = \bar{\xi}_2(d) = 0$, we have $H_2(a,y) = H_2(x,d) = 0$, and $H_2(b,d) = 1$. Thus H_2 may satisfy the necessary conditions for a cdf. Also, if ξ_1, $\bar{\xi}_1$ are increasing and $\xi_1 < F$, $\bar{\xi}_1 < G$, then $F-\xi_1$, $G-\bar{\xi}_1$ are cdf's and $H_1 = FG + \lambda_1[F-(F-\xi_1)][G-(G-\bar{\xi}_1)]$ is a member of the family (1). □

This construction applied to the family (1) gives

$$H_2 = FG + \lambda_1(F-\Phi)(G-\Psi) + \lambda_2(F-\Phi)(f-\varphi)(G-\Psi)(g-\psi) \tag{12}$$

However, in general, this family is not diagonal, in the sense that $(F-\Phi)$ and $(F-\Phi)(f-\varphi)$ are not canonical functions, as the correlation coefficient between both functions could not be zero.

Example 3. Consider $\xi_1(u) = \sin(\pi u)$, $\bar{\xi}_1(v) = \sin(\pi v)$. Then $\xi_2(u) = \xi_1(u)\xi_1'(u) = \pi \sin(\pi u)\cos(\pi u)$. The cdf is the copula

$$C = uv + \lambda_1 \sin(\pi u)\sin(\pi v) + \lambda_2 \sin(2\pi u)\sin(2\pi v)$$

The density after reparametrizing, is

$$c = 1 + \lambda \cos(\pi u)\cos(\pi v) + \mu \cos(2\pi u)\cos(2\pi v)$$

The range of the parameters (λ, μ) could not be expressed in closed form, see Figure 1 (left). If we fix the second parameter to μ_0, then the first parameter λ, depending on μ_0, should satisfy

$$\lambda \cos(\pi u)\cos(\pi v) \geq -1 - \mu_0 \cos(2\pi u)\cos(2\pi v)$$

For instance, if $\mu_0 = 1/2$ then $-\sqrt{2} \leq \lambda \leq \sqrt{2}$.

5. Second extension

Let us introduce some notations concerning the cdf's F, G, Φ and Ψ. We define

$$F_\Phi(x) = \int_a^x \Phi(t)dF(t), \quad F_{\Phi^2}(x) = \int_a^x \Phi^2(t)dF(t), \quad \Phi_{F^2}(x) = \int_a^x F^2(t)d\Phi(t)$$

and similarly $F_{F\Phi}$, $F_{F^2\Phi}$, G_Ψ, G_{Ψ^2}, Ψ_{G^2}. Integration by parts shows that

$$F(x)\Phi(x) = F_\Phi(x) + \Phi_F(x), \quad \int_a^b F(t)\Phi(t)dF(t) = \frac{1}{2} - \frac{1}{2}\Phi_{F^2}(b)$$

In particular $F_\Phi(b) + \Phi_F(b) = 1$.

We also write $\gamma = \int_a^b (F - \Phi)dF = 1/2 - F_\Phi(b)$, $\delta = \int_c^d (G - \Psi)dG = 1/2 - G_\Psi(d)$ and recall that $\alpha = \int_a^b (d\Phi/dF)^2 dF$, $\beta = \int_c^d (d\Psi/dG)^2 dG$.

5.1. Definition and properties

Another extension of (1) is the bivariate family

$$\begin{aligned} H = {} & FG + \lambda_1 (F - \Phi)(G - \Psi) \\ & + \lambda_2 \left[\frac{1}{2}F^2 + \left(F_\Phi(b) - \frac{1}{2}\right)F - F_\Phi\right]\left[\frac{1}{2}G^2 + \left(G_\Psi(d) - \frac{1}{2}\right)G - G_\Psi\right] \end{aligned} \tag{13}$$

where F, G, F_Φ, G_Ψ stand for $F(x)$, $G(y)$, $F_\Phi(x)$, $G_\Psi(y) = \int_c^y \Psi(t)dG(t)$. Then $F_\Phi(b)$ and $G_\Psi(b)$ are constant values.

The density (w.r.t. the Lebesgue measure) is

$$h = fg + \lambda_1 f\left(1 - \varphi f^{-1}\right) g\left(1 - \psi g^{-1}\right) + \lambda_2 f(F - \Phi - \gamma) g(G - \Psi - \delta) \tag{14}$$

This family reduces to the previous FGM generalizations (1) and (12) for $\lambda_2 = 0$.

Theorem 4. *The family (14) is diagonal of rank 2. The canonical correlations are*

$$\rho_1 = \lambda_1 \sqrt{(\alpha - 1)(\beta - 1)}, \quad \rho_2 = \lambda_2 \sqrt{st}$$

where α, β *are defined in Proposition 1,* $s = F_{\Phi^2}(b) + \Phi_{F^2}(b) + F_\Phi(b) - F_\Phi(b)^2 - \frac{11}{12}$, *and similarly* t.

Proof. Write (13) as

$$dH = dFdG + \lambda_1 a_1 dF b_1 dG + \lambda_2 a_2 dF b_2 dG$$

where $a_1 = 1 - d\Phi/dF$, $b_1 = 1 - d\Psi/dG$, $a_2 = (F - \Phi - \gamma)$ and $b_2 = (G - \Psi - \delta)$. It is readily proved that $E(a_1) = E(a_2) = 0$ and $E(b_1) = E(b_2) = 0$. Moreover

$$\int_a^b (1 - d\Phi/dF)(F - \Phi - \gamma) dF = 0$$

hence $E(a_1 a_2) = \int_a^b a_1 a_2 dF = 0$ and similarly $E(b_1 b_2) = 0$. Also

$$\int_a^b (F - \Phi - \gamma)^2 dF = 1/3 + F_{\Phi^2}(b) + \gamma^2 - \big(1 - \Phi_{F^2}(b)\big) - \gamma + 2\gamma F_\Phi(b)$$

$$\int_c^d (G - \Psi - \delta)^2 dG = 1/3 + G_{\Psi^2}(b) + \delta^2 - \big(1 - F_{\Phi^2}(b)\big) - \delta + 2\delta G_\Psi(b)$$

The other covariances are:

$$\begin{aligned}
\operatorname{cov}(a_1, b_1) &= \int_a^b \int_c^d a_1 b_1 (dH - dFdG) \\
&= \lambda_1 \int_a^b a_1^2 dF \int_c^d b_1^2 dG + \lambda_2 \int_a^b a_1 a_2 dF \int_c^d b_1 b_2 dG \\
&= \lambda_1 (\alpha - 1)(\beta - 1), \\
\operatorname{cov}(a_1, b_2) &= \lambda_1 \int_a^b a_1^2 dF \int_c^d b_1 b_2 dG + \lambda_2 \int_a^b a_1 a_2 dF \int_c^d b_2^2 dG \\
&= 0
\end{aligned}$$

