Research Article **Open Access**

Jorge Navarro*

Bivariate box plots based on quantile regression curves

https://doi.org/10.1515/demo-2020-0008
Received April 21, 2020; accepted June 20, 2020

Abstract: In this paper, we propose a procedure to build bivariate box plots (BBP). We first obtain the theoretical BBP for a random vector (X, Y). They are based on the univariate box plot of X and the conditional quantile curves of $Y|X$. They can be computed from the copula of (X, Y) and the marginal distributions. The main advantage of these BBP is that the coverage probabilities of the regions are distribution-free. So they can be selected by the users with the desired probabilities and they can be used to perform fit tests. Three reasonable options are proposed. They are illustrated with two examples from a normal model and an exponential model with a Clayton copula. Moreover, several methods to estimate the theoretical BBP are discussed. The main ones are based on linear and non-linear quantile regression. The others are based on empirical estimators and parametric and non-parametric (kernel) copula estimations. All of them can be used to get empirical BBP. Some extensions for the multivariate case are proposed as well.

Keywords: Median regression, quantile confidence bands, copula, kernel estimation

MSC: 62G99, 62G07

1 Introduction

The (univariate) Tukey's box plots are very useful tools to analyse and compare data (random variables) both in size (mean, median, quantiles, etc.) and in dispersion (range, interquartile range, etc.). They can also be used to detect outliers.

When we analyse bivariate (paired) data we have different options in order to define bivariate box plots (BBP). From a theoretical point-of-view, when we study the (real-valued) random vector (X, Y), we can use the regions determined by the Mahalanobis distance and the associated lower bounds for their probabilities provided by the multivariate Chebyshev's inequality (see [5, 17]). Exact probabilities can be computed for normal (Gaussian) random vectors by using the Chi-squared distribution (see p. 39 in [14]). Extensions to the multivariate case can be obtained through a standard principal component analysis (see [17]). In practice, we can use the regions determined by the estimated means, variances and covariances or that provided by the empirical distribution of the data set (see [16]).

Another popular empirical option is to use the *bagplots* proposed in [20]. They are based on the concept of *depth* and provide a region (called the *bag*) based on the *depth median* which contains the 50% of the data. This region is inflated three times to obtain the *fence* which separates *inliers* from *outliers*.

A third good option is provided by the quantile regression curves. This option is especially useful when we want to estimate Y (response variable) from X (explanatory variable). In the general case we could use the regions proposed in [6] instead. The theoretical quantile regression curves can be obtained from the copula of (X, Y) (see p. 217 in [19]). In practice we can use the empirical linear or non-linear estimators proposed in

***Corresponding Author: Jorge Navarro:** Facultad de Matemáticas, Universidad de Murcia, 30100 Murcia, Spain, E-mail: jorgenav@um.es

[10, 11]. The package `quantreg` for the statistical program R can be used for this purpose. A short introduction is given in Section 3 below.

In this paper, we use the quantile regression curves to propose a procedure to build bivariate box plots (BBP). The theoretical BBP are studied in Section 2. The main advantage of this procedure is that, by using conditional distributions, we can choose the exact probability of each region. Three reasonable options are proposed. The empirical bivariate versions are discussed in Section 3. They are obtained by using different techniques to estimate the theoretical regions. Some extensions to the multivariate case are proposed in Section 4. The conclusions are placed in Section 5.

If $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued function with n variables, then $\partial_i f$ denotes the partial derivative of f with respect to its ith variable. Analogously, $\partial_{i,j} f := \partial_i \partial_j f$ and so on. Whenever we use a partial derivative, we tacitly assume that it exists.

2 Theoretical BBP

2.1 Definitions

The Tukey's univariate box plot associated to a sample $X_1, \dots, X_n$ from a random variable X is usually determined by the inverse of the function g_n obtained connecting the points $(X_{i:n}, (i-1)/(n-1))$, $i = 1, \dots, n$ (see, e.g., [8]), where $X_{1:n} \le \cdots \le X_{n:n}$ represent the ordered data. The *box* $[Q_1, Q_3]$ is determined by the empirical quartiles $Q_1 := g_n^{-1}(1/4)$ and $Q_3 := g_n^{-1}(3/4)$. In the middle we plot the empirical median $Me = Q_2 := g_n^{-1}(2/4)$. The *whiskers* are determined by $[L_1, Q_1]$ and $[Q_3, L_2]$ where $L_1 := X_{1:n} = \min(X_1, \dots, X_n)$ and $L_2 := X_{n:n} = \max(X_1, \dots, X_n)$. To detect outliers, the limit points L_1 and L_2 are replaced with $L_1 = Q_1 - 1.5(Q_3 - Q_1)$ and $L_2 = Q_3 + 1.5(Q_3 - Q_1)$ when $X_{1:n} < Q_1 - 1.5(Q_3 - Q_1)$ or $X_{n:n} > Q_3 + 1.5(Q_3 - Q_1)$. The data beyond these limit points are considered as possible outliers.

There are different options to define the *theoretical univariate box plots*. In this case, the Tukey's box plot is replaced by the 5 points method based on the mean (or the median), two moderate quantiles and two extreme quantiles. If F is the distribution function of X, the (upper) quantile (generalized inverse) function is usually defined as

$$q(u) = F^{-1}(u) := \sup\{x : F(x) \le u\}$$

for $u \in (0, 1)$. Other authors prefer to use the so called lower quantiles defined as $q^-(u) := \inf\{x : F(x) \ge u\}$. If F is continuous and strictly increasing, they coincide for $0 < u < 1$. The quantile function q is available in many statistical programs (as R) for the usual probability models. The theoretical median and quartiles can be defined as $me = q_2 = q(1/2)$, $q_1 = q(1/4)$ and $q_3 = q(3/4)$ (other definitions can be considered as well). Hence the *theoretical box* can be defined as $[q_1, q_3]$.

One can also consider different options to define the *theoretical whiskers*. As in the empirical case, the limit points (or fences) can be defined as $l_1 := q_1 - 1.5(q_3 - q_1)$ and $l_2 := q_3 + 1.5(q_3 - q_1)$. The main problem with this definition is that the probabilities $\Pr(X \in [l_1, q_1])$ and $\Pr(X \in [q_3, l_2])$ depend on F. Instead we could consider the quantiles represented by $q_1 - 1.5(q_3 - q_1)$ and $q_3 + 1.5(q_3 - q_1)$ in a normal (Gaussian) model. In this model we have

$$F(q_1 - 1.5(q_3 - q_1)) = 1 - F(q_3 + 1.5(q_3 - q_1)) = 0.003488302.$$

Therefore, we can define the theoretical limit points for a general distribution F as $\ell_1 := q(0.0034883)$ and $\ell_2 := q(1 - 0.0034883) = q(0.9965117)$. With these definitions, we have

$$\Pr(X \in [\ell_1, q_1]) = \Pr(X \in [q_3, \ell_2]) = \frac{1}{4} - 0.0034883$$

for any continuous distribution function F. In the normal model both definitions coincide but in other models l_i and ℓ_i can be different for $i = 1, 2$. Note that

$$\Pr(X < \ell_1) + \Pr(X > \ell_2) = 2 \cdot 0.0034883 = 0.0069766.$$

Therefore, approximately, the 0.7% of the data from X will be classified as (false) outliers, that is, the theoretical box-plot will contain the 99.3% of the data. If we want different coverage probabilities, we can choose other quantiles for the 5 points method.

Now we are ready to propose a procedure to define *theoretical bivariate box plots* (BBP) for a bivariate random vector (X, Y) with an absolutely continuous distribution function $\mathbf{F}$ (this assumption can be relaxed later). To this end we are going to use the 5 points method and conditional distributions. For instance, we can consider the conditional distribution $(Y|X = x)$, that is, we consider Y as a *response* variable and X as an *explanatory* variable. Then we propose the regions defined as

$$R = [a, b] \times [\alpha, \beta], \tag{2.1}$$

where $a < b$ are real numbers and $\alpha, \beta : [a, b] \to \mathbb{R}$ are continuous functions with $\alpha \le \beta$. For this kind of regions we can state the following result.

