Available online at www.sciencedirect.com
ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 5112–5129

COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Quantile curves and dependence structure for bivariate distributions

F. Belzunce[a], A. Castaño[b], A. Olvera-Cervantes[c], A. Suárez-Llorens[c,*]

[a]Departamento de Estadística e I. O. Universidad de Murcia, Campus de Espinardo 30100 Espinardo (Murcia), Spain
[b]Departamento de Estadística e I. O. Universidad de Cádiz, Campus Universitario de Puerto Real 11510 Puerto Real Cádiz, Spain
[c]Departamento de Estadística e I. O. Universidad deCádiz, C/ Duque de Nájera 8 11002 Cádiz, Spain

Available online 7 September 2006

Abstract

Within the context of a general bivariate distribution an intuitive method is presented in order to study the dependence structure of the two distributions. A set of points—level curve—which accumulate the same probability for a fixed quadrant is considered. This procedure provides four level curves which can be considered as the boundary of a generalization of the real interquantile interval. It is shown that the accumulated probability among the level curves depends on the dependence structure of the distribution function where the dependence structure is given by the notion of copula. Furthermore, the case when the marginal distributions are independent is investigated. This result is used to find out positive or negative dependence properties for the variables. Finally, a nonparametric test for independence with a local dependence meaning is performed and applied to different data sets.

Keywords: Bivariate quantile; Copula; Positive or negative dependence; Central region; Test for independence

1. Introduction

For a one-dimensional probability distribution, the concept of quantile is frequently used for the construction of descriptive statistics and arises in all applied sciences. For example, we can find applications for the study of dispersion, skewness and detection of outliers. The usual definition of the univariate quantile function has the following well known characterizations in the literature:

- *Accumulated probability*: From the inverse of the distribution function given by

$$Q_X(u) = F_X^{-1}(u) = \inf\{x : F_X(x) \geqslant u\} \quad \forall u \in (0,1), \tag{1}$$

  the quantile function provides a point which accumulates a probability u to the left tail and $1-u$ to the right tail.
- *Minimizing distances*: Under some regular conditions, the quantile function is the argument in the following minimization problem

$$\inf_{\theta \in \mathbb{R}} E\left\{ \frac{|X-\theta| + (2\alpha-1)(X-\theta)}{2} - \frac{|X| + (2\alpha-1)X}{2} \right\}. \tag{2}$$

  Note that for $\alpha = \frac{1}{2}$ we obtain the classical characterization of the median.

[*] Corresponding author: Tel.: +34 956015481, fax: +34 956015378.
E-mail address: alfonso.suarez@uca.es (A. Suárez-Llorens).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.08.017

- *Uniform transformation*: Let U be a uniform random variable on [0, 1]. If the distribution function of X is strictly increasing then the quantile function is the only one increasing transformation of U which satisfies

$$Q_X(U) =_{st} X, \tag{3}$$

  where here *st* means equality in distribution.

In addition, the concept of a quantile is linked to the definition of the interquantile interval $IQ(p)=(Q_X(1-p), Q_X(p))$, for $\frac{1}{2} < p < 1$, also called central region for univariate variables.

Since 1976 we have in the literature several attempts towards both multidimensional generalization of univariate quantiles and the real interquantile interval. We find an excellent summary in Serfling (2002). The majority of those generalizations have been developed applying the three characterizations mentioned above. For example, using the characterization based on accumulated probabilities Rousseuw and Leroy (1987) defined the minimum volume ellipsoid with fixed probability, Nolan (1992) and Massé and Theodorescu (1994) defined multivariate quantiles as halfplanes and the central region as convex hull, Koltchinskii (1997) provided a general treatment of multivariate quantiles as inversions of mappings, Chen and Welsh (2002) used this characterization to define the NS-quantiles in $\mathbb{R}^n$ and Fernández-Ponce and Suárez-Llorens (2002) defined multivariate quantiles as level curves. On the other hand, using the characterization based on minimizing distances Tukey (1975) defined the concept of depth function, Oja (1983) gave the simplex median, Avérous and Meste (1997) defined the median balls, Liu et al. (1999) aimed to construct coherent concepts that encompass multivariate notions of ranks, quantiles, location, spread and scale, bias, skewness and kurtosis from different definitions of depth, Abdous and Theodorescu (1992) and Chaudhuri (1996) defined the spatial quantiles and Koshevoy and Mosler (1997) defined the zonoid-quantile as a generalization of a related concept, namely the expectiles. It is also worth to mention that Mosler (2002) provided an overview, with a rather complete bibliography, about multivariate quantiles and depths, with special emphasis in the zonoid-quantile approach. Finally by using the concept of uniform transformation, Fernández-Ponce and Suárez-Llorens (2003) interpreted the standard construction used in Simulation Theory as a multivariate quantile and introduced a new multivariate dispersion order between random vectors. In general, it is not difficult to enumerate the main difficulties that authors found. All authors agree that the main difficulty is the well known fact that there does not exist a natural ordering in n-dimensions, $n > 1$. Additionally, authors also found difficulties in the choice of the shape of the central region which sometimes could not be appropriated for non-symmetrical distributions, the non-parametric estimation of the new concepts and the study of the accumulated probability in the regions which most of the times is less obvious than in the real case.

As a natural generalization of the univariate quantile function Fernández-Ponce and Suárez-Llorens (2002) defined a multivariate quantile as a set of points which accumulate the same probability for a fixed orthant, they called it level curves or quantile curves. This definition also leads to a new multivariate central region concept. In particular those concepts solve some of the above described problems. It is not necessary to choose the shape of the region a priori, it works for symmetrical and non-symmetrical distributions, we can use non-parametric estimations and the central regions are ordered by inclusion as in the univariate case. However, the study of the accumulated probability in the central region and in the upper and lower set defined by the quantile curves was not clear. In particular, the relationship between those probabilities and the dependence structure of the multivariate distribution was not studied. In addition the problem of estimation was not studied in depth.

The purpose of this paper is to use the concept of quantile curve defined in Fernández-Ponce and Suárez-Llorens (2002) in order to study the dependence structure—given by the concept of copula—of a bivariate distribution function and apply this result to present an independence test for bivariate distributions. The organization of this paper is the following. First, in Section 2, we want to reinforce the knowledge of the basic definitions given in Fernández-Ponce and Suárez-Llorens (2002) with special attention to bivariate distributions. This review will help understanding the following sections. In Section 3, we recall some definitions from the concept of Copula. We will show how the accumulated probability of the central region given by the notion of quantile curves depends on the dependence structure of the underlying bivariate distribution. In other words, the dependence relationship between the marginal variables of a bivariate vector restricts the accumulated probability in the central region and outside it. In Section 4, we use the results given in Section 3 to describe a particular study for distribution functions with a positive dependence structure. In Section 5 we present a non-parametric independence test based on the results given in Sections 3 and 4. Finally, in Section 6 we present two numerical examples in order to illustrate our method.

2. Definitions

We consider only bivariate absolutely continuous random vectors, these conditions will be called *regularity conditions*.

In order to introduce some concepts we will need some notation.

Let us denote by $\varepsilon = (\varepsilon_1, \varepsilon_2)$ with $\varepsilon_i \in \{-1, 1\}$, $i = 1, 2$, four directions in $\mathbb{R}^2$ (corresponding to the extreme points of the unit sphere induced by the product topology). Those directions will be used to describe the accumulated probability on the plane. In order to simplify the notation, when we refer to a particular direction we will use the symbols $-$ and $+$ for -1 and 1, respectively. For instance, the direction ε_{--} corresponds to the vector $(-1, -1)$. We will also denote the symbols $\varDelta_-$ and $\varDelta_+$ as the inequalities "$\leqslant$" and "$\geqslant$", respectively and they will be also associated with -1 and 1, respectively.

Definition 1. Let $\mathbf{X} = (X, Y)$ be a random vector under the regularity conditions. Let (x, y) be a point on $\mathbb{R}^2$. We will denote by $F_\varepsilon(x, y)$ the accumulated probability in the quadrant defined by the direction ε, i.e.

$$F_\varepsilon(x, y) = \Pr\{X \varDelta_{\varepsilon_1} x, Y \varDelta_{\varepsilon_2} y\}.$$

For example, $F_{\varepsilon_{-+}}(x, y) = \Pr\{X \varDelta_- x, Y \varDelta_+ y\} = \Pr\{X \leqslant x, Y \geqslant y\}$. The following definition is an intuitive generalization of the univariate quantile for bivariate distributions. This definition takes each quantile as a level curve.