Similarly $\operatorname{cov}(a_2, b_1) = 0$. Moreover

$$\begin{aligned}
\operatorname{cov}(a_2, b_2) &= 0 + \lambda_2 \int_a^b a_2^2 dF \int_c^d b_2^2 dG \\
&= \lambda_2 st
\end{aligned}$$

The variances are $E(a_1^2) = \alpha - 1, E(a_2^2) = F_{\Phi^2}(b) + \Phi_{F^2}(b) + F_\Phi(b) - F_\Phi(b)^2 - 11/12$, etc. □

5.2. Conjugate family and measures of association

Let us express the cdf (13) as

$$\begin{aligned}
dH = {} & dFdG + \lambda_1 (1 - d\Phi/dF) dF (1 - d\Psi/dG) dG \\
& + \lambda_2 \left[F - \Phi + F_\Phi(b) - \frac{1}{2} \right] dF \left[G - \Psi + G_\Psi(d) - \frac{1}{2} \right] dG
\end{aligned}$$

The conjugate family is

$$dH_* = d\Phi d\Psi + \lambda_1(1-dF/d\Phi)d\Phi(1-d\Psi/dG)d\Psi + \lambda_2\left[\Phi-F+\Phi_F(b)-\frac{1}{2}\right]d\Phi\left[\Psi-G+\Psi_G(d)-\frac{1}{2}\right]d\Psi$$

From $F_\Phi(b)-\frac{1}{2} = -\Phi_F(b)+\frac{1}{2}$, we have

$$dH-dFdG = dH_*-d\Phi d\Psi$$

Hence the dependence structure of H and H_* is quite similar. The covariance is the same and Spearman's rho for H is

$$\rho_S(H) = 12\lambda_1\left[\frac{1}{2}-F_\Phi(b)\right]\left[\frac{1}{2}-G_\Psi(d)\right] + \lambda_2\left[-\frac{1}{12}+F_\Phi(b)-I\right]\left[-\frac{1}{12}+\Phi_F(b)-J\right]$$

where $I = \int_a^b F_\Phi dF$ and $J = \int_a^b \Phi_F d\Phi$. We similarly obtain $\rho_S(H_*)$.

From $F_\Phi(b)+\Phi_F(b) = 1$ and

$$I = F_\Phi(b)-\int_a^b F\Phi dF, \quad J = \Phi_F(b)-\int_a^b F\Phi d\Phi$$

$\rho_S(H)$ and $\rho_S(H_*)$ have similar expressions, which may coincide in some particular cases.

Kendall's tau is given by

$$\begin{aligned}\tau(H) = {} & 8\lambda_1\left[\frac{1}{2}-F_\Phi(b)\right]\left[\frac{1}{2}-G_\Psi(d)\right] \\ & + \lambda_2\left[-\frac{13}{24}+\frac{1}{3}F_\Phi(b)+\frac{1}{2}F_{F^2\Phi}(b)\right]\left[-\frac{13}{24}+\frac{1}{3}G_\Psi(d)+\frac{1}{2}G_{G^2\Psi}(d)\right] \\ & + \lambda_2\left[\frac{1}{4}-\frac{1}{2}F_\Phi(b)+F_{F\Phi}(b)\right]\left[\frac{1}{4}-\frac{1}{2}G_\Psi(d)+G_{G\Psi}(d)\right] \\ & + \lambda_1\lambda_2\left[-\frac{5}{6}-\frac{1}{2}F_\Phi(b)-F_{F\Phi}(b)\right]\left[-\frac{5}{6}-\frac{1}{2}G_\Psi(d)-G_{G\Psi}(d)\right] \\ & + \lambda_1\lambda_2\left[\frac{1}{12}+F_{\Phi^2}(b)-2F_{F\Phi}(b)-F_\Phi(b)^2+F_\Phi(b)\right] \\ & \times\left[\frac{1}{12}+\Phi_{F^2}(d)-2\Phi_{\Phi F}(d)-\Phi_F(d)^2+\Phi_F(d)\right]\end{aligned}$$

6. Associated copulas

Here we find the copulas corresponding to the above families.

Lemma 4. *If F and Φ are two cdf's with support in $[a,b]$, then $Q = \Phi(F^{-1})$ is a cdf with support in $[0,1]$.*

Proof. Q is not decreasing and $Q(0) = \Phi[F^{-1}(0)] = \Phi(a) = 0$, $Q(1) = \Phi[F^{-1}(1)] = \Phi(b) = 1$. □

The copula corresponding to (12) is

$$C = uv + \lambda_1(u-Q)(v-R) + \lambda_2(u-Q)(1-q)(v-R)(1-r) \tag{15}$$

where $Q = \Phi(F^{-1}), q = Q', R = \Psi(G^{-1}), r = R'$. Both Q, R are cdf's with support in $[0,1]$.

Since $F_\Phi = \int_a^b \Phi dF = \int_0^1 Q(t)dt = 1-\mu_Q$, where μ_Q is the mean of the r.v. with cdf Q, and similarly μ_R, the copula corresponding to (13) is

$$\begin{aligned} C = {} & uv + \lambda_1(u-Q)(v-R) \\ & + \lambda_2\left[\left(\frac{1}{2}u^2 + \left(\frac{1}{2}-\mu_Q\right)u - \int_0^u Q(t)dt\right]\left[\left(\frac{1}{2}v^2 + \left(\frac{1}{2}-\mu_R\right)v - \int_0^v R(t)dt\right]\right. \end{aligned} \tag{16}$$

Proposition 6. *For $\Phi = F^2, \Psi = G^2$, the families (12) and (13) have the same copula*

$$C = uv + \lambda_1 u(1-u)v(1-v) + \bar{\lambda}_2(2u^3-3u^2+u)(2v^3-3v^2+v) \tag{17}$$

where $\bar{\lambda}_2 = \lambda_2/36$.