Theorem 2.1. *If the region R in* (2.1) *is defined by using the quantile functions $\alpha(x) = F^{-1}_{2|1}(u|x)$ and $\beta(x) = F^{-1}_{2|1}(v|x)$ for $F_{2|1}(y|x) := \Pr(Y \le y|X = x)$ and fixed quantiles u and v such that $0 < u < v < 1$, then*

$$\Pr((X, Y) \in R) = (v - u)\Pr(a \le X \le b).$$

Proof. From eq. (2.1)

$$\Pr((X, Y) \in R) = \Pr(a \le X \le b)\Pr(\alpha \le Y \le \beta|a \le X \le b) = p_1 p_2,$$

where we assume $p_1 := \Pr(a \le X \le b) > 0$ (if $p_1 = 0$ the result is trivial),

$$p_2 := \Pr(\alpha \le Y \le \beta|a \le X \le b) = \int_a^b \Pr(\alpha(x) \le Y \le \beta(x)|X = x)\frac{f_1(x)}{p_1}dx$$

and f_1 is the probability density function (pdf) of X. By using now that α and β are quantile functions of the conditional random variable $(Y|X = x)$, we get

$$p_2 = \int_a^b (v - u)\frac{f_1(x)}{p_1}dx = (v - u)\int_a^b \frac{f_1(x)}{p_1}dx = v - u$$

which concludes the proof. □

The preceding theorem is the key tool for defining the theoretical bivariate box plots. Note that here we can use the copula approach to determine the conditional quantile functions $F^{-1}_{2|1}$ (see, e.g., [19, p. 217]). The result can be stated as follows.

Proposition 2.1. *If C is the copula function of (X, Y), then the conditional (or regression) quantile function $q_{2|1}(v|x) := F^{-1}_{2|1}(v|x)$ of $(Y|X = x)$ can be obtained as*

$$q_{2|1}(v|x) = F_2^{-1}(d^{-1}_{F_1(x)}(v)) \tag{2.2}$$

for $v \in (0, 1)$, where F_2^{-1} is the quantile function of Y, d_u^{-1} is the inverse function of $d_u(v) := \partial_1 C(u, v)$ and we assume $\lim_{v\to 0} \partial_1 C(u, v) = 0$ for all $0 < u < 1$.

Proof. If F_1, F_2 represent the absolutely continuous marginal distribution functions of (X, Y), then from Sklar's theorem, we know that there exists a unique copula function C such that

$$\mathbf{F}(x, y) := \Pr(X \le x, Y \le y) = C(F_1(x), F_2(y))$$

for all x, y. Hence a joint pdf $\mathbf{f}$ of (X, Y) is

$$\mathbf{f}(x, y) := f_1(x)f_2(y)c(F_1(x), F_2(y))$$

for all x, y, where $f_1 = F_1'$ and $f_2 = F_2'$ represent the marginal pdf functions and $c = \partial_{1,2}C$ is the pdf of the copula C. Therefore, the conditional pdf of $(Y|X = x)$ is

$$\mathbf{f}_{2|1}(y|x) = \frac{\mathbf{f}(x,y)}{f_1(x)} = f_2(y)c(F_1(x), F_2(y))$$

for x such that $f_1(x) > 0$. Then the conditional distribution function can be represented as

$$F_{2|1}(y|x) = \int_{-\infty}^{y} \mathbf{f}_{2|1}(z|x)dz = \int_{-\infty}^{y} f_2(z)\partial_{1,2}C(F_1(x), F_2(z))dx = d_{F_1(x)}(F_2(y))$$

where $d_u(v) := \partial_1 C(u, v)$ for $0 < v < 1$, $d_u(0) := 0$ and $d_u(1) := 1$, since we assume $\lim_{v\to 0} \partial_1 C(u, v) = 0$ for all $0 < u < 1$ (see [18] or [19, p. 217]). Therefore 2.2 holds. □

The function $d_u : [0, 1] \to [0, 1]$ defined in the preceding proof is a *distortion function* for all $0 < u < 1$, that is, it is increasing from $d_u(0) = 0$ to $d_u(1) = 1$ (see [18]). Note that d_u only depends on the copula C. In many copula models, it is continuous and strictly increasing in $[0, 1]$. Therefore its inverse function d_u^{-1} can be obtained easily (by using analytical and/or numerical methods). Properties for the conditional quantile functions (also known as Conditional Value at Risk or CoVaR) of different copulas can be seen in, e.g., [2, 3, 9]. Note that here we are fixing one of the infinitely many versions of the conditional distribution $Y|X$. In some cases, this selection might affect the shape of BBP defined below.

The preceding proposition can be used to obtain different quantile curves. For example, the conditional median (or the median regression) curve can be obtained as

$$me_{2|1}(x) := q_{2|1}(1/2|x) = F_{2|1}^{-1}(0.5|x) = F_2^{-1}(d_{F_1(x)}^{-1}(0.5)).$$

If we want to predict Y for a specific value of $X = x$, this curve is a good alternative to classical mean regression curve $m_{2|1}(x) := E(Y|X = x)$. Moreover, it can be used to define a conditional median vector as follows.

Definition 2.1. *The conditional median vector for X and $Y|X$ is defined as*

$$\mathbf{me}_{2|1} = (me_X, me_{2|1}(me_X)),$$

where $me_X := F_1^{-1}(0.5)$ is the median of X and $me_{2|1}$ is the median regression curve of $Y|X$.

The conditional median vector $\mathbf{me}_{1|2}$ for Y and $X|Y$ can be defined in a similar way. It is in general different from $\mathbf{me}_{2|1}$ (some examples will be given later) and both are different from the mean vector (or the conditional mean vectors, defined similarly).

Analogously, the conditional (theoretical) quartile functions are defined by

$$q_{2|1}^{(i)}(x) := q_{2|1}(i/4|x) = F_{2|1}^{-1}(i/4|x) = F_2^{-1}(d_{F_1(x)}^{-1}(i/4))$$

for $i = 1, 2, 3$. Of course, $q_{2|1}^{(2)} = me_{2|1}$.

As mentioned above, the function d_u only depends on the copula. For many copulas, we can obtain an explicit expression for d_u^{-1}. For example, the explicit expression for Archimedean copulas was given in [19, p. 218]. If we cannot obtain an explicit expression, we can use numerical methods to approximate d_u^{-1}. If we do not want to use copulas, we can choose models with known conditional distributions (e.g. a normal distribution) or models built by using specific conditional distributions (see [1]).

The regions given in eq. (2.1) with appropriate conditional quantile functions α and β can be used to define the theoretical BBP. We propose here three reasonable options. In all of them, we use the 5 points method applied to the marginal distribution of X and the conditional distribution of $Y|X$. Other options can be considered as well.

Definition 2.1: BBP, option 1. In this option, the central region R_{cc} is obtained with $p_1 = p_2 = 1/2$. In this case, the interval $[a, b]$ coincides with the theoretical (univariate) box for X (i.e. $a = q_1(X) = F_1^{-1}(1/4)$ and

$b = q_3(X) = F_1^{-1}(3/4)$) and the curves α and β coincide with the first and the third quartile regression curves (i.e., $\alpha = q_{2|1}^{(1)}$ and $\beta = q_{2|1}^{(3)}$). In the middle of the region we can plot the median regression curve $me_{2|1} = q_{2|1}^{(2)}$ (see the examples below). Hence, from Theorem 2.1, the region R_{cc} will contain (approximately) the 25% of the data since $p_1 p_2 = 1/4$.

Next we can define the central-top R_{ct} and central-bottom R_{cb} regions with $p_1 = 1/2$ and $p_2 = 1/4 - 0.0034883 = 0.2465117$. Both are obtained with $[a, b]$ as above (i.e. the box of X). Then R_{ct} is obtained with $\alpha = q_{2|1}^{(3)}$ and $\beta(x) = q_{2|1}(0.9965117|x)$ and R_{cb} is obtained with $\alpha(x) = q_{2|1}(0.0034883|x)$ and $\beta = q_{2|1}^{(1)}$ (i.e. the limit points of the theoretical whiskers of $Y|X$). Then $p_1 p_2 = 0.1232559$.

Analogously, the left-central R_{lc} and right-central R_{rc} regions are obtained with $p_2 = 1/2$ and $p_1 = 1/4 - 0.0034883$. The first one is obtained with $a = \ell_1$ and $b = q_1(X)$ and the second with $a = q_3(X)$ and $b = \ell_2$ (i.e. the limit points of the theoretical whiskers of X). Then, in both cases, we take $\alpha = q_{2|1}^{(1)}$ and $\beta = q_{2|1}^{(3)}$. Therefore $p_1 p_2 = 0.1232559$ as in the preceding regions. The four regions $R_{ct}, R_{cb}, R_{lc}, R_{rc}$ together will contain (approximately) the 49.30234% of the data.