Definition 2. Let $\mathbf{X} = (X, Y)$ be a bivariate random vector under the regularity conditions, and let $p \in [0, 1]$. We define the pth bivariate quantile set or quantile curve for the direction ε, denoted by $Q_{\mathbf{X}}(p, \varepsilon)$, as

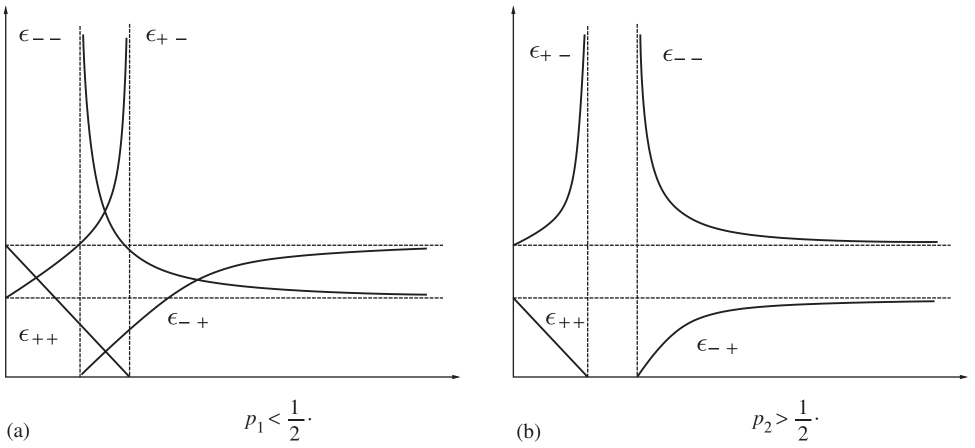
$$Q_{\mathbf{X}}(p, \varepsilon) = \{(x, y) \in \mathbb{R}^2 : F_\varepsilon(x, y) = p\}.$$

Therefore for each $p \in [0, 1]$ we have four quantile curves, where each one can be described by an equation.

Example 3. Let $\mathbf{X}$ be a bivariate random vector with independent and exponentially distributed components with parameter $\lambda = 1$. Given p in the interval $(0, 1)$ it is easy to compute the four quantile curves just solving the equation $F_\varepsilon(x, y) = p$ for all directions. For instance, if we take the direction ε_{++} the solution is given by the quantile curve $y = -x - \ln p$. We show in Fig. 1 (a) and (b) all bivariate quantile curves for $p_1 < \frac{1}{2}$ and $p_2 > \frac{1}{2}$, respectively.

In order to simplify, each quantile curve is just denoted by its direction.

From the definition of the accumulated probability, it is clear to show that all quantiles curves tend asymptotically to the marginal univariate quantile functions. Hence the lines $y = Q_Y(p)$ and $y = Q_Y(1 - p)$ are asymptotes for the



(a) $p_1 < \frac{1}{2}$.

(b) $p_2 > \frac{1}{2}$.

Fig. 1. Bivariate quantile curves: (a) $p_1 < \frac{1}{2}$, (b) $p_2 > \frac{1}{2}$.

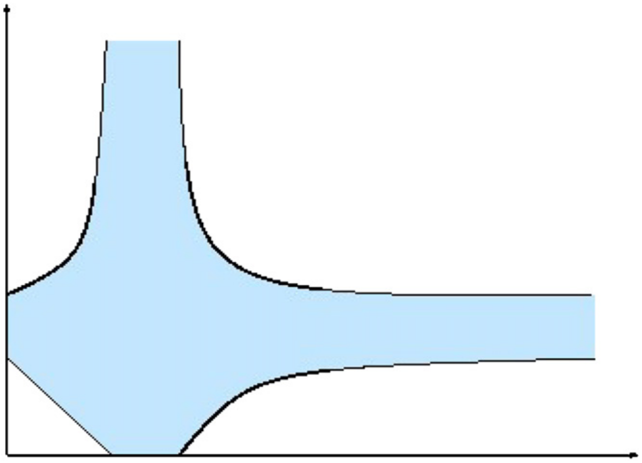


Fig. 2. The central region.

quantile curves given by the pair of directions ε_{--}, ε_{+-} and ε_{++}, ε_{-+}, respectively. Analogously, the lines $x = Q_X(p)$ and $x = Q_X(1-p)$ are asymptotes for the quantile curves given by the pair of directions ε_{--}, ε_{-+} and ε_{++}, ε_{+-}, respectively. It is also clear from the definition that if we take $p > \frac{1}{2}$ the intersection of all quantile curves is empty.

As in the univariate case, linked to the quantile definition, we have a natural definition of central region.

Definition 4. Let $\mathbf{X} = (X, Y)$ be a bivariate random vector and let $p \in [\frac{1}{2}, 1]$. Then we define the central region, denoted $\Omega_{\mathbf{X}}(p)$, as follows:

$$\Omega_{\mathbf{X}}(p) = \{(x, y) \in \mathbb{R}^2 : F_\varepsilon(x, y) < p, \ \forall \varepsilon\}. \tag{4}$$

Roughly, the central region $\Omega_{\mathbf{X}}(p)$ corresponds to the points on the plane which accumulate a probability less than p in all quadrants, i.e. the points among all quantile curves. Note that $\Omega_{\mathbf{X}}(p)$ is a clear generalization of the real interquantile interval and if we increase p the central regions are ordered by inclusion, i.e. $\Omega_{\mathbf{X}}(p) \subset \Omega_{\mathbf{X}}(q)$ for all $p < q$. Finally, the shape of the region is not a priori chosen thus it can be applied to symmetrical and non-symmetrical distributions.

Example 5. To continue with Example 3, the filled up region in Fig. 2 represents the central region for a bivariate exponential random variable with independent components and $p > \frac{1}{2}$ fixed.

In the univariate case we have two tails outside the interquantile interval which are often used to study skewness and detect outliers. The following definition generalizes those tails.

Definition 6. Let $\mathbf{X} = (X, Y)$ be a bivariate random vector and let p in the interval $(0, 1)$. We define the lateral region with order p in the direction ε, denoted by $L_{\mathbf{X}}(p, \varepsilon)$, as

$$L_{\mathbf{X}}(p, \varepsilon) = \{(x, y) : F_\varepsilon(x, y) > p\}. \tag{5}$$

Roughly, the quantile curves can be seen as the boundary of the lateral regions.

To finalize, the quantile curves can be described in a parametric form. Due to the regularity conditions of F, the equation $F_\varepsilon(x, y) = p$ represents a curve on the plane. A straightforward computation shows that those curves can be expressed by means of the quantiles for the conditional distributions $[Y|X \leqslant x]$ and $[Y|X \geqslant x]$ as follows:

$$Q_{\mathbf{X}}(p, \varepsilon_{--}) \to \{(Q_X(u), Q_{Y|X \leqslant Q_X(u)}(p/u)) : u > p\},$$

$$Q_{\mathbf{X}}(p, \varepsilon_{+-}) \to \{(Q_X(u), Q_{Y|X \geqslant Q_X(u)}(p/(1-u))) : u < 1-p\},$$

$$Q_{\mathbf{X}}(p, \varepsilon_{-+}) \to \{(Q_X(u), Q_{Y|X \leqslant Q_X(u)}(1 - p/u)) : u > p\}$$

and

$$Q_{\mathbf{X}}(p, \varepsilon_{++}) \rightarrow \{(Q_X(u), Q_{Y|X \geqslant Q_X(u)}(1 - p/(1-u))) : u < 1 - p\}. \tag{6}$$

Observe that we can also obtain a similar parametric form if we interchange the marginal variables X and Y.

3. The relationship between the level curves and the Copula

The description of a multivariate distribution can be split in two factors: the marginal distributions and the dependence structure among them. It is well known that different marginal distributions could have the same dependence structure. For example, two bivariate normals with the same correlation coefficient. On the other hand, we could have the opposite, the same two marginal distributions could have different dependence structure. For example, normal margins cannot have the dependence structure of a multivariate normal distribution. The dependence structure for a multivariate distribution can be represented by the concept of copula. This concept allows us to separate the effect of the dependence from effects of the marginal distributions, see Nelsen (1999).