Proof. $Q = u^2$, $1-q = 1-2u$, $\mu_Q = 1/3$ and $\int_0^u Q(t)dt = u^3/3$. Then (15) and (16) reduce to (17). □

Another interesting particular case is $\Phi = F^k$, $\Psi = G^k$. The copula corresponding to (13) is

$$\begin{aligned} C = {} & uv + \lambda_1(u-u^k)(v-v^k) \\ & + \lambda_2\left[\frac{1}{2}u^2 + (k-1)/(2(k+1))u - u^{k+1}/(k+1)\right] \times (\text{similar term in } v) \end{aligned} \tag{18}$$

and Spearman's correlation is

$$\rho_S = 3\lambda_1\left(\frac{k-1}{k+1}\right)^2 + \frac{\lambda_2}{12}\left[\frac{6k-(k+1)(k+2)}{(k+1)(k+2)}\right]^2$$

It is difficult to find analytically the region of the parameters for which (18) is a copula, see Figure 1 (right). But if we fix the first parameter, a closed form is possible.

Example 4. Consider the family (17) and fix $\lambda_1 = \lambda_0$. Then $\bar{\lambda}_2$ should satisfy

$$\bar{\lambda}_2(6u^2-6u+1)(6v^2-6v+1) \geq -1-\lambda_0(2u-1)(2v-1)$$

Thus, if $\lambda_1 = 1/2$ then $-1/2 \leq \bar{\lambda}_2 \leq 2$.

7. Relating two cdf's

Let $H_a, H_b \in \mathcal{F}(F,G)$ two cdf's with the same marginals. In this section we present some ways of measuring the proximity between H_a and H_b.

Definition 4. The chi-square distance between H_a and H_b is

$$\delta^2(H_a, H_b) = \int_a^b \int_c^d \left(\frac{dH_a - dH_b}{dFdG}\right)^2 dFdG$$

Definition 5. Pearson's affinity between H_a and H_b is

$$\phi(H_a, H_b) = \int_a^b \int_c^d \left(\frac{dH_a}{dFdG} \times \frac{dH_b}{dFdG} \right) dFdG$$

It is clear that $\delta^2(H, FG)$ is the Pearson contingency coefficient ϕ_t^2, see (4), It can be proved that, among all distributions H with fixed Spearman's correlation $\lceil \rho_0 \rceil < 1/3$, the closest cdf H to the independence FG, in the sense that $\delta^2(H, FG)$ is minimized, is the FGM cdf (Nelsen 1994).

The relation between distance and affinity is

$$\delta^2(H_a, H_b) = \phi(H_a, H_a) + \phi(H_b, H_b) - 2\phi(H_a, H_b)$$

Since $\phi(H_a, H_b)$ is an inner product, Cauchy-Schwarz inequality $\phi(H_a, H_b)^2 \leq \phi(H_a, H_a)\phi(H_b, H_b)$ holds, which suggests the association coefficient

$$A(H_a, H_b) = \frac{\phi(H_a, H_b)^2}{\phi(H_a, H_a)\phi(H_b, H_b)}$$

We have $0 \leq A(H_a, H_b) \leq 1$, and $A(H_a, H_b) = 1$ if $H_a = H_b$.

Suppose that the following SVD exists:

$$\frac{dH_a}{dFdG} - \frac{dH_b}{dFdG} = \sum_{n \geq 1} \lambda_n a_n b_n$$

Then $dH_a - dH_b = \sum_{n \geq 1} \lambda_n a_n b_n dFdG$. It is next proved that $\int_a^b a_i a_j dF = \int_c^d b_i b_j dG = \delta_{ij}$ (Kronecker's delta).

Theorem 5. *If (X, Y) $\sim H_a$ and $(X^*, Y^*) \sim H_b$ then $E[a_n(X)] = E[b_n(Y)] = 0$ and*

$$\operatorname{cov}\big(a_m(X), b_n(Y)\big) - \operatorname{cov}\big(a_m(X^*), b_n(Y^*)\big) = \begin{cases} \lambda_n & \text{if } m = n \\ 0 & \text{if } m \neq n \end{cases}$$

Proof. $E[a_n(X)] - E[a_n(X^*)] = \int_a^b \int_c^d a_n dH_a - \int_a^b \int_c^d a_n dH_b = \int_a^b a_n dF - \int_a^b a_n dF = 0$. Therefore

$$\int_a^b \int_c^d a_n (dH_a - dH_b) = \sum_{k \geq 1} \lambda_k \int_a^b a_k a_n dF \int_c^d b_n dG$$

$$= \lambda_n \int_c^d b_n dG$$

hence $\int_c^d b_n dG = E[b_n(Y)] = 0$. Similarly $E[a_n(X)] = 0$. The difference of covariances is

$$\int_a^b \int_c^d a_m b_n (dH_a - dH_b) = \sum_{k \geq 1} \lambda_k \int_a^b a_m a_k dF \int_c^d b_n b_k dG$$

where $\{a_k\}, \{b_k\}$ are orthogonal on F, G, respectively. □

8. Reducing a bivariate cdf to a simpler one

Let $H_a \in \mathcal{F}(F, G)$ and suppose that the diagonal expansion $dH_a = dFdG + \sum_{n \geq 1} \rho_n a_n dF b_n dG$ exists, where a_n, b_n are unitary canonical functions. Let $H_t \in \mathcal{F}(F, G)$

the "true" cdf of two observable r.v.'s (X, Y). Given a positive integer k, we are interested in approximating H_t by means of a finite linear combination of canonical functions obtained from H_a:

$$dH_t \simeq dFdG + \sum_{i=1}^{k} \lambda_i a_i dF b_i dG$$

In a more precise way, we seek the approximation

$$\frac{dH_t}{dFdG} \simeq 1 + \sum_{i=1}^{k} \lambda_i a_i b_i$$

where $\lambda_1, ..., \lambda_k$ are real coefficients such that

$$\int_a^b \int_c^d \left(\frac{dH_t - dFdG}{dFdG} - \sum_{i=1}^{k} \lambda_i a_i b_i \right)^2 dFdG \tag{19}$$

is minimized. If the densities h_t, f, g exist, then h_t is approximated by $\widehat{h}_t = fg(1 + \sum_{i=1}^{k} \lambda_i a_i b_i)$.