Finally, we define in a similar way the four corner regions $R_{lt}, R_{lb}, R_{rt}, R_{rb}$ (left-top, left-bottom, right-top and right-bottom) with $p_1 = p_2 = 1/4 - 0.0034883$ and $p_1 p_2 = 0.06076802$. Hence, the four corner regions together contain the 24.30721% of the data. The data out of these nine regions can be considered as possible outliers and $1 - 0.9860955 = 0.01390453$ is the probability of a false outlier (i.e. approximately the 1.4% of the data from (X, Y) will be considered as possible outliers).

Some examples of these regions are given below (see Figures 1 and 5, left). There we add some simulated data (right plots). We also perform fit tests based on the Chi-squared distribution and these regions.

Definition 2.2: BBP, option 2. If we want a central region R_{cc} containing the 50% of the data, we can choose $p_1 = p_2 = 1/\sqrt{2} = 0.7071068$. However, in this case, the interval $[a, b]$ does not coincide with the theoretical (univariate) box for X, since

$$a = F_1^{-1}\left(\frac{1 - 1/\sqrt{2}}{2}\right) = F_1^{-1}(0.1464466) < q_1(X) = F_1^{-1}(0.25)$$

and

$$b = F_1^{-1}\left(1 - \frac{1 - 1/\sqrt{2}}{2}\right) = F_1^{-1}(0.8535534) > q_3(X) = F_1^{-1}(0.75).$$

The same happen with the curves α and β which do not coincide with the quartile regression curves. They are defined as $\alpha(x) = q_{2|1}(0.1464466|x)$ and $\beta(x) = q_{2|1}(0.8535534|x)$. In the middle of the region, as in option 1, we plot the median regression curve $me_{2|1}$. Hence, from Theorem 2.1, the region R_{cc} defined in this way will contain (approximately) the 50% of the data (since $p_1 p_2 = 1/2$).

To complete this option we just need to define the other regions as in option 1 but changing the coverage probabilities to get simple regions. For example, a reasonable option could be to define the limit points as $\ell_1 = F_1^{-1}(0.05)$ and $\ell_2 = F_1^{-1}(0.95)$. In this case $\Pr(\ell_1 \le X \le \ell_2) = 0.9$. We can do the same with the conditional distributions by choosing the quantile functions $q_{2|1}(0.05|x)$ and $q_{2|1}(0.95|x)$. Hence the nine regions of the BBP will contain (approximately) the 81% of the data and the 9% of the data will be classified as possible (false) outliers. The four regions $R_{ct}, R_{cb}, R_{lc}, R_{rc}$ will contain the 6.819805% of the data and all together the 27.27922%. Analogously, the corner regions $R_{lt}, R_{lb}, R_{rt}, R_{rb}$ will contain the 0.9301948% and all together the 3.720779%. Some examples are given below (see Figures 2 and 5, right).

Definition 2.3: BBP, option 3. Another option to get a central region R_{cc} containing the 50% of the data, is to choose $p_1 = 1$ and $p_2 = 0.5$, that is, $a = -\infty$, $b = \infty$, $\alpha(x) = q_{2|1}(0.25|x)$ and $\beta(x) = q_{2|1}(0.75|x)$. This option is not new since it is a popular option in quantile regression plots (see [10, 11]) and the resulting region can be considered as a 50% confidence band for the predictions for Y obtained by using the median regression curve. Note that they also coincide with the union of all the theoretical boxes for $Y|X = x$ for $x \in \mathbb{R}$.

Here we also have several options for the whiskers. The most habitual one (**option 3.1**) in quantile regression plots is to use $q_{2|1}(0.05|x)$ and $q_{2|1}(0.95|x)$ for the lower and upper limits (as in option 2 above). With this choice the three regions all together will contain the 90% of the data (as in option 2) and they can also

be considered as a 90% confidence band for $Y|X$. The 10% of the data will be considered as false outliers. Instead (**option 3.2**) we can use the limits for the theoretical whiskers (used in option 1 above) obtained with the limit curves $q_{2|1}(0.0034883|x)$ and $q_{2|1}(0.9965117|x)$. This is equivalent to use the theoretical box plots for $Y|X = x$ at every point $x \in \mathbb{R}$. In all these choices we maintain $a = -\infty$ and $b = \infty$. Thus, the main advantage of option 3 is that we just have three regions (as in the univariate box plots). The main disadvantage is that we will not be able to detect outliers due to extremes values of X but close to the median regression curve (see Figure 3) since we are not using a univariate box plot for X to determine $[a, b]$.

2.2 Main properties

The BBP defined above have the following properties. The proofs are immediate from Theorem 2.1 and the properties of the regression quantile functions.

Proposition 2.2. *The BBP are equivariant under monotone increasing transformations, that is, if h_1, h_2 are increasing functions and $R = (a, b)\times(\alpha, \beta)$ is a $p = p_1p_2$ confidence region for (X, Y) based on quantile regression curves, then*

$$R^* = (h_1(a), h_1(b)) \times (h_2(\alpha(h_1^{-1})), h_2(\beta(h_1^{-1})))$$

is a $p = p_1p_2$ confidence region for $(h_1(X), h_2(Y))$.

Proof. From (2.2), the quantile regression curves of $h_2(Y)|h_1(X) = x$ satisfy

$$q_{h_2(Y)|h_1(X)}(v|x) = h_2(q_{Y|X}(v|h_1^{-1}(x)))$$

for all increasing functions h_1, h_2 and $0 < v < 1$. This property is based on Proposition 2.1 and the well known fact that the copula (and so d_u) does not change under increasing transformations. Therefore, R^* is a $p = p_1p_2$ confidence region for $(h_1(X), h_2(Y))$ based on quantile regression curves. □

In particular, the BBP proposed above are equivariant under scale-location transformations with a positive scale parameter (i.e., when we change X and Y with $aX + b$ and $cY + d$ for $a, c > 0$).

However, the BBP are not in general equivariant for interchanging X and Y (since they are based on quantile regression curves). If we want equivariant plots in this sense we should use bagplots (see [20]) or the confidence regions based on the multivariate Chebyshev's inequality (see [16]).

Proposition 2.3. *The coverage probabilities and the expected values for the regions of the BBP in the different options considered above are distribution-free.*

The proof is immediate from Theorem 2.1. The definitions given above of the different regions can be modified to get the desired (theoretical) coverage probabilities.

By prolonging the curves/lines used to define the 9 regions proposed in options 1 and 2 above, the possible outliers can be also classified in 16 different regions that indicate how extreme they are (the more extreme ones will be that included in the four corner regions).

2.3 Two theoretical examples

Let us see how to compute the theoretical BBP in two different models. A normal (Gaussian) model (next example) and an exponential model with a Clayton copula (Example 2.2).

Example 2.1. A typical (relevant) case is to consider a normal model. In this case the conditional distributions are known and the quantile regression curves are straight lines (this fact will simplify the estimation procedure considered in the next section). For example, we can consider two standard normal distributions

with correlation $\rho = 0.7$. Then the conditional distribution $(Y|X = x)$ has a normal distribution with mean

$$m_{2|1}(x) = E(Y|X = x) = \rho x = 0.7x$$

and a constant variance $\sigma^2 = Var(Y|X = x) = 1 - \rho^2 = 1 - 0.49 = 0.51$ (see, e.g., [14, p. 63]). Moreover, the conditional median regression curve coincides with the regression curve, that is,

$$me_{2|1}(x) = m_{2|1}(x) = \rho x.$$

In this case, the conditional medians coincide and are equal to the mean vector, that is, $\mathbf{me}_{2|1} = \mathbf{me}_{1|2} = (0, 0)$. Analogously, the conditional quantile function is

$$q_{2|1}(v|x) = \rho x + q_{norm}(v, 0, \sigma),$$

where $q_{norm}(v, 0, \sigma)$ is the quantile function of a normal distribution with zero mean and standard deviation $\sigma = \sqrt{0.51} = 0.7141428$.

The first and third quartiles of X (and Y) satisfy $-q_1 = q_3 = 0.6744898$ and the limits for the whiskers $-\ell_1 = \ell_2 = 2.697959$. As mentioned above, in this case we have $\ell_1 = q(0.0034883) = l_1 = q_1 - 1.5(q_3 - q_1)$ and $\ell_2 = q(0.9965117) = l_2 = q_3 + 1.5(q_3 - q_1)$.