In mathematical terms, a copula C is any multivariate distribution function with uniformly distributed marginals on $[0, 1]$. Furthermore, it has been shown that if F is a n-dimensional distribution function, with marginal distribution functions $F_1, \ldots, F_n$ then there exists a n-copula C such that for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$, it holds that $F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n))$. Moreover, if $F_1, \ldots, F_n$ are continuous then C is unique and it is given by the expression

$$C(u_1, \ldots, u_n) = F(Q_{X_1}(u_1), ..., Q_{X_n}(u_n)). \tag{7}$$

Observe that C represents the distribution function of the random variable $(U_1, \ldots, U_n) = (F_{X_1}(X_1), \ldots, F_{X_n}(X_n))$ and it easily holds that

$$\mathbf{X} =_{st} (Q_{X_1}(U_1), \ldots, Q_{X_n}(U_n)), \tag{8}$$

where *st* means the same distribution. A copula C represents an entire family of random vectors having the same dependence structure. In addition, the copula is invariant under increasing transformations of the marginal distributions. For example, the copula of independence is well known in the literature as the one associated with n independent variables which is given by

$$C(u_1, \ldots, u_n) = \prod_{i=1}^{n} u_i. \tag{9}$$

For more details about the general theory of copula, see Nelsen (1999). With these settings, we can formulate the following results.

We will show that the accumulated probability in the central region and the lateral regions defined in (4) and (5), respectively, depend on the copula of the underling random variable, $\mathbf{X}$. First we prove the following technical lemma.

Lemma 7. *Let $\mathbf{X} = (X, Y)$ be a bivariate random vector under the regularity conditions with a copula C and let $u_i \in (0, 1)$ for $i = 1, 2$. Then the probability $F_\varepsilon(Q_X(u_1), Q_Y(u_2))$ depends only on the copula for each direction ε.*

Proof. From the expression (7) for the bivariate case we obtain

$$F_\varepsilon(Q_X(u_1), Q_Y(u_2)) = \begin{cases} C(u_1, u_2), & \varepsilon = \varepsilon_{--}, \\ u_1 - C(u_1, u_2), & \varepsilon = \varepsilon_{-+}, \\ u_2 - C(u_1, u_2), & \varepsilon = \varepsilon_{+-}, \\ 1 - u_1 - u_2 - C(u_1, u_2), & \varepsilon = \varepsilon_{++}. \end{cases} \qquad \square$$

Theorem 8. *Let $\mathbf{X} = (X, Y)$ be a bivariate random vector under the regularity conditions with a copula C. Then the accumulated probability in the central region* $\Pr\{\mathbf{X} \in \Omega_{\mathbf{X}}(p)\}$ *depends solely on the copula.*

Proof. From Definition 4 it holds that

$$\Pr\{\mathbf{X} \in \Omega_{\mathbf{X}}(p)\} = \Pr_{\mathbf{X}}\{(x, y) : F_\varepsilon(x, y) < p, \forall \varepsilon\}.$$

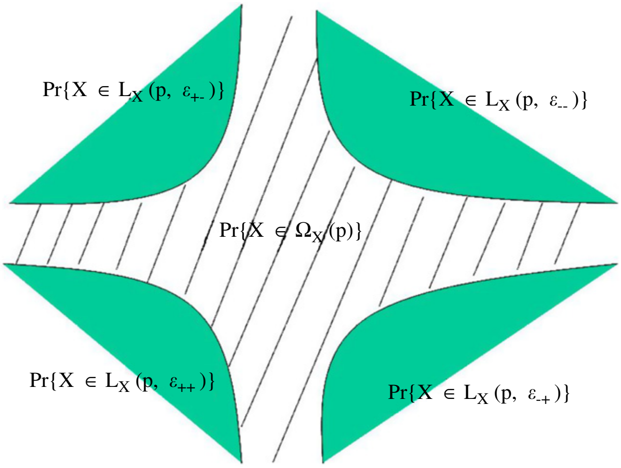


Fig. 3. The accumulated probability.

Using expression (8) for the bivariate case we consider $\mathbf{U}=(U_1,U_2)$ with distribution function C such that $\mathbf{X}{=}_{st}(Q_X(U_1),Q_Y(U_2))$. Then

$$\Pr\{\mathbf{X}\in\Omega_{\mathbf{X}}(p)\}=\Pr\{(Q_X(U_1),Q_Y(U_2))\in\Omega_{\mathbf{X}}(p)\},\tag{10}$$

and it follows, easily, that (10) is equal to

$$\Pr_{\mathbf{U}}\{(u_1,u_2):F_\varepsilon(Q_X(u_1),Q_Y(u_2))<p,\forall\varepsilon\}.$$

The result follows from Lemma 7. □

By the same reasoning a similar result is established for the four lateral regions, and therefore the proof is omitted.

Theorem 9. *Let $\mathbf{X}=(X,Y)$ be a bivariate random vector under the regularity conditions with a copula C. Then the accumulated probability in the lateral region with order p in the direction* ε, $\Pr\{\mathbf{X}\in L_{\mathbf{X}}(p,\varepsilon)\}$, *depends solely on the copula.*

The central region divides the plane in five regions namely the central region and four lateral regions as we can see in Fig. 3. Theorems (8) and (9) state that the accumulated probability in all those regions depends solely on the copula of **X**.

At this point it is natural to wonder about the behavior of the accumulated probabilities for different dependence structures. First we analyze the case of independence.

Corollary 10. *Let $\mathbf{X}=(X,Y)$ be a bivariate random vector under the regularity conditions with independent components. Then*

$$\Pr\{\mathbf{X}\in L_{\mathbf{X}}(p,\varepsilon)\}=1-p+\ln(p),\tag{11}$$

for all direction ε, $p\in[0,1]$. *In addition, for* $p>\frac{1}{2}$ *it holds that*

$$\Pr\{\mathbf{X}\in\Omega_{\mathbf{X}}(p)\}=4p(1+\ln(p))-3.\tag{12}$$

Proof. Without lack of generality we consider the direction $\varepsilon=\varepsilon_{--}$. By hypothesis **X** has a copula given by expression (9) for $n=2$. Then using Theorems 8 and 9 the bivariate random vectors **X** and $\mathbf{U}=(U_1,U_2)$, where U_1 and U_2 are independent and uniformly distributed random variables on [0, 1], have the same accumulated probabilities for the central region and lateral regions. In particular, the probability $\Pr\{\mathbf{X} \in L_{\mathbf{X}}(p, \varepsilon_{--})\}$ is equal to

$$\Pr\{\mathbf{U} \in L_{\mathbf{U}}(p, \varepsilon_{--})\} = \Pr\{U_1 U_2 > p\} = 1 - p + p\ln(p).$$

From previous equality and the fact that lateral regions correspond to the complement of the central region it easily holds that

$$\Pr\{\mathbf{X} \in \Omega_{\mathbf{X}}(p)\} = 1 - 4(1 - p + p\ln(p)) = 4p(1 + \ln(p)) - 3. \qquad \square$$

Now we compare the accumulated probability in the lateral regions for a general bivariate distribution with the corresponding probabilities for a bivariate distribution with independent components. The following result will be key in the next section.

Lemma 11. *Let $\mathbf{X} = (X, Y)$ be a bivariate random vector under the regularity conditions. Then the accumulated probability in the lateral regions satisfies that*

$$\Pr\{(X, Y) \in L_{\mathbf{X}}(p, \varepsilon)\} = \Pr\{F_\varepsilon(X, Y) > p\}, \tag{13}$$

for each direction ε, with corresponding probabilities

$$L_{\mathbf{X}}(p, \varepsilon_{--}) \to PI_p + \int_p^1 p/u - F_{Y|X=Q_X(u)}(Q_{Y|X \leqslant Q_X(u)}(p/u))\,\mathrm{d}u,$$

$$L_{\mathbf{X}}(p, \varepsilon_{++}) \to PI_p + \int_0^{1-p} p/(1-u) - \bar{F}_{Y|X=Q_X(u)}(Q_{Y|X \geqslant Q_X(u)}(1 - p/(1-u)))\,\mathrm{d}u,$$

$$L_{\mathbf{X}}(p, \varepsilon_{+-}) \to PI_p + \int_0^{1-p} p/(1-u) - F_{Y|X=Q_X(u)}(Q_{Y|X \geqslant Q_X(u)})(p/(1-u))\,\mathrm{d}u$$

and

$$L_{\mathbf{X}}(p, \varepsilon_{-+}) \to PI_p + \int_p^1 p/u - \bar{F}_{Y|X=Q_X(u)}(Q_{Y|X \leqslant Q_X(u)}(1 - p/u))\,\mathrm{d}u,$$

where $PI_p = 1 - p + p\ln p$ is the probability in case of independence and $\bar{F}(\cdot) = 1 - F(\cdot)$.