Theorem 6. *Suppose (X, Y) ~H_t. The coefficients minimizing (19) are $\lambda_i = r_i$, where*

$$r_i = cor\big(a_i(X), b_i(Y)\big), \quad i = 1, ..., k$$

Then $\sum_{i=1}^{k} r_i^2 \leq \phi_t^2$, where ϕ_t^2 is the Pearson contingency coefficient of H_t, and the minimum is

$$\phi_t^2 - \sum_{i=1}^{k} r_i^2 \tag{20}$$

Proof. Write $z = (dH_t - dFdG)/(dFdG)$. Since $\int_a^b \int_c^d a_i b_i (dH_t - dFdG) = r_i$ is the correlation between a_i, b_i, we have

$$\begin{aligned}
\int_a^b \int_c^d \left(z - \sum_{i=1}^{k} \lambda_i a_i b_i \right)^2 dFdG &= \phi_t^2 + \int_a^b \int_c^d \sum_{i=1}^{k} \lambda_i^2 a_i^2 b_i^2 dFdG \\
&\quad - 2 \int_a^b \int_c^d \sum_{i=1}^{k} \lambda_i a_i b_i)(dH_t - dFdG) \\
&\quad + \sum_{i \neq j=1}^{k} \lambda_i \lambda_j \int_a^b a_i dF \int_c^d b_j dG \\
&= \phi_t^2 + \sum_{i=1}^{k} \lambda_i^2 - 2 \sum_{i=1}^{k} \lambda_i r_i
\end{aligned}$$

Taking the partial derivative w.r.t. λ_i, on the right hand side of this equation, and equaling to zero, we obtain $\lambda_i = r_i$, $i = 1, ..., k$ and (20) is the minimum. The maximal property of the canonical correlations shows that $\sum_{i=1}^{k} r_i^2 \leq \sum_{i \geq 1} \rho_n^2$. □

Table 2. Estimated correlations and fit for the Gumbel-Barnett copula using trigonometic and polynomial functions, see (21) and (17). η measures the maximum difference between the true and the fitted copula.

	Trigonometric			Polynomial		
θ	r_1	r_2	η	r_1	r_2	η
0.25	−0.1589	0.0025	0.0082	−0.1676	0.0057	0.0068
0.5	−0.2934	0.0387	0.0109	−0.3050	0.0520	0.0087
0.75	−0.4091	0.0882	0.0107	−0.4222	0.1106	0.0081
1	−0.5100	0.1411	0.0095	−0.5238	0.1698	0.0060

Table 3. Estimated correlations with the polynomial model, see (17), and fit for the AMH copula. η measures the maximum difference between the true and the fitted copula.

θ	r_1	r_2	η
1	0.4784	0.2337	0.0261
0.5	0.1924	0.0223	0.0032
−0.5	−0.1489	0.0080	0.0017
−1	−0.2711	0.0216	0.0055

Note that, using the diagonal expansion and canonical correlations of H_t, we always can find the above approximation. This result is useful if we know H_a. Also note that r_i is the correlation between the canonical variables a_i, b_i of H_a, but this correlation is computed w.r.t. the "true" cdf H_t.

We can choose k such that $\sum_{i=1}^{k} r_i^2$ is close to ϕ_t^2. In the examples below, $k=2$ is a good choice.

This approximation of a cdf for a simpler one is as follows:

1. $H_t \in \mathcal{F}(F, G)$ is the true or real cdf but unknown.
2. Take suitable unitary functions a_i, b_i, on F, G, and parameters ρ_i, $i = 1, ..., k$, such that $dH_a = dFdG + \sum \rho_i a_i b_i dFdG$ gives $H_a \in \mathcal{F}(F, G)$.
3. Compute the correlation coefficients $r_i = \text{cor}(a_i, b_i)$ w.r.t. H_t.
4. Construct the cdf $\widehat{H}_t$ such that $d\widehat{H}_t = dFdG + \sum_{i=1}^{k} r_i a_i b_i dFdG$.

The canonical correlations of H_t are not necessary and the r_i should be obtained by statistical estimation. In order to choose canonical functions, we may consider the univariate expansions of the marginal variables and take the first principal dimensions. See Cuadras and Fortiana (1995), Cuadras and Lahlou (2000), Cuadras (2014).

9. Examples

9.1. Gumbel-Barnett

Suppose that the true cdf of (U, V) is the Gumbel-Barnett family of copulas (see Hutchinson and Lai 1991; Nelsen 2006):

$$C_t = uv \exp(-\theta \ln u \ln v), \quad 0 \le \theta \le 1$$

The marginals are (0, 1) uniform.

A system of orthogonal principal components of U is $\{1-\cos(n\pi U))\}$ (Cuadras and Fortiana 1995). Centering and normalizing the first two components, we get $a_1 =$

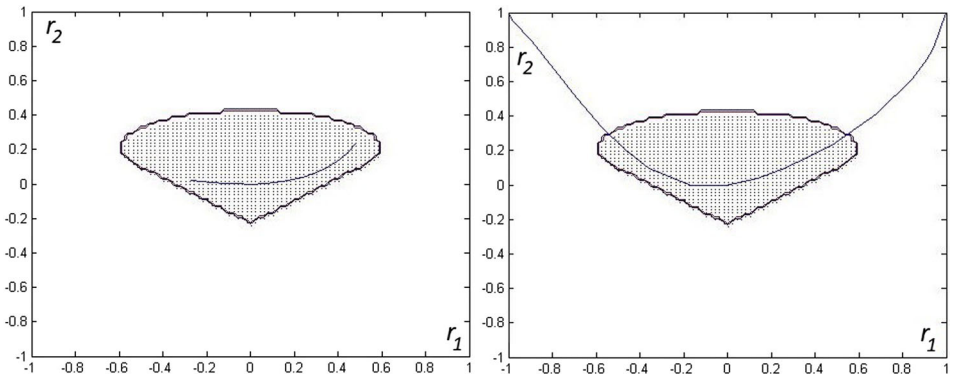

Figure 2. Region of the correlations (parameters) for which the density is positive, for the polynomial cdf of degree two (left), see (17), and the polynomial cdf of degree three (right), see (18) with $k = 3$. See also (22). The lines indicate the possible correlations under the AMH copula (left) and Clayton-Oakes copula (right). For the AMH copula the line is inside the admissible region. For the Clayton-Oakes copula the left and right parts of the line are outside the admissible region.

$\sqrt{2}\cos(\pi U)$, $a_2 = \sqrt{2}\cos(2\pi U)$, which play the role of canonical functions. This suggests the copula with density:

$$c = 1 + \rho_1 2\cos(\pi u)\cos(\pi v) + \rho_2 2\cos(2\pi u)\cos(2\pi v)$$

The canonical correlations, interpreted as parameters (possibly negative), should belong to the region $\mathcal{R} = \{(\rho_1, \rho_2) | c \geq 0\}$, see Figure 1. The copula (already introduced in Example 3), is

$$C_a = uv + \rho_1\left(2/\pi^2\right)\sin(\pi u)\sin(\pi v) + \rho_2\left[1/(2\pi^2)\right]\sin(2\pi u)\sin(2\pi v) \tag{21}$$

Next, using (7), we compute the covariance (or correlation) between the normalized variables $\sqrt{2}\cos(\pi U)$ and $\sqrt{2}\cos(\pi V)$, i.e.