The regions for the BBP in option 1 for this example are plotted in Figure 1, left. The central region R_{cc} (blue) is determined by the four blue lines. The black line plotted in the middle represents the conditional median curve $me_{2|1}$ (in this model it is also the regression curve). The blue and red lines are used to determine the other regions. For example, the central-top region R_{ct} (top red) is determined by the vertical blue lines (the box of X), the top blue line and the top red line (the top whisker of $Y|X$). The other regions are determined in a similar way. The blue region contains the 25% of the data, each red region the 12.32559% (all together the 49.30234%) and each grey region the 6.076802% (all together the 24.30721%). The conditional median $\mathbf{m}_{2|1} = (0, 0)$ point is represented by the + symbol. The data outside these nine regions can be considered as possible outliers. Note that these outliers can also be classified in different regions. For example, a data on the left (i.e. with $X < \ell_1$) is just an outlier in the box plot of X. If it is also in the top (i.e. $Y > \beta$), then it is also an outlier in the conditional distribution $Y|X$.

In the right plot of Figure 1, we add $n = 100$ simulated data from this model. With these data we obtain the following estimations: $Me(X) = -0.16317$ (of $me = 0$), $Q_1(X) = -0.72181$ (of $q_1 = -0.6744898$) and $Q_3(X) = 0.55324$ (of $q_3 = 0.6744898$). In this plot we have two possible (false) outliers. The first one (right-top) is also an outlier in the box plot of X since $X = 3.134841 > L2 = Q_3(X) + 1.5(Q_3(X) - Q_1(X)) = 2.465815$. The second one (bottom) with values $(-0.8450186, -2.621403)$ is not detected as an outlier for X. However, it is classified as a possible (false in this case) outlier in the conditional distribution of $Y|X$ since $Y < \alpha(X)$ (by a small margin). The counts of the data in each region are given in Table 1. There, we also provide the expected data values in each region obtained as np_1p_2. Thus, we can perform a Chi-squared fit test with the statistic

$$T = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

where O_i represent the observed data in each region, E_i the expected data and m the number of regions. We can consider $m = 10$ regions by including a common region for the outliers. It is well known that, with the BBP obtained by using the correct (exact) distribution (null hypothesis), the asymptotic distribution of T (when $n \to \infty$) is a χ^2_{m-1}. The P-value for this test is $P - value = \Pr(T > T_o)$, where T_o is the observed value for T. The approximation with the Chi-squared distribution gives a good approximation if all the expected values E_i are greater than 5 (see, e.g., [7]). So, if it is necessary, some regions can be joined. If the distribution contains k unknown parameters and they are estimated with a maximum likelihood method, then the asymptotic distribution of T is χ^2_{m-k-1}. With these simulated data and the ten regions of option 1, we obtain (see Table 1)

$$T_o = \frac{(4 - 6.076802)^2}{6.076802} + \cdots + \frac{(2 - 6.076802)^2}{6.076802} + \frac{(2 - 1.39045)^2}{1.39045} = 7.679805$$

and $P - value = \Pr(T > 7.679805) = 0.5667081$, where T has (approx.) a χ^2_9 distribution. Hence, as expected, we can confirm that the data come from this model (we have an expected value less than 5 but the P-value is

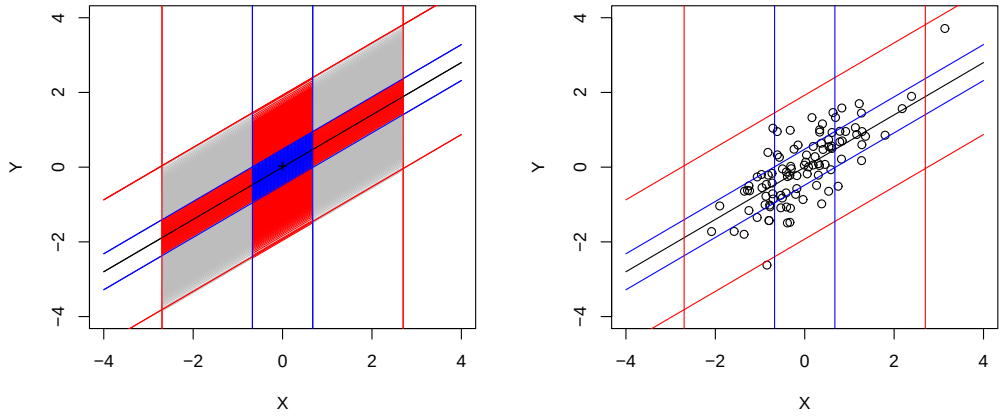

Figure 1: BBP in option 1 (left) for the normal model considered in Example 2.1. The blue region contains the 25% of the data, the four red regions the 49.30234% and the four grey regions the 24.30721%. In the right plot we add 100 simulated data from this model.

large enough to confirm this decision). Note that with the BBP in option 1 (and the correct distribution), the expected values in the regions will only depend on n for any joint distribution $\mathbf{F}$ (i.e. they are distribution-free).

Table 1: Observed and expected data in the regions determined by the BBP in option 1 for the normal model in Example 2.1.

$O_i\|E_i$	Left	Central	Right	Sum
Top	4|6.076802	12|12.32558	4|6.076802	20|24.47919
Central	17|12.32558	30|25	13|12.32558	60|49.65117
Bottom	6|6.076802	10|12.32558	2|6.076802	18|24.47919
Sum	27|24.4796	52|49.65117	19|24.4796	98|98.60955

The BBP in options 2 and 3.1 can be obtained in a similar way. They are plotted in Figures 2 and 3, respectively. They can also be used to perform fit tests. As expected, in the BBP of option 2 we have several (14) false outliers but we just have an extreme outlier (right corner), the point with values (3.134841, 3.711933) which is out of both the regions determined by the right vertical red line (ℓ_2) and the top red line (β). Instead, the central (blue) region R_{cc} contains 55 data (we expect 50). In the BBP of option 3.1 we also have many false outliers, 5 above the top red line and 4 below the bottom red line (we expect 10). In this case the more extreme data is the point with values (−0.8450186, −2.621403) because it is the farthest point from the median regression estimation

$$\hat{Y} := me_{2|1}(-0.8450186) = 0.7(-0.8450186) = -0.591513.$$

If we use the limits in option 3.2 (red dashed lines in Figure 3), then this point is the unique outlier. In this case, the point in the right-top corner is not detected as an outlier (remember that it was classified as outlier in the univariate box plot of X). The same will happen with extremes values of X or Y close to the median regression curve (black line).

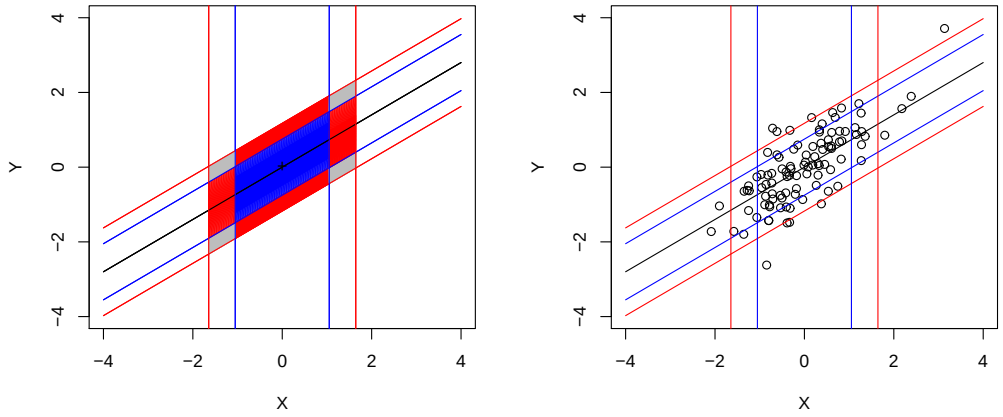

Figure 2: BBP in option 2 (left) for the normal model considered in Example 2.1. The blue region contains the 50% of the data, the four red regions the 27.27922% and the four grey regions the 3.720779%. In the right plot we add 100 simulated data from this model.