Proof. First, note that from the definition of lateral region it is clear (13). We will show the case of the direction ε_{--}. The proof for the rest of directions is analogous. It holds that

$$\begin{aligned}
\Pr\{\mathbf{X} \in L_{\mathbf{X}}(p, \varepsilon_{--})\} &= \Pr_{\mathbf{X}}\{F(X, Y) > p\}, \\
&= \int_{Q_X(p)}^{\infty} \Pr\{F_{Y|X \leqslant x}(Y|X = x) > p/F_X(x)\} f_X(x)\,\mathrm{d}x, \\
&= \int_{Q_X(p)}^{\infty} 1 - F_{Y|X=x}(Q_{Y|X \leqslant x}(p/F_X(x))) f_X(x)\,\mathrm{d}x, \\
&= \int_p^1 1 - F_{Y|X=Q_X(u)}(Q_{Y|X \leqslant Q_X(u)}(p/u))\,\mathrm{d}u, \\
&= PI_p + \int_p^1 p/u - F_{Y|X=Q_X(u)}(Q_{Y|X \leqslant Q_X(u)}(p/u))\,\mathrm{d}u,
\end{aligned}$$

where the second equality is obtained from $F(x, y) = F_X(x) F_{Y|X \leqslant x}(y)$ for a fixed value x of the marginal distribution X. A straightforward computation leads to the third equality. The fourth one is obtained describing the support of X as $x = Q_X(u)$ where it is well known that the derivative of the quantile function corresponds to the inverse of the density function evaluated at the quantile, i.e. $1/f_X(Q_X(u))$. Last equality is obtained by adding and subtracting the expression PI_p for independent marginal variables. □

Remark 12. Observe that we obtain a similar result if we interchange the marginal variables X and Y in Lemma 11.

From Lemma 11 the probability of all lateral regions is a sum of two factors namely the probability in the independence case and a second factor that modifies this one. If we pay attention to the second one it is computed by evaluating the distribution function of the conditional distribution $\{Y|X=x\}$ at the quantile function of the conditional distribution $\{Y|X\leqslant x\}$ or $\{Y|X\geqslant x\}$, depending on the direction ε. Obviously, the second factor is zero for independent marginal distributions. In the next section we study how the accumulated probability in the lateral regions is modified with respect to the probability of independence for families of bivariate distributions with a positive dependence structure.

4. The relationship to positive dependence concepts

We have shown that the accumulated probability in the central region and the four lateral regions depends on the dependence structure—copula—of the bivariate random variable. We have also shown, that, in the case of independence, all lateral regions have the same accumulated probability. Now it is natural to wonder how the accumulated probabilities are modified for some particular dependence structures. An excellent exposition of the most usual dependence structures can be found in Nelsen (1999). There is a great number of dependence structures and in this section we will pay attention to bivariate random variables with a particular dependence structure known as positive dependence structure. The case of negative ones is similar just reversing all inequalities.

For a bivariate random vector (X, Y) we mean by positive dependence that X and Y are likely to be large or to be small together. We find an excellent exposition of all positive dependence concepts in Joe (1997). It is not our interest to be exhaustive in mentioning all positive dependence concepts that have ever been proposed in the literature. We focus our attention on some of the most useful that can be easily interpreted under the stochastic order. First we need the definition of the stochastic order.

Let Z and W be two random variables and denote by F_Z and F_W their distribution functions, respectively. We will say that Z and W are ordered in the usual stochastic order, denoted by $Z\prec_{st}W$, if and only if

$$F_Z(x)\geqslant F_W(x),$$

for all x. Roughly the above expression says that Z is less likely than W to take on large values, where "large" means any value greater than x, and that this is the case for all x's. We find in Shaked and Shanthikumar (1994) the following characterization which gives a meaningful interpretation. It holds that $Z\prec_{st}W$, if and only if

$$\mathbb{E}(\phi(Z))\leqslant\mathbb{E}(\phi(W)),$$

for all ϕ non-decreasing function.

It is clear from the definition of the stochastic order that the quantiles are ordered, that is to say,

$$Z\prec_{st}W \rightarrow Q_Z(u)\leqslant Q_W(u) \quad \forall u\in[0,1]. \tag{14}$$

Most of the dependence concepts in the literature can be given in terms of the stochastic comparisons among the conditional distributions $[X_2|X_1\leqslant x]$, $[X_2|X_1=x]$ and $[X_2|X_1>x]$ and symmetrically but different if we change X_2 and X_1-.

The first positive dependence concept we deal with was called by Lehmann (1966) positive regression dependence (*PRD*) but most of the authors use the term stochastically increasing (*SI*). We will say that (X, Y) is $PRD(Y|X)$ if

$$[Y|X=x_1]\prec_{st}[Y|X=x_2] \quad \forall x_1\leqslant x_2. \tag{15}$$

Clearly (15) is a positive dependence notion.

Remark 13. Most usual models in simple regression analysis satisfy the *PRD* property. Let us consider the bivariate random vector (X, Y) where X represents an explanatory variable and Y the dependent variable given by $Y=\Phi(X)+Z$, such that $\Phi(X)$ is an increasing transformation of X and Z represents a random error independent of X. A straightforward computation shows that $[Y|X=x_1]\prec_{st}[Y|X=x_2]$ for $x_1<x_2$. Hence (X,Y) is always $PRD(Y|X)$ for all $\Phi(\cdot)$ and Z described as before.

Other well known positive dependence concepts in the statistical literature are the left-tail decreasing (*LTD*) and the right-tail increasing (*RTI*) ones. We will say that (X,Y) is $LTD(Y|X)$ if

$$[Y|X\leqslant x_1]\prec_{st}[Y|X\leqslant x_2] \quad \forall x_1\leqslant x_2. \tag{16}$$

Similarly, (X,Y) is $RTI(Y|X)$ if

$$[Y|X>x_1]\prec_{st}[Y|X>x_2] \quad \forall x_1\leqslant x_2. \tag{17}$$

The reason why (16) and (17) are positive dependence conditions is that, for (16) Y is more likely to take on smaller values as X_1 decreases, and, for (17) Y is more likely to take on larger values as X_1 increases. Note that in both definitions the conditional distribution increases or decreases as a truncated information which has a different interpretation from (15).

For further information about the $PRD(Y|X)$, $LTD(Y|X)$ and $RTI(Y|X)$ concepts see Joe (1997). In particular it holds that

$$PRD(Y|X)\rightarrow LTD(Y|X) \quad \text{and} \quad RTI(Y|X), \tag{18}$$

and there is no implication between $LTD(Y|X)$ and $RTI(Y|X)$.

Now we present two equivalent concepts to *LTD* and *RTI*, respectively. The following definitions will be key later on.

We will say that (X,Y) is left conditional decreasing, denoted by $LCD(X|Y)$, if

$$[Y|X\leqslant x]\prec_{st}[Y|X=x] \quad \forall x. \tag{19}$$

Similarly, (X,Y) is right conditional increasing, denoted $RCI(Y|X)$ if

$$[Y|X=x]\prec_{st}[Y|X>x] \quad \forall x. \tag{20}$$

With an equivalent argument as in the previous concepts, the *LCD* and *RCI* properties have a clear interpretation in terms of positive dependence. We show an illustrative example, let (X,Y) be a bivariate random vector interpreting Y as the life of a son and X the life of a father, both sharing a common genetical disease. If (X,Y) is $RCI(Y|X)$ the expected life of the son whose father is alive at age x is larger than the expected life of the son whose father dies at age x.

The *LCD* and *RCI* positive dependence concepts can be found in the literature, without name, as conditions to establish some criteria for tail monotonicity in terms of the partial derivatives of the copula. From Theorem 5.2.5 and Corollary 5.2.6 in Nelsen (1999) can be easily shown the following two lemmas. The reason why we use the characterizations given by (19) and (20) is because they provide a meaningful interpretation for our interest as we will see later on.