$$\begin{aligned} r_1 &= \int_0^1\int_0^1 (C_t - uv)d\left[\sqrt{2}\cos(\pi u)\right]d\left[\sqrt{2}\cos(\pi v)\right] \\ &= 2\pi^2\int_0^1\int_0^1 \left[uv\exp(-\theta\ln u\ln v) - uv\right]\sin(\pi u)\sin(\pi v)dudv \end{aligned}$$

We similarly compute r_2. Then C_t can be approximated by a particular version of C_a :

$$\widehat{C}_t = uv + r_1\left(2/\pi^2\right)\sin(\pi u)\sin(\pi v) + r_2\left[1/(2\pi^2)\right]\sin(2\pi u)\sin(2\pi v)$$

Of course, we may use other functions. Let us take the approximation of Gumbel-Barnett to the generalized FGM copula (17). The density for this copula is

$$c = 1 + \frac{\lambda_1}{3}\sqrt{3}(1-2u)\sqrt{3}(1-2v) + \frac{\lambda_2}{5}\sqrt{5}(6u^2-6u+1)\sqrt{5}(6v^2-6v+1)$$

The canonical functions are $a_1 = \sqrt{3}(1-2u)$, $a_2 = \sqrt{5}(6u^2-6u+1)$ and similarly b_1, b_2. To be sure that c is a density, the canonical correlations must belong to the region $\mathcal{R} = \{(\rho_1, \rho_2) | c \geq 0\}$, see Figure 1.

Table 4. Sample sizes of six simulations A, B, C, D, E, F, estimation of the correlations and fit for the Gaussian copula with parameter rho. The last line S reports the fit to a stock data set. η measures the maximum difference between the true and the fitted copula. However η can not be computed for the simulations E, F, because the true copulas are out of the admissible region, see Figure 3 (right).

	n	ρ	r_1	r_2	η
A	100	0.2	0.2710	0.0133	0.0229
B	80	0.3	0.3425	0.1936	0.0209
C	120	−0.2	−0.1337	0.0425	0.0092
D	200	−0.5	−0.5546	0.1518	0.0090
E	120	0.8	0.7756	0.5695	–
F	180	−0.9	−0.8950	0.7161	–
S	100	0.3684	0.3124	0.1047	0.0090

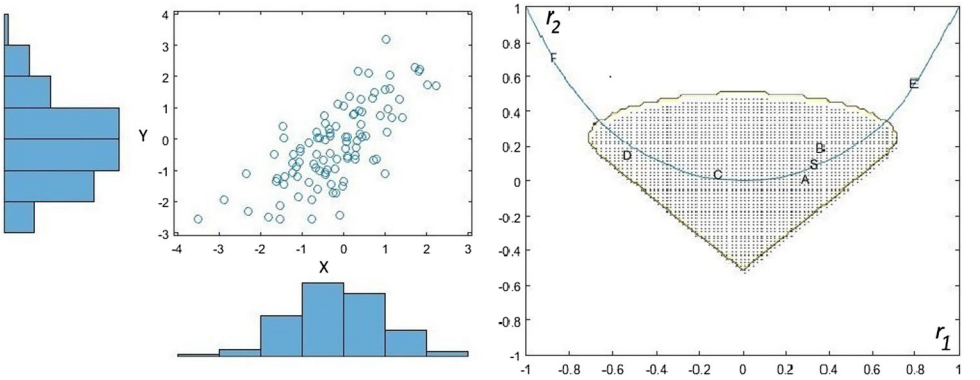

Figure 3. Scatterplot of the stocks data (left), region for the trigonometric copula (right), see (21), and line giving the range of the correlations (r_1, r_2) for the Gaussian copula and the correlations obtained with six simulations (right). The points A, B, C, D indicate correct approximation. The points E, F indicate that the correlations are out of the admissible region. The point S, clearly on the line, corresponds to the stocks data set.

We compute

$$r_1 = \int_0^1 \int_0^1 (C_t - uv) d\left[\sqrt{3}(1-2u)\right] d\left[\sqrt{3}(1-2v)\right]$$
$$= 12 \int_0^1 \int_0^1 \left[uv \exp\left(-\theta \ln u \ln v\right) - uv\right] du dv$$

and similarly r_2. Then we consider the approximation

$$\widehat{C}_t = uv + r_1 3u(1-u)v(1-v) + r_2 5(2u^3 - 3u^2 + u)(2v^3 - 3v^2 + v)$$

The results are reported in the Table 2, where the measure of fit is $\eta = \max|C_t(u,v) - \widehat{C}_t(u,v)|$.

Figure 1 shows the set of points (r_1, r_2) for which the density is positive for the two approximations considered here. The percentage of the areas of the regions in $[-1,1]^2$ are 21% (trigonometric) and 12% (polynomial). Thus, fitting the true cdf to a trigonometric one may be easier.

Nevertheless, with the trigonometric model and under Gumbel-Barnett cdf, we have $(r_1, r_2) \in \mathbf{A}$, where $\mathbf{A}$ is a curve inside the region of the admissible correlations for the

Table 5. Estimated correlations and fit for a new copula (left), see (23), and an Archimedean copula (right), see (24). η measures the maximum difference between the true and the fitted copula.

θ	r_1	r_2	η	θ	$\bar{r}_1$	$\bar{r}_2$	η
0	0	0	0	1	0.3822	0.2738	0.0351
0.25	0.0792	0.0004	0.0043	1.5	0.7006	0.5673	0.0303
0.5	0.1616	0.0012	0.0096	2	0.8258	0.7200	0.0318
1	0.3369	0.0034	0.0226	4	0.9552	0.9168	0.0370

two functions defining C_a. Thus, this approximation always works with the Gumbel-Barnett cdf. Similarly, with the polynomial model, the fit also works for any value of the parameter. See Figure 1.

The Pearson contingency coefficient for $\theta = 1$ is $\phi_t^2 = 0.3221$. Then, combining (11) and (20), we can see that the polynomial cdf $\widehat{C}_t$ accounts for by the 94% of the true cdf $C_t = uv\exp(-\ln u \ln v)$.

9.2. Ali-Mikhail-Haq

Here we study the approximation to the AMH copula also using the generalized FGM copula, but considering the exact computation of the estimated correlations. We should calculate the correlations $r_1 = \text{cor}(U, V)$ and $r_2 = \text{cor}(U^2 - U, V^2 - V)$. From (7) we have:

$$r_1 = 12\int_0^1\int_0^1 C(u,v)dudv - 3$$
$$r_2 = 180\int_0^1\int_0^1 C(u,v)(4uv - 2u - 2v + 1)dudv - 5$$

Taylor's expansion of the AMH copula (the "true" copula C_t) and using the beta function, we can obtain exact expressions for r_1, r_2, see Cuadras and Diaz (2012).