Example 2.2. There are more models (copulas) whose quantile regression curves are straight lines (see, e.g., Example 5.24 in [19, p. 218]). However, in this second example, we want to choose a model without this property (this fact will hinder the estimation procedures studied in the next section). Thus, we consider the following Clayton copula

$$C(u,v) = \frac{uv}{u+v-uv},$$

which induces a positive dependence between X and Y. As C is the distribution function of $U = F_1(X)$ and $V = F_2(Y)$ we can just consider the BBP for (U, V). The similar plots for (X, Y) will be obtained with the transformations $X = F_1^{-1}(U)$ and $Y = F_2^{-1}(V)$ for any continuous marginal distributions F_1 and F_2 and Proposition 2.2.

As the Clayton copulas belong to the wide family of Archimedean copulas, the quantile regression curves can be obtained from the expression given in Example 5.25 of [19, p. 218]. For the copula C given above we can also use a direct calculation as follows. The distortion function d_u for the conditional distribution $Y|X$ is

$$d_u(v) := \partial_1 C(u,v) = \frac{v^2}{(u+v-uv)^2}$$

for $0 \le v \le 1$ and $0 < u < 1$. Note that d_u is the distribution function of $V|U = u$ for $0 < u < 1$. It is plotted in Figure 4, left, for different values of u. Its inverse function is

$$d_u^{-1}(v) = \frac{u\sqrt{v}}{1-(1-u)\sqrt{v}}$$

for $0 \le v \le 1$ and $0 < u < 1$. Hence, from eq. (2.2), the quantile regression curves are

$$q_{2|1}(y|x) = F_2^{-1}\left(d_{F_1(x)}^{-1}(y)\right) = F_2^{-1}\left(\frac{F_1(x)\sqrt{y}}{1-(1-F_1(x))\sqrt{y}}\right)$$

for $0 < y < 1$ and x such that $f_1(x) > 0$. In particular, for (U, V), which have uniform marginal distributions in $(0, 1)$, we get

$$q_{2|1}(v|u) = d_u^{-1}(v) = \frac{u\sqrt{v}}{1-(1-u)\sqrt{v}}$$

for $0 < u < 1$. Thus, for example, the median regression curve for $V|U = u$ is

$$me_{2|1}(u) := q_{2|1}(0.5|u) = \frac{u\sqrt{0.5}}{1-(1-u)\sqrt{0.5}}$$

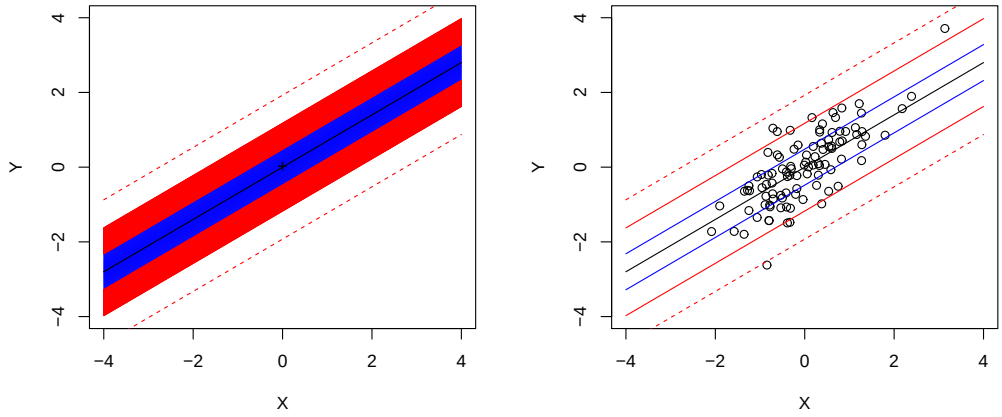

Figure 3: BBP in option 3.1 (left) for the normal model considered in Example 2.1. The blue region contains the 50% of the data and the two red regions the 40%. The dashed lines represent the BBP in option 3.2. In the right plot we add 100 simulated data from this model.

for $0 < u < 1$. It is plotted in Figure 4, right (black line) jointly with the first and third quartile functions (blue lines) and the 5 and 95 percentile functions (red lines). These curves determine the BBP in option 3.1. Note that C is symmetric (i.e. (U, V) has an exchangeable distribution) but that these regions are not. The conditional median vector of $V|U$ is

$$\mathbf{m}_{2|1} = (0.5, me_{2|1}(0.5)) = \left(0.5, 0.5\sqrt{0.5}1 - 0.5\sqrt{0.5}\right) = (0.5, 0.5469182).$$

By the symmetry of the model, we have $\mathbf{m}_{1|2} = (0.5469182, 0.5)$ and so, in this case, they do not coincide. Note that they are also different from the mean and median vectors which coincide in the point $(E(U), E(V)) = (me(U), me(V)) = (0.5, 0.5)$.

The quantile regression curve $q_{2|1}$ can also be used to get the BBP in options 1 and 2. They are plotted in Figure 5 jointly with 100 simulated data from this model. These simulated data were obtained by using the *conditional standard method*. Thus we first generate a 100 random data $u_1, \ldots, u_{100}$ from U (i.e. 100 random data in $(0, 1)$), and then we get one random data from $V|U = u_i$ by using $q_{2|1}$ for $i = 1, \ldots, 100$ (i.e. with the inverse transform method). As in the preceding example the BBP could be used to perform fit tests for the copula (when we are not sure about the exact copula).

If, for example, the marginal distributions are standard exponentials, that is, $F_1(x) = F_2(x) = 1 - e^{-x}$ for $x \geq 0$, then by applying the inverse transform $F_i^{-1}(y) = -\ln(1 - y)$ for $i = 1, 2$, we obtain the data and the BBP (options 1 and 2) plotted in Figure 6.

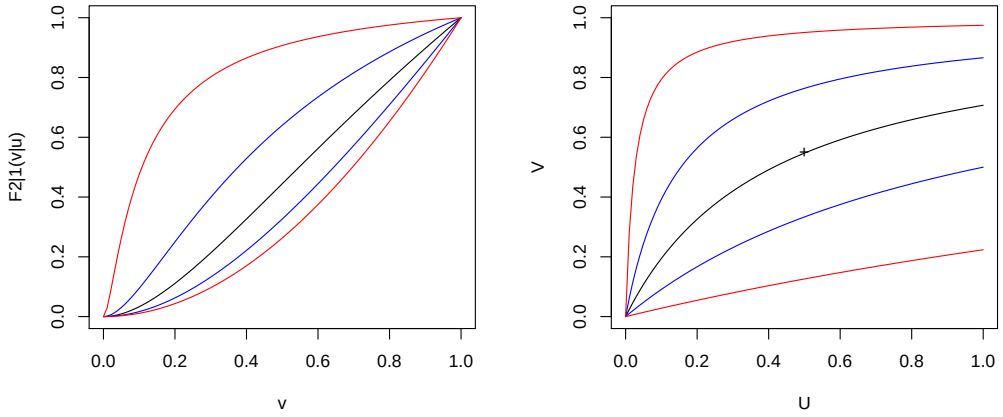

Figure 4: Conditional distribution functions (left) of $V|U = u$ for the Clayton copula in Example 2.2 and $u = 0.5, 0.25, 0.75, 0.05, 0.95$ (black, blue top, blue bottom, red top, red bottom). Quantile regression curves $q_{2|1}(v|u)$ (right) for $V|U = u$ for $v = 0.5, 0.25, 0.75, 0.05, 0.95$ (black, blue bottom, blue top, red bottom, red top). The + point represents the conditional median $\mathbf{m}_{2|1} = (0.5, 0.5469182)$.

3 Practical BBP

As mentioned at the beginning of Section 2, the univariate box plots are an empirical tool. Therefore, it is very important to obtain the practical BBP associated to the theoretical ones proposed in the preceding section. To this end we can maintain the definitions (options) and just estimate the theoretical regions or to modify the definitions by using the sample. In this section $(X_1, Y_1), \ldots, (X_n, Y_n)$ represent a sample of $i.i.d.$ random vectors from (X, Y).

As the theoretical BBP in option 1 of the preceding section is based on the univariate box plot for X, in practice we have two options for the limit points ℓ_1 and ℓ_2.

Option 1.a (five points method): Here we can maintain the definitions and use the function g_n obtained from the data $X_1, \ldots, X_n$ (see the first paragraph of Section 2) to estimate ℓ_1 and ℓ_2 by $\hat{\ell}_1 := g_n^{-1}(0.0034883)$ and $\hat{\ell}_2 := g_n^{-1}(0.9965117)$.

Option 1.b (Tukey's method): Alternatively, we can just use the empirical (Tukey's) box plot for X, that is, we can use the limits L_1 and L_2 for the whiskers as vertical limits for the BBP.