Lemma 14. *Let (X,Y) be a bivariate random variable and denote by F its absolutely continuous distribution function. Then (X,Y) is $LTD(Y|X)$ if and only if (X,Y) is $LCD(Y|X)$.*

Lemma 15. *Let (X,Y) be a bivariate random variable and denote by F its absolutely continuous distribution function. Then (X,Y) is $RTI(Y|X)$ if and only if (X,Y) is $RCI(Y|X)$.*

There exist stronger and weaker positive dependence concepts in the literature than the previous ones, see again Joe (1997). We are interested in the above concepts because they depend on the permutation of the marginal distributions X and Y. This fact will give us a particular interpretation of the accumulated probabilities in the lateral regions as we will see in Proposition 16. Note that there exist some weaker positive notions that do not depend on the permutation of the marginal variables, see the concepts of association and positive quadrant dependence in Joe (1997). On the other hand, we emphasize that the *PRD*, *LTD*, *RTI* concepts satisfy the majority of the desirable properties related with a positive dependence concept. For example, they are all invariant with respect to the copula, that is if (X, Y) has one of the mentioned positive dependence property then all bivariate random vectors with a common copula as (X, Y) have also the same property. They also satisfy that the usual Kendall's tau and Spearman's rho coefficients are positive where it is well known that these coefficients are invariant under the same copula. And finally they also imply that the Pearson's correlation coefficient is greater or equal than zero. It is worth to mention that the Pearson's correlation coefficient is not invariant with respect to the same copula and it is more desirable as a measure of positive dependence for bivariate normal distributions. There are many families of bivariate distribution functions which have a positive dependence structure, for example it is well known that a bivariate normal distribution (X, Y) is $PRD(Y|X)$ and $PRD(X|Y)$ if and only if the Pearson's correlation coefficient is greater than zero. We can find many illustrative examples in Joe (1997), Nelsen (1999) and Drouet and Kotz (2001).

Now we analyze how the accumulated probability in the lateral regions is modified under a positive dependence structure.

Proposition 16. *Let* (X, Y) *be a bivariate random vector under the regularity conditions. Then it holds the following implications*:

$$(X, Y) \text{ is } LCD(Y|X) \to \begin{cases} \Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{--})\} > PI_p, \\ \Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{-+})\} < PI_p. \end{cases}$$

$$(X, Y) \text{ is } RCI(Y|X) \to \begin{cases} \Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{++})\} > PI_p, \\ \Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{+-})\} < PI_p. \end{cases}$$

Similarly, *if we interchange X and Y*, *it holds that*

$$(X, Y) \text{ is } LCD(X|Y) \to \begin{cases} \Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{--})\} > PI_p, \\ \Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{+-})\} < PI_p. \end{cases}$$

$$(X, Y) \text{ is } RCI(Y|X) \to \begin{cases} \Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{++})\} > PI_p, \\ \Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{-+})\} < PI_p. \end{cases}$$

for all p in the interval [0, 1].

Remark 17. From Eq. (13) and Corollary 10 Proposition 16 provides us a stochastic comparison between the distribution function of the univariate variable $F_\varepsilon(X, Y)$ and $F_X(X')F_Y(Y')$, where X', Y' are independent copies of X and Y, respectively. For instance, let us consider the direction ε_{--}, the inequality $\Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{--}\} > PI_p$ holds for all p if and only if $F_X(X')F_Y(Y') \prec_{st} F(X, Y)$. These kind of stochastic comparisons are clearly related with the positive K-dependence concept as defined in Capéraà et al. (1997). The univariate random variable given by the transformation $F(X, Y)$ is known in the literature as the probability integral transformation and has been studied in depth by Genest and Boies (2003) and Genest et al. (2006).

Remark 18. Observe that when we interchange X and Y we obtain a similar result for the directions ε_{--} and ε_{++} but not for ε_{-+} and ε_{+-}.

Proof. The proof is equivalent for each direction ε. Then we will only prove the result for the lateral region of order p in the direction ε_{--}. Using expressions in Lemma 11 it holds that

$$\Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{--})\} = PI_p + \int_p^1 p/u - F_{Y|X=Q_X(u)}(Q_{Y|X \leqslant Q_X(u)}(p/u)\,\mathrm{d}u.$$

By hypothesis (X, Y) is $LCD(Y|X)$. Then by definition $[Y|X \leqslant x] \prec_{st} [Y|X = x]$ for all x. Using (14) for the stochastic comparison it easily holds that the conditional distribution function $[Y|X=x]$ evaluated at the quantiles of the conditional distribution function $[Y|X \leqslant x]$ satisfies that $F_{[Y|X=x]}(Q_{[Y|X \leqslant x]}(u)) \leqslant u$, for all u. Hence just looking at the second term of the above equality it easily holds that $\Pr\{(X, Y) \in L_{(X,Y)}(p, \varepsilon_{--})\} > PI_p$.

Similarly, if we interchange Y and X the accumulated probability in the lateral region can be expressed as

$$\Pr\{(X,Y)\in L_{(X,Y)}(p,\varepsilon_{--})\}=PI_p+\int_p^1 p/u-F_{X|Y=Q_Y(u)}(Q_{X|Y\leqslant Q_Y(u)}(p/u)\,\mathrm{d}u.$$

If we assume now that (X,Y) is $LCD(X|Y)$ we obtain the same result with a similar argument. □

At this point it is interesting to show the maximum and minimum values for the accumulated probabilities in the lateral regions in order to provide a bound for them.

Proposition 19. *Let us consider the lateral region with order p in the direction ε. Then its accumulated probability can be bounded by*

$$0\leqslant \Pr\{(X,Y)\in L_{(X,Y)}(p,\varepsilon)\}\leqslant 1-p.$$

Proof. The proof is equivalent for all lateral regions. We will only consider the direction ε_{--}. Looking at the proof of Lemma 11 it holds that

$$\Pr\{(X,Y)\in L_{(X,Y)}(p,\varepsilon_{--})\}=\int_p^1 1-F_{Y|X=Q_X(u)}(Q_{Y|X\leqslant Q_X(u)}(p/u)\,\mathrm{d}u.$$

The result follows easily from the inequalities $0\leqslant F_{Y|X=Q_X(u)}(y)\leqslant 1$. □

Observe that Proposition 19 is valid for whatever dependence structure. Using Propositions 16 and 19 we obtain the lower and upper bounds for the accumulated probability in the lateral regions in case of positive dependence structure. Observe that these bounds are achieved in the case of maximum positive dependence, i.e. if Y is an increasing function of X then it is clear that the probabilities $\Pr\{(X,Y)\in L_{(X,Y)}(p,\varepsilon_{--})\}$ and $\Pr\{(X,Y)\in L_{(X,Y)}(p,\varepsilon_{++})\}$ are equal to $(1-p)$ and $\Pr\{(X,Y)\in L_{(X,Y)}(p,\varepsilon_{-+})\}$ and $\Pr\{(X,Y)\in L_{(X,Y)}(p,\varepsilon_{+-})\}$ are equal to 0.

From Propositions 16 and 19 and the posterior discussions the accumulated probability in the lateral regions can be seen as measures of positive dependence. Note that if (X,Y) has a positive dependence structure then the accumulated probability in the lateral regions with order p in the pair of directions ε_{--}, ε_{++} and ε_{+-}, ε_{-+} would be expected to be greater or smaller than in the case of independence, respectively. Also all those probabilities do not necessarily have to be modified simultaneously and they depend on the structure of the conditional variable when we truncate to the left or to the right. On the other hand, for instance if (X,Y) has a strong positive dependence property as $PRD(Y|X)$ or $PRD(X|Y)$ then it is expected that the four lateral regions modify their probabilities simultaneously. We also emphasize that the accumulated probability depends on the value of p. Therefore, they can also be seen as local measures of dependence, where local means that we could find the dependence relationship between the variables only for some values of p. Note that it is easy to describe situations where the classical global measures of dependence do not take successfully into account the association between two variables. The reason for this is that sometimes the dependence does not exist on the whole plane.

Remark 20. Note that from Proposition 19 the accumulated probability in each lateral region is always between 0 and $1-p$ corresponding to case of maximum positive or negative dependence. In case of independence it holds that this probability is $1-p+p\ln p$ where the term $p\ln p$ seems to be associated with a particular measure of the entropy which we have not been able to give a meaning, “it is intuitive that random values spread out on the plane achieve the maximum entropy in case of independence”.

5. Test for independence and estimation

The purpose of this section is to present a test for independence based on the quantile curves and the accumulated probability in the lateral regions. The problem of independence has been dealt deeply in the literature, some references in this direction are Deheuvels (1979,1981a,b,c).

Let (X,Y) be a bivariate random vector under the regularity conditions and let $p\in[\frac{1}{2},1)$. Let (X_i,Y_i), $i=1,\ldots,n$, be n independent and identically distributed copies of (X,Y). If we consider each lateral region and the central region associated with p as five categories we can formulate the chi-squared test where we compare the expected frequencies in each category in the case of independence, given by (11) and (12), with the observed probabilities.