However, as finding r_2 is quite difficult, we propose a numerical alternative computing

$$r_2 = \int_0^1\int_0^1 (C_t - uv)d\left[\sqrt{5}(6u^2 - 6u + 1)\right]d\left[\sqrt{5}(6v^2 - 6v + 1)\right]$$
$$= 180\int_0^1\int_0^1 \left\{uv/[1 - \theta(1-u)(1-v)] - uv\right\}(2u-1)(2v-1)dudv$$

Thus, the copula AMH can be approximated by

$$C_2 = uv + r_1 3u(1-u)v(1-v) + r_2 5(2u^3 - 3u^2 + u)(2v^3 - 3v^2 + v)$$

Again, a measure of fit is $\eta = \max|C_t(u,v) - C_2(u,v)|$, where $0 < u, v < 1$. Table 3 reports a numerical illustration, showing that the fit is quite good.

For $\theta = 0.5$, the Pearson contingency coefficient is $\phi_t^2 = 0.0386$. Again, combining (11) and (20), we can see that $\widehat{C}_t$ accounts for by the 97% of the "true" cdf $C_t = uv/[1 - 0.5(1-u)(1-v)]$.

Figure 2 (left) shows the set of points (r_1, r_2) for which the density is positive for the polynomial approximation considered here. The percentage of the area of the region in $[-1, 1]^2$ is only 12%. However, if the underlying cdf is AMH, then $(r_1, r_2) \in \mathbf{B}$, where $\mathbf{B}$

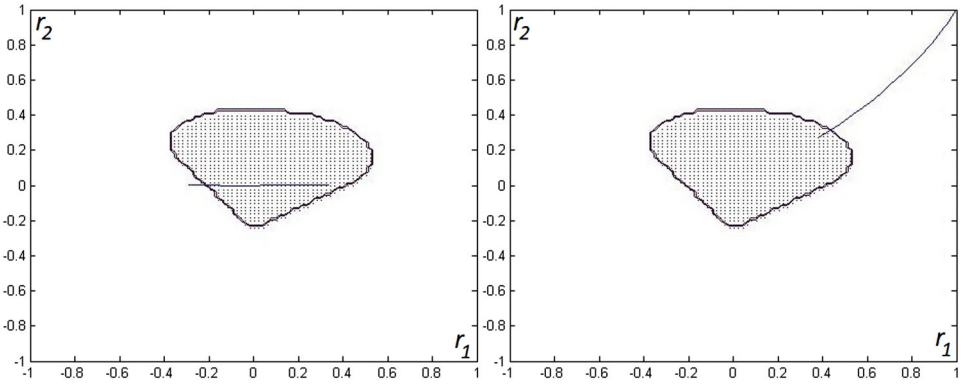

Figure 4. Region of the correlations (parameters) for which the density is positive, for the polynomial cdf of degree three, see (22). The lines indicate the possible correlations under the new copula (left), see (23), and a specific Archimedean copula (right), see (24). For this copula most of the line is outside the admissible region.

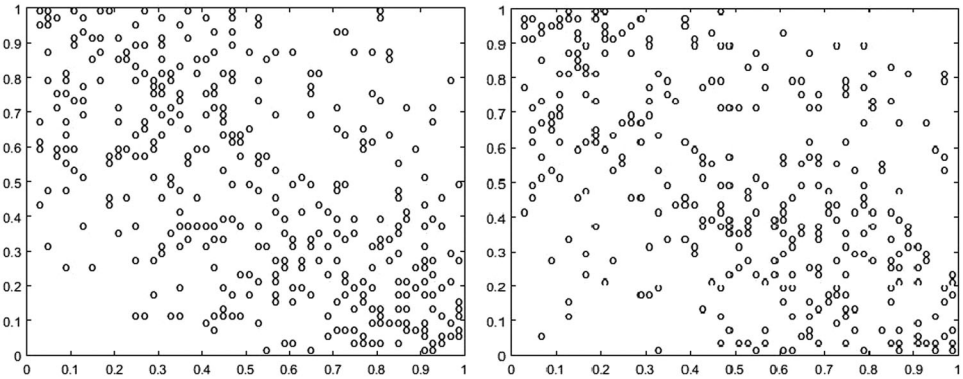

Figure 5. Scatterplot for Gumbel-Barnett copula, $\theta = 1$ (left), and the polynomial approximation of degree two, see (17), with $r_1 = -0.5238$, $r_2 = 0.1698$ (right).

is a curve inside this set of admissible correlations. Thus, this approximation will work for any value of the parameter θ, if the cdf is truly AMH.

9.3. Bivariate normal

The bivariate normal distribution is a probability model frequently used in the applications. If (X, Y) follows this distribution with correlation coefficient ρ, the uniform transformation $U = F(X)$, $V = G(Y)$ provides (U, V) with cdf the Gaussian copula. We simulate this copula for several values of ρ and the sample size n. Then we study the fit to the trigonometric copula (21). Table 4 reports the correlations $r_1 = \text{cor}(\cos \pi U, \cos \pi V)$, $r_2 = \text{cor}(\cos 2\pi U, \cos 2\pi V)$, and the fit η measuring the maximum difference between the cdf's of the Gaussian copula and the trigonometric copula (21). Figure 3 shows that the simulations A, B, C, D can be approximated by the trigonometric copula. However, E and F, out of the region, reveals that the fit is not possible for both data sets.

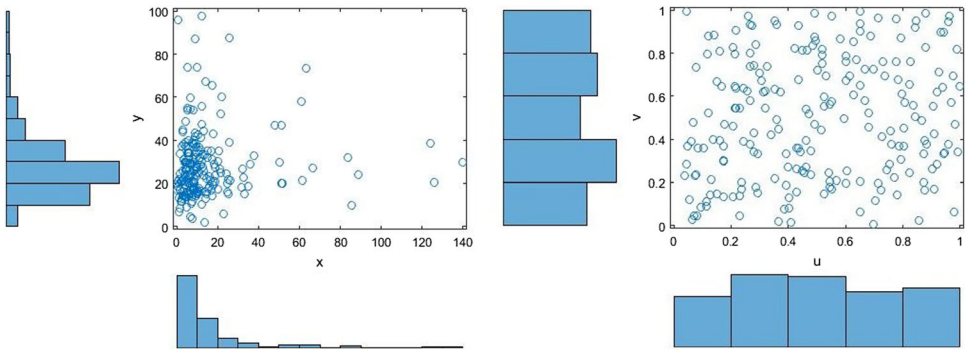

Figure 6. Scatterplots of the initial cancer data (left), and the same data transformed to have uniform marginals (right).