Note that this second option cannot be applied to the conditional distributions $Y|X = x$ since, in continuous models, we just have a data (X_i, Y_i) for each value $x = X_i$ for $i = 1, \ldots, n$ and we do not have data for the other values of x. Hence we need to estimate the conditional quantile function $q_{2|1}$ and use the quantile limits proposed in options 1-3. Fortunately, we have several techniques available in the literature (and in some statistical programs) for this purpose. Let us see some of them.

3.1 Linear quantile regression

The linear quantile regression was proposed by Koenker and Basset in [11] (see also [10]) as an alternative to the well known linear regression model. The key idea is to replace conditional means by conditional medians. It is well known that the empirical median $\hat{m}_X = g_n^{-1}(0.5)$ of $X_1, \ldots, X_n$ minimizes the mean absolute error $MAE(x) = \frac{1}{n}\sum_{i=1}^{n} |X_i - x|$ while the empirical mean $\overline{X}$ minimizes the mean squared error $MSE(x) = \frac{1}{n}\sum_{i=1}^{n}(X_i - x)^2$.

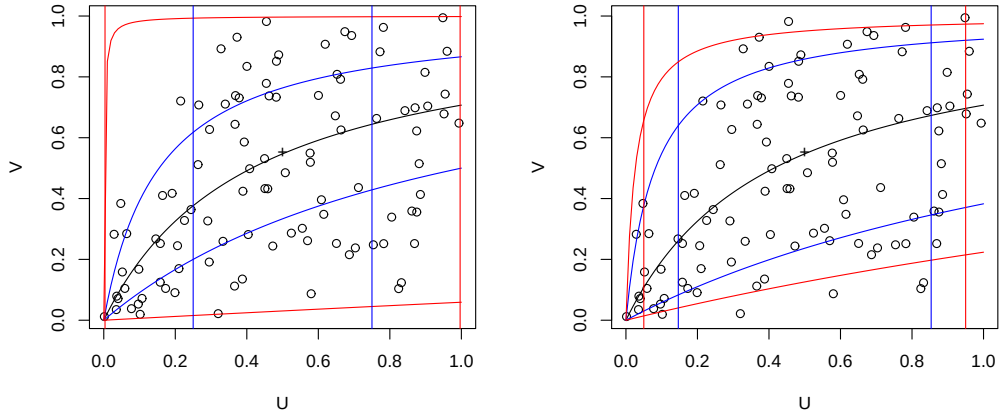

Figure 5: BBP in options 1 (left) and 2 (right) for the Clayton copula in Example 2.2 with 100 simulated data.

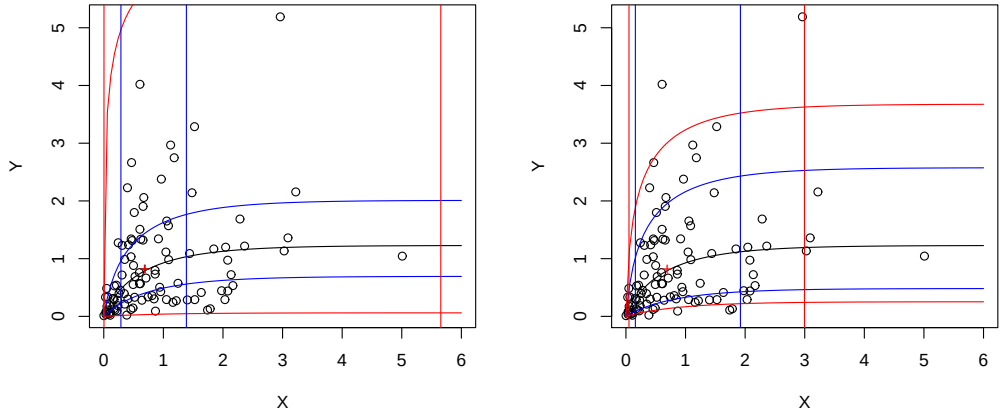

Figure 6: BBP in options 1 (left) and 2 (right) for the exponential model with the Clayton copula in Example 2.2 and 100 simulated data.

Thus, in a linear median regression model, we consider the linear function $y = a + bx$ which minimizes the

$$MAE(a, b) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - a - bX_i|$$

instead of the MSE used in the linear regression model. The empirical linear median regression curve is then defined as $\hat{m}_{2|1}(x) = \hat{a} + \hat{b}x$, where $\hat{a}, \hat{b} \in \mathbb{R}$ are the (one) solutions of this minimization problem. This definition can be extended when the random variable X is replaced with a random vector (see [11]).

Analogously, the qth sample quantile $\hat{x}_q$ of X is defined as the (one) solution of

$$\min_x \sum_{i:X_i>x} q|X_i - x| + \sum_{i:X_i<x} (1-q)|X_i - x|$$

for $0 < q < 1$. Note that this definition does not coincide with that used above (and in the default method of R) to get the quartiles Q_1 and Q_3 and to built the univariate box plots (see the first paragraph of Section 2 or [8]). This minimization problem can be used to define the *linear regression quantile* function as $\hat{q}_{2|1}(v|x) = \hat{a}_v + \hat{b}_v x$, where $\hat{a}_v, \hat{b}_v \in \mathbb{R}$ are the (some) solutions of

$$\min_{a,b} \sum_{i:Y_i>aX_i+b} v|Y_i - a - bX_i| + \sum_{i:Y_i<aX_i+b} (1-v)|Y_i - a - bX_i|$$

for a fixed $v \in (0, 1)$. These linear regression quantiles can be obtained in R with the package: `quantreg` (see [13]). After installing this package, the command: `rq(Y~X,v)` provides the estimated linear regression quantiles for $0 < v < 1$, where `X` and `Y` are two columns containing the paired data (X_i, Y_i).

Clearly, $\hat{q}_{2|1}(v|x)$ can be considered as an estimator of the theoretical curve $q_{2|1}(v|x)$ defined in the preceding section. Of course, it will be a better estimator if the theoretical curves are straight lines.

Thus, if we use the data in Example 2.1 (obtained from a normal model), we obtain the BBP (option $1.b$) in Figure 7. The continuous lines represent the theoretical regression quantile functions (plotted also in Figure 1) and the dashed lines the associated estimations. The symbols + represent the conditional median vector (black) and its estimation (red). The theoretical median (and mean) regression curve $me_{2|1}(x) = 0.7x$ (continuous black line) is estimated with the median regression line $\hat{me}_{2|1}(x) = 0.05705845+0.76947487x$ (dashed red line). In this case it can also be estimated with the linear regression line $\hat{m}_{2|1}(x) = 0.02156 + 0.82231x$ (dashed green line). In the right, we just plot the empirical BBP and we add the data. Note that the worst estimation is obtained in the lower quantile function (bottom red lines) which fits to the lower data. That is because the lower dashed red line is a regression quantile with probability 0.0035 and so it is obviously attracted by (and even goes through) the lower outliers. As a result, the fences might not identify these outliers (but they remark them as possible outliers). Also note that the greater data of X is considered as an outlier and so it is replaced as the upper limit of X with $L_2 = Q_3 + 1.5(Q_3 - Q_1) = 2.465815$ (right red dashed vertical lines). It is a good estimation of $\ell_2 = q_{norm}(0.9965117, 0, \sigma) = 2.697959$ (right red vertical line). The dashed lines can be used to define the nine empirical regions $\hat{R}$ as in Section 2. The counts of the data in these regions are given in Table 2. As they depend on the data, some regions may contain more data than the theoretical ones (for this sample). However, if they are used to study the data in an independent sample, the expected values should be the same.

Analogously, if we use the data in Example 2.2 (obtained from a Clayton copula), then we obtain the BBP (option $1.b$) in Figure 8 for uniform (left) and exponential (right) marginals. The purpose of this figure is to show that the estimated quantile regression lines can be quite far from the theoretical ones when they are not straight lines. In this case they are not so bad in the left plot. Note that in the right plot we have five outliers in X due to the fact that the exponential model is quite far from the normal model. So, in the classic univariate box (Tukey's) plot of X (vertical lines), the maximum value has been replaced with $L_2 = Q_3 + 1.5(Q_3 - Q_1) = 2.56428$ (detecting five false outliers). In these cases (far from the normal model), it is better to use option $1.a$ (i.e. the empirical 5 points method). Moreover, with this option, we maintain the expected values in the regions. The counts for the empirical regions in the left plot are in Table 3. They do not coincide with the counts for the regions in the right plot due to the change in the right limit of X.