Note that for this test, the null hypothesis H_0 is that of independence, and the alternative H_1 is that there exist some dependence. This test can be stated for every p and this fact will provide a special local dependence meaning as we mentioned at the end of Section 4. On the other hand, if we reject the null hypothesis we will have in Proposition 16 a useful result to find some positive or negative dependence structure. In this paper we are not interested in studying the power of the test which could be dealt with a particular alternative hypothesis, for some recent literature in this direction see for example De Martini and Vespa (2005) and Rödel and Kössler (2004).

The above method has a main difficulty. Although no estimation is necessary to compute the expected frequencies. However, we cannot count directly the observed frequencies. So our next objective is to give an approximation to the observed proportions in each region in order to justify the use of a classical chi-squared test.

Without lack of generality we now focus on the lateral region with order p in the direction ε_{--}, where the observed frequency can be obtained as follows:

Let us introduce the random variable Z_i defined as,

$$Z_i = \begin{cases} 1, & F(X_i, Y_i) > p, \\ 0, & F(X_i, Y_i) \leqslant p. \end{cases}$$

The observed frequency is given by $\sum_{i=1}^{n} Z_i$, which follows a binomial distribution with parameters n and success probability $Pr\{F(X, Y) > p\}$. Obviously the observed probability in that category is $\sum_{i=1}^{n} Z_i/n$.

Using the empirical distribution function of the bivariate random sample denoted by $F_n(x, y)$ we propose the following method in order to provide and approximation of the observed proportion:

Let us consider the Bernoulli variables

$$Z_{n,i} = \begin{cases} 1, & F_n(X_i, Y_i) > p, \\ 0, & F_n(X_i, Y_i) \leqslant p, \end{cases} \tag{21}$$

that are identically distributed for $1 \leqslant i \leqslant n$ and fixed n, according to a Bernoulli distribution with parameter $\theta_n = Pr\{F_n(X_1, Y_1) > p\}$.

Then we propose as a natural estimator of the observed proportion the statistic given by $\sum_{i=1}^{n} Z_{n,i}/n$.

We first note that $\sum_{i=1}^{n} Z_{n,i}/n$ is a function of the empirical copula. Note that it is a well known fact that the empirical copula is a sufficient statistic of the copula, see for example, De Martini and Vespa (2005) for a nice exposition of the properties of this kind of estimators.

In order to study convergence properties of this estimator it will be necessary to propose some previous results.

It is well known that $F_n(x, y)$ is a strongly consistent estimator of $F(x, y)$ for all $(x, y) \in \mathbb{R}^2$, that is $F_n(x, y) \xrightarrow{\text{a.s.}} F(x, y)$ as $n \to \infty$. This result holds if we substitute the fixed values (x, y) by a bivariate random vector (X', Y'), identically distributed with (X, Y) and independent with the considered random sample from (X, Y), as we show in the next lemma. The proof follows directly from the definition of the convergence and has been omitted.

Lemma 21. *Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from a bivariate random vector (X, Y), with distribution function $F(x, y)$ and $F_n(x, y)$ its empirical distribution function. Consider X', Y' a copy of X, Y and independently distributed with the random sample. Then*

$$F_n(X', Y') \xrightarrow{\text{a.s.}} F(X', Y').$$

As a consequence we have,

Corollary 22. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate distribution $F(x, y)$, then with the ith component fixed we have*

$$F_n(X_i, Y_i) \xrightarrow{\text{a.s.}} F(X_i, Y_i),$$

as $n \to \infty$.

Proof. It is sufficient to apply Lemma 21 to the expression

$$F_n(X_i, Y_i) = \frac{1}{n} + \frac{n-1}{n} F_{n-1}(X_i, Y_i),$$

where $F_{n-1}(X_i, Y_i)$ represents the empirical distribution function based on the random sample of size $n-1$ formed for all the observations except for the ith component. □

To continue with the convergence properties, we first analyze the behavior of the sequence $\{Z_{n,i}\}_n$ defined in (21).

Lemma 23. *Consider the random variable Z_i and $Z_{n,i}$ defined as before, then the sequence of random variables $\{Z_{n,i}\}_n$ with $1 \leqslant i \leqslant n$ fixed, converges in mean square to Z_i, denoted by $Z_{n,i} \xrightarrow{2} Z_i$, that is*

$$\lim_{n\to\infty} E\{(Z_{n,i} - Z_i)^2\} = 0 \quad \forall i, \ 1 \leqslant i \leqslant n.$$

Proof. From definitions of $Z_{n,i}$ and Z_i, easily we have

$$E\{(Z_{n,i} - Z_i)^2\} = Pr\{|Z_{n,i} - Z_i| = 1\}. \tag{22}$$

Therefore, it is enough to show

$$\lim_{n\to\infty} Pr\{|Z_{n,i} - Z_i| = 1\} = 0. \tag{23}$$

By definition, $Pr\{|Z_{n,i} - Z_i| = 1\}$ is equal to

$$Pr\{F_n(X_i, Y_i) > p, F(X_i, Y_i) \leqslant p\} + Pr\{F_n(X_i, Y_i) \leqslant p, F(X_i, Y_i) > p\}.$$

We only show that the first term of the above sum converges to zero as n tends to infinite, the study for the second term follows under a similar argument.

Consider $\delta > 0$,

$$\begin{aligned} Pr\{F_n(X_i, Y_i) > p, F(X_i, Y_i) \leqslant p\} = {} & Pr\{F_n(X_i, Y_i) > p, F(X_i, Y_i) \leqslant p - \delta\} \\ & + Pr\{F_n(X_i, Y_i) \geqslant p + \delta, F(X_i, Y_i) \leqslant p\} \\ & - Pr\{F_n(X_i, Y_i) \geqslant p + \delta, F(X_i, Y_i) \leqslant p - \delta\} \\ & + Pr\{p < F_n(X_i, Y_i) \leqslant p + \delta, p - \delta < F(X_i, Y_i) \leqslant p\}. \end{aligned}$$

The first three terms of the above equality verify

$$\Pr\{F_n(X_i, Y_i) > p, F(X_i, Y_i) \leqslant p - \delta\} \leqslant \Pr\{|F_n(X_i, Y_i) - F(X_i, Y_i)| > \delta\},$$

$$\Pr\{F_n(X_i, Y_i) \geqslant p + \delta, F(X_i, Y_i) \leqslant p\} \leqslant \Pr\{|F_n(X_i, Y_i) - F(X_i, Y_i)| > \delta\}$$

and

$$\Pr\{F_n(X_i, Y_i) \geqslant p + \delta, F(X_i, Y_i) \leqslant p - \delta\} \leqslant \Pr\{|F_n(X_i, Y_i) - F(X_i, Y_i)| > \delta\}.$$

Then using Corollary 22, the common right-hand side of these inequalities tends to zero as n tends to infinite.

With respect to the fourth term, it holds

$$Pr[p < F_n(X_i, Y_i) \leqslant p + \delta, p - \delta < F(X_i, Y_i) \leqslant p] \leqslant Pr[p - \delta < F(X_i, Y_i) \leqslant p].$$

Therefore, taking into account that the parent random vector is under regularity conditions, this term tends to 0 when δ tends to 0. □

Lemma 24. *Also,*

$$\left| \frac{\sum_{i=1}^n Z_{n,i}}{n} - \frac{\sum_{i=1}^n Z_i}{n} \right| \xrightarrow{2} 0,$$

as $n \to \infty$.

Proof. Note that $\{Z_{n,i}\}_{1\leqslant i\leqslant n}$ are identically distributed, then $\{Z_{n,i}-Z_i\}$ are identically distributed.

In order to study the convergence in mean square, we have

$$E\left|\frac{\sum_{i=1}^n Z_{n,i}}{n}-\frac{\sum_{i=1}^n Z_i}{n}\right|^2 \leqslant E\left[\frac{\sum_{i=1}^n |Z_{n,i}-Z_i|}{n}\right]^2 = \frac{\sum_{i=1}^n E|Z_{n,i}-Z_i|^2}{n^2}+\frac{\sum_{i\neq j}E|Z_{n,i}-Z_i||Z_{n,j}-Z_j|}{n^2},$$

where the first term verifies

$$\frac{\sum_{i=1}^n E|Z_{n,i}-Z_i|^2}{n^2}=\frac{nE|Z_{n,1}-Z_1|^2}{n^2}, \tag{24}$$

since $\{Z_{n,i}-Z_i\}$ have the same expectation. And taking limit in (24) as $n\to\infty$, the first term vanishes.