Next, we fit the stocks data provided by the Matlab package, available with the sentence load stockreturns. During the course of $n = 100$ weeks, the change in stock prices of 10 companies has been recorded. The first four companies are classified as primarily technology. We choose the first and third companies as the variables X and Y, The fit to the bivariate normal is good, see Figure 3 (left). Then the uniform transformation of the data may follow the Gaussian copula. The trigonometric approximation, see Table 4, last line, is quite good. This stocks data set is summarized in the point S, see Figure 3 (right). S is just on the curved line. This line contains the possible values (r_1, r_2) for the bivariate normal distribution.

9.4. Clayton-Oakes

The Clayton-Oakes family of copulas (Nelsen 2006) is defined by

$$C = \left[\max\left(u^{-\theta} + v^{-\theta} - 1, 0\right)\right]^{-1/\theta}, -1 \leq \theta < \infty$$

The computations of the correlation r_1 and r_2 have been obtained numerically. The fit works for θ between –0.5 and 1. However, for other θ the results can provide polynomial approximations which are not copulas, i.e., the density is negative for some values of $0 \leq u, v \leq 1$. See Figure 2 (right). Then we should take $(r_1^*, r_2^*) \in \mathcal{R}$ with smaller Euclidean distance to (r_1, r_2). The fit is acceptably good, especially for intermediate values of the parameter, see Cuadras and Diaz (2012). See the life example below.

9.5. Other distributions

Here we consider the second FGM extension with $\Phi = F^3$, $\Psi = G^3$. The associated copula is

$$C_a = uv + \lambda_1(u - u^3)(v - v^3) + \lambda_2(u^4 - 2u^2 + u)(v^4 - 2v^2 + v) \tag{22}$$

The canonical functions are $a_1 = \sqrt{5/4}(3u^2 - 1)$, $a_2 = 4\sqrt{105/23}(4u^3 - 4u + 1)$, and similarly b_1, b_2. We wish to approximate the (possibly new) copula (Cuadras 2017)

$$uv \exp\left\{\sin[\theta(1-u)(1-v)]\right\}, \quad -1 \leq \theta \leq 1 \tag{23}$$

to C_a. We also approximate the Archimedean copula, Equation (4.2.12) in Nelsen (2006),

$$\left\{1+\left[(u^{-1}-1)^{\theta}+(v^{-1}-1)^{\theta}\right]^{1/\theta}\right\}^{-1},\ \theta\geq 1 \tag{24}$$

to C_a.

We compute $r_1=\text{cor}(a_1,b_1)$, $r_2=\text{cor}(a_2,b_2)$ w.r.t. these "true" copulas. For instance,

$$r_2=\int_0^1\int_0^1 (C_t-uv)d\left[4\sqrt{105/23}(4u^3-4u+1)\right]d\left[4\sqrt{105/23}(4v^3-4v+1)\right]$$

Some results are reported in Table 5.

The fit for the first copula is quite good. However, for this copula and the values of θ close to -1, and specially for the Archimedean copula, the approximation $\widehat{C}_t$ using a polynomial of degree 3 may provide distributions with negative mass in a region of the support, see Figure 4. Indeed, the fit using polynomials works efficiently when the support of the true distribution is the full square $[a,b]\times[c,d]$. The support of this Archimedean copula is a subset of $[0,1]^2$, see Figure 4.6 in Nelsen (2006). As in the Clayton-Oakes cdf (see the previous section), this problem can be overcome by taking a proper copula near to $\widehat{C}_t$. See Cuadras and Diaz (2012).

Finally, we perform some simulations. Figure 5 shows a sample of points (u, v) simulated from a Gumbel-Barnett copula with $\theta=1$ and the corresponding polynomial approximation. The scatterplots are quite similar.

9.6. Example with life data

We illustrate the fit to the trigonometric copula (21) with the data set involving patients of the cancer of the prostate studied in Hosner and Lemeshow (2000). This data set is available on line: ftp://ftp.wiley.com/public/sci_tech_med/logistic/

We consider the variables Prostatic Specific Antigen (PSA in mg/ml),and Tumor Volume (TV in cm^3), labeled X and Y, respectively. There are 380 patients, but we discard the subjects with zero TV and fit the copula to the data of the remaining 213 cases. Then the copula related to (X, Y) fits quite well to a Clayton-Oakes copula. However, we suppose the "true" copula unknown and fit the data to the trigonometric copula (21).

We use the Matlab function *ksdensity* for transforming X, Y into U, V with (0, 1) uniform distribution. Next, we compute the sample correlations $r_1=\text{cor}(\cos\pi U,\cos\pi V)$, $r_2=\text{cor}(\cos 2\pi U,\cos 2\pi V)$ and the fit measure $\eta=\max|C_e-\widehat{C}_t|$, where C_e is the empirical copula (see Nelsen, 2006) and $\widehat{C}_t$ is given in (21). Note that $\sqrt{2}\cos(\pi U)$, $\sqrt{2}\cos(2\pi U)$, play the role of canonical functions for this copula.

For the cancer data we find

$$r_1=0.1249,\ \ r_2=-0.0203,\ \ \eta=0.0392$$

The average of $|C_e-\widehat{C}_t|$ is 0.0076. The fit is acceptably good. Clearly (r_1,r_2) belongs to the admissible region for the trigonometric model, see Figure 1 (left). Therefore, $\widehat{C}_t$ is a proper copula.

Figure 6 (left) shows the scatterplo of this life data. Figure 6 (right) shows the scatterplot of this bivariate data transformed to have uniform marginals.

Acknowledgments

The authors are indebted to an associate editor and two anonymous referees for useful suggestions and comments.

Funding

Work supported in part by grants MTM-2015-65016-C2-2-R (MINECO/FEDER), RTI2018-095518-B-C22 (MCIU) and FONDECYT 1160429.

References

Ali, M. M., N. N. Mikhail, and M. S. Haq. 1978. A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis* 8 (3):405–12. doi:10.1016/0047-259X(78)90063-5.

Amblard, C., and S. Girard. 2002. Symmetry and dependence properties within a semiparametric family of bivariate copulas. *Nonparametric Statistics* 14 (6):715–27. doi:10.1080/10485250215322.

Amblard, C., and S. Girard. 2009. A new extension of bivariate FGM copulas. *Metrika* 70 (1): 1–17. doi:10.1007/s00184-008-0174-7.

Ash, R. B. 1965. *Information theory*. New York: Interscience Pub., John Wiley.

Ash, R. B. 1972. *Real analysis and probability*. New York: Academic Press.