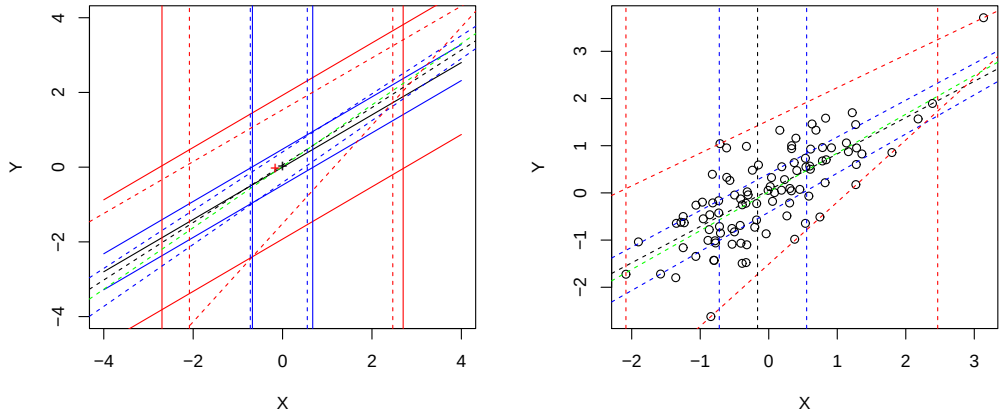

Figure 7: Empirical BBP (dashed lines) in option 1.*b* for the data in Example 2.1. The continuous lines represent the theoretical BBP and the dashed green lines the empirical regression line.

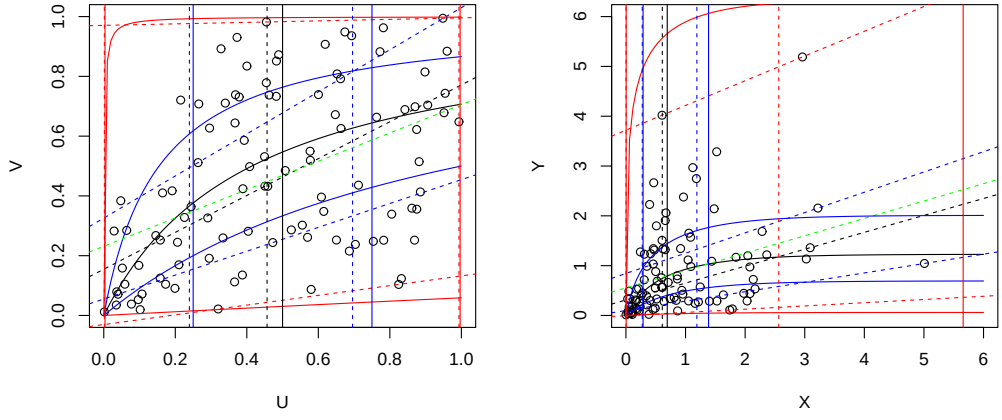

Figure 8: Empirical BBP (dashed lines) in option 1.*b* for the data in Example 2.2 with uniform (left) and standard exponential (right) marginals. The continuous lines represent the theoretical BBP and the dashed green lines the empirical regression line.

Table 2: Observed and expected values for the data in Example 2.1 and the regions determined by the empirical BBP in option 1.

$O_i\|E_i$	Left	Central	Right	Sum
Top	7|6.076802	11|12.32558	6|6.076802	24|24.47919
Central	12|12.32558	27|25	12|12.32558	51|49.65117
Bottom	6|6.076802	12|12.32558	5|6.076802	23|24.47919
Sum	25|24.4796	50|49.65117	23|24.4796	98|98.60955

Table 3: Observed and expected values for the data from (U, V) in Example 2.2 and the regions determined by the empirical BBP in option $1.b$.

$O_i\|E_i$	Left	Central	Right	Sum
Top	2|6.076802	19|12.32558	3|6.076802	24|24.47919
Central	14|12.32558	23|25	12|12.32558	49|49.65117
Bottom	8|6.076802	7|12.32558	8|6.076802	23|24.47919
Sum	24|24.4796	49|49.65117	23|24.4796	96|98.60955

3.2 Non-linear quantile regression

The basic theory for the non-linear quantile regression can be seen in [12]. They are also included in the R-package: `quantreg`. Of course, this method is especially useful for models with non-linear regression curves (as the model studied in Example 2.2). However, it can also be applied to models with linear quantile regression curves. Note that the linear model can also be applied to linear expressions based on X. For example, we can obtain quadratic approximations by using X and X^2. One can think that the fits will be better in this case because we have an additional parameter by using $y = a + bx + cx^2$. However, this is not always true. Thus, for example, the quadratic median regression curve obtained with this method will be closer to the data than the linear one but it can be a worse estimation of the theoretical median regression curve. In an extreme case we might add n parameters to get a curve which contains all the data, but this curve will not be a good approximation of the theoretical one. This fact can be observed in Figure 9 where we plot the quadratic estimations obtained from the data in Example 2.1. They can be compared with the linear estimations given in Figure 7. However, if we use the data in Example 2.2, we obtain the BBP in Figure 10 (uniform and exponential marginals). Clearly, in these cases, the quadratic approximations are better than the linear approximations given in Figure 8.

3.3 Empirical estimators

If we prefer to have "empirical estimators" we just need to estimate the conditional median (quantiles) for $Y|X = X_i$. For example, in a **three points method**, we can first sort the data from X obtaining $X_{1:n}, \ldots, X_{n:n}$. Then, to estimate the quartiles of $Y|X = X_{i:n}$, we can consider the three values of the Y, associated with $X_{i-1:n}, X_{i:n}, X_{i+1:n}$. They are called **concomitans** and are usually represented as $Y_{[i-1:n]}, Y_{[i:n]}, Y_{[i+1:n]}$. These values are not necessarily ordered, so we need to sort them and then to estimate the quartiles of $Y|X = X_{i:n}$ with these three ordered values $Q_1^i \le Q_2^i \le Q_3^i$ (the median is estimated with the value in the middle). For the extreme points $X_{1:n}$ and $X_{n:n}$, we can use respectively the points $X_{1:n}, X_{1:n}, X_{2:n}$ and $X_{n-1:n}, X_{n:n}, X_{n:n}$ (or just use two points). For these points we have the following property.

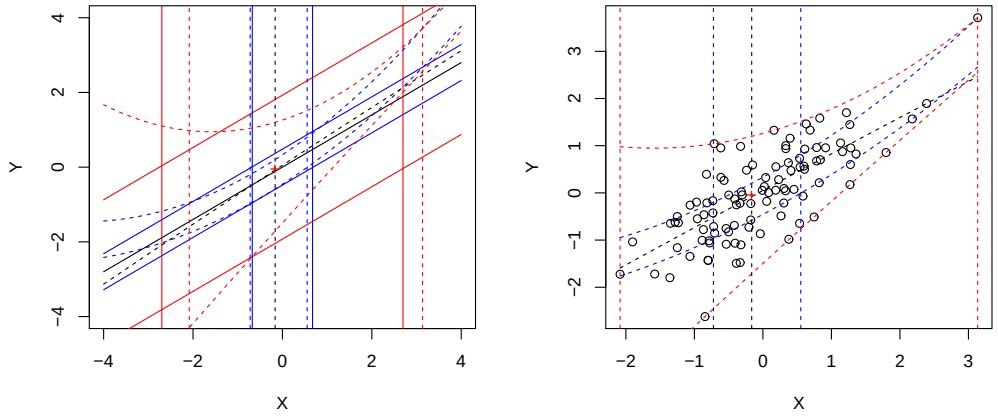

Figure 9: Quadratic empirical BBP (dashed lines) in option $1.b$ for the data in Example 2.1. The continuous lines represent the theoretical BBP.