With respect to the second term, applying the Cauchy–Schwarz inequality it holds

$$\frac{\sum_{i\neq j}E|Z_{n,i}-Z_i||Z_{n,j}-Z_j|}{n^2}\leqslant\frac{\sum_{i\neq j}\{E|Z_{n,i}-Z_i|^2\}^{1/2}\{E|Z_{n,j}-Z_j|^2\}^{1/2}}{n^2}, = \frac{n(n-1)}{n^2}E|Z_{n,1}-Z_1|^2,$$

which converges to zero as $n\to\infty$ from Lemma 23. □

Therefore, using the proposed approximation to the observed proportions, we can approximate the test of independence as a chi-squared test with four degree of freedom.

Respect to the estimation of the level curves, we directly estimate the parametric form given by (6) where we only have to estimate the conditional and marginal quantiles. Let $x_{(1)}\leqslant\cdots\leqslant x_{(n)}$ be the ordered values of the i.i.d. sample of X. We denote $\hat{Q}_{Y|X\leqslant x_{(i)}}(.)$ the estimator of the conditional quantile defined in terms of the empirical distribution. Then, the set

$$\hat{Q}_{\mathbf{X}}(p,(-1-1))=\{\mathbf{p}_i=(x_{(i)},\hat{Q}_{Y|X\leqslant x_{(i)}}(n*p/i)):i/n>p\},$$

provides a non-parametric estimation of $Q_{\mathbf{X}}(p,\varepsilon_{--})$ by a family of points. Note that when $n\to\infty$ we have guaranteed the convergence. We also represent the lines connecting the points $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$ to interpret more easily the different regions.

6. Numerical examples

Example 25. First we illustrate our method with real data. For this purpose, we consider the sulfur oxide (SO_2) and nitrogen oxides (NO_x) concentration level in the lowest latitude of the south of Spain. The observations were supplied by the corresponding local government of the council of Cádiz, Spain. The contamination variables, SO_2 and NO_x, were measured each day from a monitoring network system during 1994 and expressed in mg/m^3, with 300 observations. We show in Fig. 4 the dispersion diagram where the sulfur oxide is in the horizontal axis and the nitrogen oxide is in the vertical axis.

The presence of an outlier is obvious. If we take into account the outlier, the Spearman's rho takes the value 0.165 which is statistically significative with p-value <0.05 and the Pearson's correlation takes the value 0.0975 and it is not statistically significative, p-value >0.1. If we delete the outlier then we obtain a similar result for the Spearman's rho. However, the Pearson's correlation takes the value 0.1615 being now statistically significative p-value <0.05. It is worth to mention that Wilcox (2005) discusses the influence of outliers in classical coefficients of dependence for different data structures. In our case, both coefficients suggest a slight positive dependence. The Spearman's rho and the Pearson's correlation are global measures of dependence. In other words, we summarize the dependence structure in just a measure.

Fig. 5(a) and (b) and Fig. 6(a) represent the estimations of the central and lateral regions for $p=0.5$, $p=0.6$ and $p=0.7$, respectively. Tables 1–3 show the result to compute a chi-squared test for $p=0.5$, $p=0,6$ and $p=0,7$,

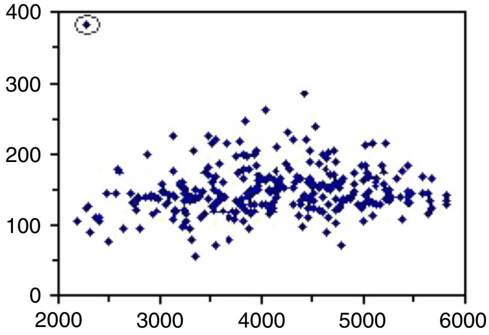


Fig. 4. (SO_2, NO_x).

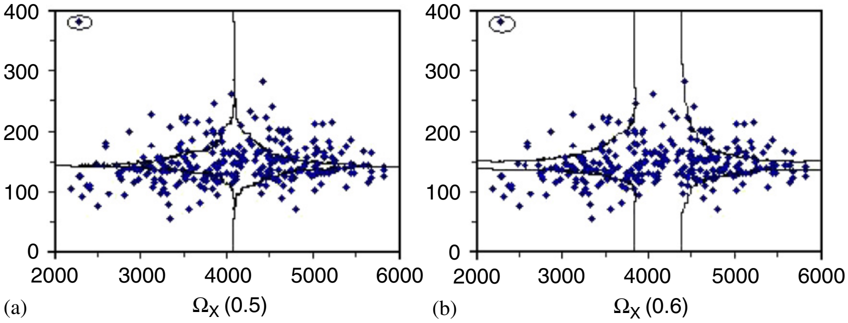


Fig. 5. Central regions: (a) $\Omega_{\mathbf{X}}(0.5)$; (b) $\Omega_{\mathbf{X}}(0.6)$.

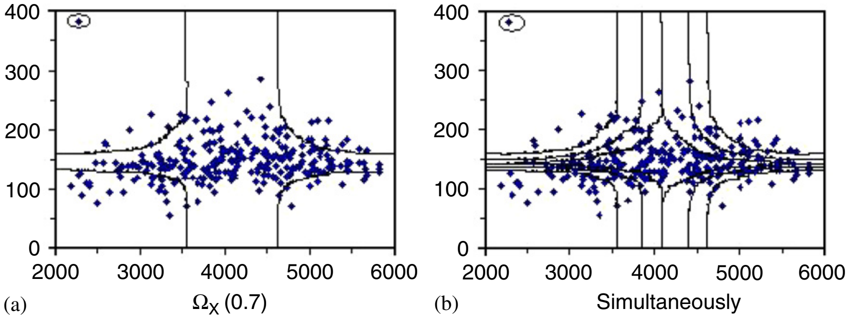


Fig. 6. Central regions: (a) $\Omega_{\mathbf{X}}(0.7)$; (b) simultaneously.

Table 1
Test of independence for (SO_2, NO_X), $p = 0.5$

$p = 0.5$	$L_{(X,Y)}(p, \varepsilon_{--})$	$L_{(X,Y)}(p, \varepsilon_{-+})$	$L_{(X,Y)}(p, \varepsilon_{+-})$	$L_{(X,Y)}(p, \varepsilon_{++})$	$\Omega_{(X,Y)}(p)$
Observed	0.158450704	0.137323944	0.10915493	0.200704225	0.394366197
Expected	0.15342641	0.15342641	0.15342641	0.15342641	0.386294361
Chi-squared	4 d.f.	8.34	p-Value	$0.05 < p < 0.1$	

Table 2
Test of independence for $(SO_2, NO_X), p = 0.6$

$p = 0.6$	$L_{(X,Y)}(p, \varepsilon_{--})$	$L_{(X,Y)}(p, \varepsilon_{-+})$	$L_{(X,Y)}(p, \varepsilon_{+-})$	$L_{(X,Y)}(p, \varepsilon_{++})$	$\Omega_{(X,Y)}(p)$
Observed	0.091549296	0.066901408	0.052816901	0.13028169	0.658450704
Expected	0.093504626	0.093504626	0.093504626	0.093504626	0.625981497
Chi-squared	4 d.f.	11.775	p-Value	$p < 0.05$	Reject

Table 3
Test of independence for $(SO_2, NO_X), p = 0.7$

$p = 0.7$	$L_{(X,Y)}(p, \varepsilon_{--})$	$L_{(X,Y)}(p, \varepsilon_{-+})$	$L_{(X,Y)}(p, \varepsilon_{+-})$	$L_{(X,Y)}(p, \varepsilon_{++})$	$\Omega_{(X,Y)}(p)$
Observed	0.038732394	0.028169014	0.028169014	0.091549296	0.813380282
Expected	0.050327539	0.050327539	0.050327539	0.050327539	0.798689843
Chi-squared	4 d.f.	15.965	p-Value	$p < 0.05$	Reject

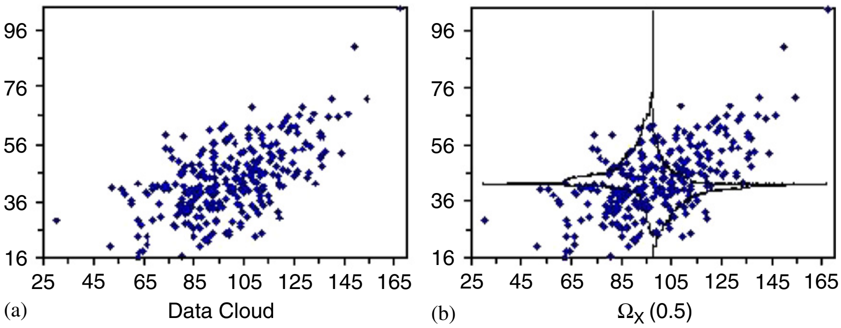


Fig. 7. Simulated example: (a) Data cloud; (b) $\Omega_{\mathbf{X}}(0.5)$.

respectively. The first row in all tables provides the observed accumulated probability and the second one the theoretical probability in case of independence given by $1 - p + \ln p$. Note that we have not deleted the outlier because just a few points are not significative in the accumulated probability.