Balakrishnan, N., and C. D. Lai. 2009. *Bivariate continuous distributions*. New York: Springer.

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *The Annals of Statistics* 18 (3):1032–69. doi:10.1214/aos/1176347739.

Celebioglu, S. 1997. A way for generating comprehensive copulas. *Journal of the Institute of Science and Technology of Gazi University* 10:57–61.

Csiszár, I. 1975. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability* 3:146–58.

Cuadras, C. M. 2002. On the covariance between functions. *Journal of Multivariate Analysis* 81 (1):19–27. doi:10.1006/jmva.2001.2000.

Cuadras, C. M. 2006. The importance of being the upper bound in the bivariate family. *SORT* 30:55–84.

Cuadras, C. M. 2009. Constructing copula functions with weighted geometric means. *Journal of Statistical Planning and Inference* 139 (11):3766–72. doi:10.1016/j.jspi.2009.05.016.

Cuadras, C. M. 2014. Nonlinear principal and canonical directions from continuous extensions of multidimensional scaling. *Open Journal of Statistics* 4:132–49.

Cuadras, C. M. 2015. Contributions to the diagonal expansion of a bivariate copula with continuous extensions. *Journal of Multivariate Analysis* 139:28–44. doi:10.1016/j.jmva.2015.02.015.

Cuadras, C. M. 2016. Corrigendum to "Contributions to the diagonal expansion of a bivariate copula with continuous extensions" [J. of Multivariate Analysis, 139 (2015) 28-44]. *Journal of Multivariate Analysis* 147:315. doi:10.1016/j.jmva.2016.02.003.

Cuadras, C. M. 2017. A note on canonical copulas. University of Barcelona. hal-01598825v2.

Cuadras, C. M., and J. Augé. 1981. A continuous general multivariate distribution and its properties. *Communications in Statistics-Theory and Methods* A10:339–53. doi:10.1080/03610928108828042.

Cuadras, C. M., and D. Cuadras. 2006. A parametric approach to correspondence analysis. *Linear Algebra and Its Applications* 417 (1):64–74. doi:10.1016/j.laa.2005.10.029.

Cuadras, C. M., and D. Cuadras. 2008. Eigenanalysis on a bivariate covariance kernel. *Journal of MultivariateAnalysis* 99 (10):2497–507. doi:10.1016/j.jmva.2008.02.039.

Cuadras, C. M., and W. Diaz. 2012. Another generalization of the bivariate FGM distribution with two-dimensional extensions. *Acta et Commentationes Universitatis Tartuensis de Mathematica* 16 (1):3–12.

Cuadras, C. M., and J. Fortiana. 1995. A continuous metric scaling solution for a random variable. *Journal of Multivariate Analysis* 52 (1):1–14. doi:10.1006/jmva.1995.1001.

Cuadras, C. M., J. Fortiana, and M. J. Greenacre. 2000. Continuous extensions of matrix formulations in Correspondence Analysis, with applications to the FGM family of distributions. In: *Innovations in multivariate statistical analysis*, eds. R.D.H. Heijmans, D.S.G. Pollock and A. Satorra, 101–116. Kluwer Ac. Publ., Dordrecht.

Cuadras, C. M., and Y. Lahlou. 2000. Some orthogonal expansions for the logistic distribution. *Communications in Statistics-Theory and Methods* 29 (12):2643–63. doi:10.1080/03610920008832629.

Diaz, W., and C. M. Cuadras. 2017. On a multivariate generalization of the covariance. *Communications in Statistics-Theory and Methods* 46 (9):4660–9. doi:10.1080/03610926.2015.1056368.

Drouet-Mari, D., and S. Kotz. 2001. *Correlation and dependence*. London: Imperial College Press.

Eagleson, G. K. 1979. Orthogonal expansions and U-statistics. *Australian Journal of Statistics* 21 (3):221–37. doi:10.1111/j.1467-842X.1979.tb01141.x.

Farlie, D. J. G. 1960. The performance of some correlation coefficients for a general bivariate distribution. *Biometrika* 47 (3–4):307–23. doi:10.1093/biomet/47.3-4.307.

Greenacre, M. J. 1984. *Theory and applications of correspondence analysis*. London: Academic Press.

Hannan, E. J. 1961. The general theory of canonical correlation and its relation to functional analysis. *Journal of the Australian Mathematical Society* 2 (2):229–42. doi:10.1017/S1446788700026707.

Hosner, D. W., and S. Lemeshow. 2000. *Applied logistic regression*. 2nd ed. New York: John Wiley.

Huang, J. S., and S. Kotz. 1999. Modifications of the Farlie-Gumbel-Morgenstern distributions. A tough hill to climb. *Metrika* 49 (2):135–45. doi:10.1007/s001840050030.

Hutchinson, T. P., and C. D. Lai. 1991. *The engineering statistician's guide to continuous bivariate distributions*. Adelaide: Rumsby Scientific Pub.

Joe, H. 1997. *Multivariate models and dependence concepts*. London: Chapman and Hall.

Kotz, S., N. Balakrishnan, and N. L. Johnson. 2000. *Continuous multivariate distributions*. New York: Wiley.

Lai, C. D., and M. Xie. 2000. A new family of positive quadrant dependent bivariate distributions. *Statistics & Probability Letters* 46:359–64. doi:10.1016/S0167-7152(99)00122-4.

Lancaster, H. O. 1958. The structure of bivariate distributions. *The Annals of Mathematical Statistics* 29 (3):719–36. doi:10.1214/aoms/1177706532.

Letac, G. 2008. Comment: Lancaster Probabilities and Gibbs Sampling. *Statistical Science* 23: 87–191.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate analysis*. New York: Academic Press.

Nelsen, R. B. 1994. A characterization of Farlie-Gumbel-Morgenstern distributions via Spearman's rho and chi-square divergence. *Sankhya, Series A* 56:476–9.

Nelsen, R. B. 2006. *An introduction to copulas*. 2nd ed. New York: Springer.

Rodríguez-Lallena, J. A., and M. Úbeda-Flores. 2004. A new class of bivariate copulas. *Statistics & Probability Letters* 66:315–25. doi:10.1016/j.spl.2003.09.010.

Ruiz-Rivas, C., and C. M. Cuadras. 1988. Inference properties of a one-parameter curved exponential family of distributions with given marginals. *Journal of Multivariate Analysis* 27 (2): 447–56. doi:10.1016/0047-259X(88)90141-8.

Sklar, A. 1959. Fonctions de repartition à *n* dimensions et les marges, *Publications de l' Institute. d'Statistique de l'Université de Paris* 8: 229–31.