Observe that we reject that both variables are independent. For $p = 0.5$ it is not so clear but for $p = 0.6$ and $p = 0.7$ we reject with a p-value smaller than 0.05, this last gives a particular local dependence meaning to the accumulated probabilities, i.e. the dependence could not exist on the whole plane.

If we pay attention to the tables, we note that the biggest differences between the observed accumulated probabilities and the expected values in case of independence are achieved for the directions ε_{++} and ε_{+-}. However, this difference does not seem to be significative for the variation ε_{--}, for $p = 0.5$ and $p = 0.6$. Note that a classical regression model, as described in Remark 13, could not be appropriated here. Those models are *PRD* by construction and from (18) also *LTC* and *RTI*. Hence using the characterizations shown in Lemmas 14, 15 and Proposition 16, it would be reasonable to expect that the accumulated probability in the lateral region given by the variation ε_{--} to be bigger than PI_p. From the above discussion and Proposition 16 we could expect a weaker positive dependence property, such as $RCI(NO_x|SO_2)$, and we could think that NO_x increases its concentration when we know that SO_2 has exceeded a threshold amount.

Example 26. Now we present a simulated example. For our purpose, we consider the classical regression model shown in Remark 13. Let (X, Y) be a random vector where the explanatory variable is given by $Y = \Phi(X) + Z$, such that $\Phi(X)$ is an increasing transformation of X and Z represents a random error independent of X. Hence (X, Y) is $PRD(Y|X)$ and from (18) also $LTD(Y|X)$ and $RTI(Y|X)$. Therefore using Lemma 14, Lemma 15 and Proposition 16 it easily holds that the accumulated probabilities in the lateral regions for the pair of directions ε_{--}, ε_{++} and ε_{+-}, ε_{-+} are expected to be greater and smaller than in the case of independence, respectively.

Table 4
Test of independence for $(X, \Phi(X) + Z)$, $p = 0.5$

$p = 0.5$	$L_{(X,Y)}(p, \varepsilon_{--})$	$L_{(X,Y)}(p, \varepsilon_{-+})$	$L_{(X,Y)}(p, \varepsilon_{+-})$	$L_{(X,Y)}(p, \varepsilon_{++})$	$\Omega_{(X,Y)}(p)$
Observed	0.281690141	0.038732394	0.035211268	0.278169014	0.366197183
Expected	0.15342641	0.15342641	0.15342641	0.15342641	0.386294361
Chi-squared	4 d.f.	109.771	p-Value	$p < 0.05$	Reject

Under the above discussion, we have simulated 300 data of a normal distribution $X \sim N(100; 20)$ and a normal error distribution $Z \sim N(0; 10)$ and we have taken an increasing transformation $\Phi(X) = 30 + \exp(0.025X)$. Fig. 7 represents the data collection and the estimation of the central and lateral regions for $p = 0.5$.

In Table 4 we obviously reject the independence. We want to emphasize that the observed accumulated probabilities differ significatively from the expected values in case of independence for all lateral regions.

Acknowledgments

We would like to thank two referees and Rand Wilcox whose careful reading and marking in the first version of this manuscript led to a number of corrections which greatly improved this article. Suárez-Llorens research was supported by the Spanish National Program of Research, grant SEJ 2005-06678-ECON.

References

Abdous, B., Theodorescu, R., 1992. Note on the spatial quantile of a random vector. Statist. Probab. Lett. 13, 333–336.

Avérous, J., Meste, M., 1997. Median balls: an extension of the interquantile intervals to multivariate distributions. J. Multivariate Anal. 63, 222–241.

Capéraà, P., Fougères, A.L., Genest, C., 1997. A stochastic ordering based on a decomposition of Kendall's Tau. In: Benes, V., Stepan, J. (Eds.), Distributions with given Marginals and Moment Problems. Kluwer, Boston, pp. 81–86.

Chaudhuri, 1996. On a geometric notion of quantiles for multivariate data. J. Amer. Statist. Assoc. 91 (434), 862–872.

Chen, L.A., Welsh, A.H., 2002. Distribution-function-based bivariate. J. Multivariate Anal. 83, 208–231.

Deheuvels, P., 1979. La Fonction de Dépendence Empirique et ses Propiétés. Un Test non Paramétrique d'Indépendence. Acad. Roy. Belg. Bull. CL. Sci. 65 (5), 274–292.

Deheuvels, P., 1981a. A Kolmogorov–Smirnov type test for independence and multivariate samples. Rev. Roumaine Math. Pures Appl. 26, 213–226.

Deheuvels, P., 1981b. A non parametric test for independence. Publ. Inst. Statist. Univ. Paris 26, 29–50.

Deheuvels, P., 1981c. Multivariate tests of independence. Analytical Methods in Probability (Oberwolfach, 1980). Lecture Notes in Mathematics 861 (Springer-Verlag, Berlin), 42–50.

De Martini, D., Vespa, E., 2005. Copula-based models for the power of independence test. Commun. Statist.-Theory Methods 34, 2283–2297.

Drouet, D., Kotz, S., 2001. Correlation and Dependence. Imperial Collegue Press, UK.

Fernández-Ponce, J.M., Suárez-Llorens, A., 2002. Central regions for bivariate distributions. Austrian J. Statist. 31 (2–3), 141–156 On Line Avaiable.

Fernández-Ponce, J.M., Suárez-Llorens, A., 2003. A multivariate dispersion ordering based on quantiles more widely separated. J. Multivariate Anal. 85, 40–53.

Genest, C., Boies, J.-C., 2003. Detecting dependence with Kendall plots. Amer. Statist. 57 (4), 275–284.

Genest, C., Quessy, J.-F., Rémillard, B., 2006. Goodness-of-fit procedures for copula models based on the probability integral transformation. Scandinavian J. Statist. 33 (2), 337–366.

Joe, H., 1997. Multivariate Models and Dependence Concepts. Chapman and Hall, London.

Koltchinskii, V., 1997. M-estimation, convexity and quantiles. Ann. Statist. 25, 435–477.

Koshevoy, G., Mosler, K., 1997. Zonoid trimming for multivariate distributions. Ann. Statist. 25 (5), 1998–2017.

Lehmann, E.L., 1966. Some concepts of dependence. Ann. Math. Statist. 37, 1137–1153.

Liu, R.Y., Parelius, J.M., Singh, K., 1999. Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann. Statist. 27 (3), 783–858.

Massé, J.C., Theodorescu, R., 1994. Halfplane trimming for bivariate distributions. J. Multivariate Anal. 48, 188–202.

Mosler, K., 2002. Multivariate Dispersion, Central Regions and Depth. Lecture Notes in Statistics, Springer, Berlin.

Nelsen, R.B., 1999. An Introduction to Copulas. Springer, New York.

Nolan, D., 1992. Asymptotics for multivariate trimming. Stochastic Processes and their Appl. 42, 157–169.

Oja, H., 1983. Descriptive statistics for multivariate distributions. Statist. Probab. Lett. 10, 407–410.

Rödel, E., Kössler, W., 2004. Linear rank tests for independence in bivariate distributions-power comparisons by simulation. Comput. Statist. Data Anal. 46 (4), 645–660.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.
Serfling, R., 2002. Quantile functions for multivariate analysis: approaches and applications. Statistica Neerlandica 56 (2), 214–232.
Shaked, M., Shanthikumar, J.G., 1994. Stochastic Orders and Their Applications. Academic Press, New York.
Tukey, J.W., 1975. Mathematics and the picturing of data. Congr. Math. 2, 523–531.
Wilcox, R., 2005. Introduction to Robust Estimation and Hypothesis Testing. Academic Press, Burlington.