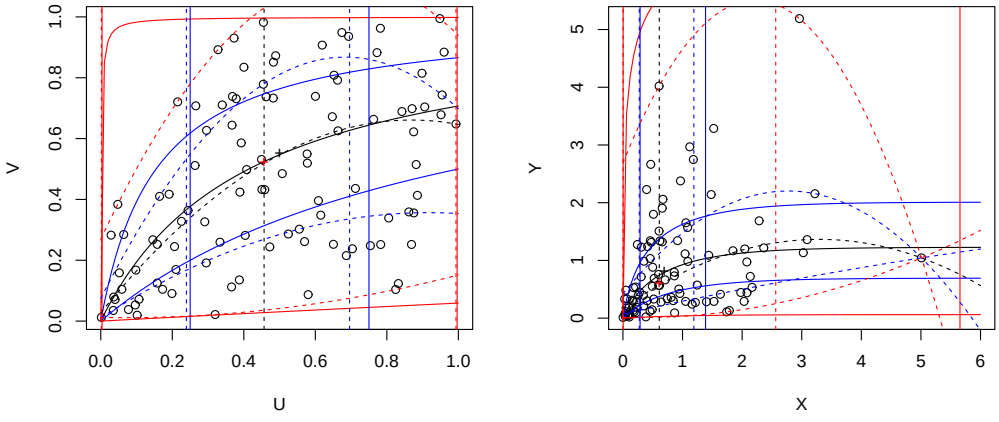

Figure 10: Quadratic empirical BBP (dashed lines) in option $1.b$ for the data in Example 2.2 with uniform (left) and standard exponential (right) marginals. The continuous lines represent the theoretical BBP.

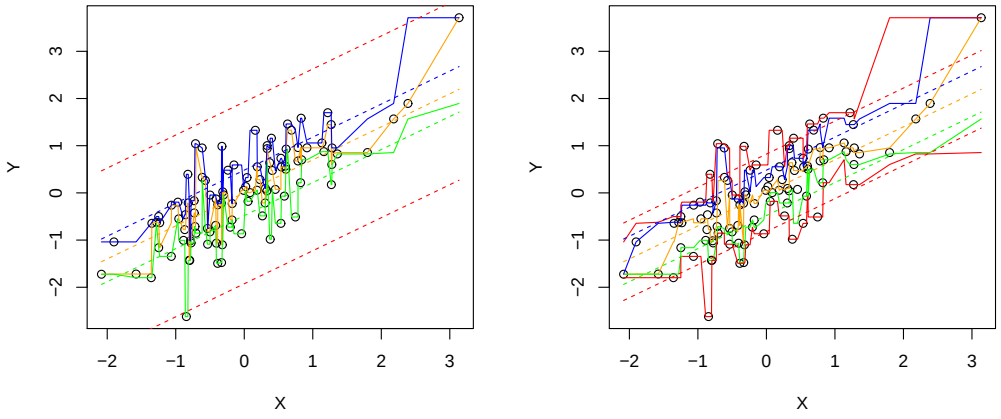

Figure 11: BBP in option 3 for the data in Example 2.1 estimated with the three-points (left) and the seven-points (right) methods. The dashed lines represent the theoretical BBP. The orange lines represent the median curves, the blue lines, the third quartile curves, and the green lines, the first quartile curves. The red lines in the right plot can be used to get the whiskers.

Proposition 3.1. *With the notation introduced above,*

$$\lim_n E(F_{2|1}(Q_j^i|X_{i:n})) = E(U_{j:3}) = j/4$$

for $j = 1, 2, 3$, *where* $U_{1:3}, U_{2:3}, U_{3:3}$ *are the order statistics from a standard uniform distribution.*

Proof. As $n \to \infty$, the three ordered values $Q_1^i \le Q_2^i \le Q_3^i$ can be seen as three ordered values from the conditional distribution $Y|X = X_{i:n}$. Then $U_{1:3} := F_{2|1}(Q_1^i|X_{i:n})$, $U_{2:3} := F_{2|1}(Q_2^i|X_{i:n})$ and $U_{3:3} := F_{2|1}(Q_3^i|X_{i:n})$ can be seen as three ordered values from a standard uniform distribution. Their distributions are well known in the literature. For example, the distribution of $U_{3:3}$ is

$$F_{3:3}(u) = \Pr(\max(U_1, U_2, U_3) \le u) = \Pr(U_1 \le u)\Pr(U_2 \le u)\Pr(U_3 \le u) = u^3$$

for $0 \le u \le 1$, the pdf is $f_{3:3}(u) = 3u^2$ and so

$$E(U_{3:3}) = \int_0^1 u f_{3:3}(u)du = \int_0^1 3u^3 du = \frac{3}{4}.$$

Analogously, it can be proved easily that $E(U_{1:3}) = 1/4$ and $E(U_{2:3}) = 1/2$. □

By applying this method to the data in Example 2.1, we obtain the estimations (continuous lines) plotted in Figure 11, left. As we can see the estimations are not very good. If we have more data we could use a **five-points** or a **seven-points** methods. In the first case, the five closest values can be used to get the box-plots but note that they are biased (except in the case of the median) since $E(U_{j:5}) = j/6$ for $j = 1, 2, 3, 4, 5$. However, in the second, we have $E(U_{j:7}) = j/8$ for $j = 1, \ldots, 7$ and so the second, the fourth and the sixth values of the ordered Ys are unbiased estimators for the quartiles at this point. The first and the seventh values can be used to plot the whiskers. They are plotted in Figure 11, right. There we have changed the theoretical limits in option 1 with the quantile regression curves with $q = 1/8$ and $q = 7/8$ (dashed red lines). In discrete populations we can use all the data with $X = X_i$ for a fixed i. In the following subsections we propose other techniques to get "smooth" estimations based on copula estimations.

3.4 Parametric estimation

As we have seen in Section 2, the conditional regression curves depend on the partial derivative of the copula C and the marginal distributions F_1 and F_2 (see eq. (2.2)). The last ones can be estimated with the empirical (or kernel) distribution functions from X and Y. These estimations $\hat{F}_1$ and $\hat{F}_2$, can be used to transform the data into $U_i = \hat{F}_1(X_i)$ and $V_i = \hat{F}_2(Y_i)$ for $i = 1, \dots, n$, which can be considered as a sample from the copula C. Then we just need to use these data to estimate C.

In this subsection, we assume a parametric form for the copula C, that is, we assume that $C = C_\theta$ for given family of copulas C_θ with an unknown parameter θ. Typically, θ represents the grade of dependency between X and Y. Thus, in Example 2.2, we might assume that the data have a Clayton copula

$$C_\theta(u, v) = \left(\max(0, u^{-\theta} + v^{-\theta} - 1)\right)^{-1/\theta} \tag{3.1}$$

(see, e.g., [19, p. 116]) with an unknown dependence parameter $\theta \in [-1, 0) \cup (0, \infty)$. If $\theta \to 0$, then C_θ goes to the product copula $\Pi(u, v) = uv$ (i.e. X and Y are independent). So, by definition, we take $C_0 := \Pi$.

A typical procedure in copula theory to estimate the dependence parameter θ is to consider a dependence (or concordance) coefficient invariant under monotone increasing transformations. The most usual coefficients are the Kendall's tau τ and the Spearman's rho ρ_S (see [19, p. 158, 167]). Then we estimate this coefficient from the data and we choose the value of θ to get this coefficient in C_θ. For example, the Kendall's tau τ coefficient for the above Clayton copula is $\tau_\theta = \theta/(2 + \theta)$ for $\theta \geq -1$ (see [19, p. 163]). Note that $\tau_\theta \in [-1, 1)$. The empirical Kendall's tau $\hat{\tau}$ is defined (see (5.1.1) in [19, p. 158]) as

$$\hat{\tau} = \frac{c - d}{\binom{n}{2}},$$

where c is the number of concordant and d is the number of discordant pairs of data and $\binom{n}{2}$ is the number of pairs. We say that the pairs (X_i, Y_i) and (X_j, Y_j) $(j \neq i)$ are concordant (resp. discordant) if $(X_i - X_j)(Y_i - Y_j) > 0$ (resp. < 0). If we apply this formula to the data in Example 2.2, we get $\hat{\tau} = 0.339798$. Then by solving

$$\tau_\theta = \frac{\theta}{2 + \theta} \approx 0.339798 = \hat{\tau}$$

we obtain

$$\hat{\theta} := \frac{2\hat{\tau}}{1 - \hat{\tau}} = 1.029376$$

which a good estimation of the theoretical value $\theta = 1$. Finally, under the assumption of a Clayton copula, we use the conditional quantile regression curves of $C_{\hat{\theta}}$ to estimate the theoretical ones and get the practical BBP. A straightforward calculation from eq. (3.1) gives

$$d_u^{-1}(v) = \left(1 - u^{-\theta} + u^{-\theta} v^{-\frac{\theta}{1+\theta}}\right)^{-1/\theta}$$

for $0 < u < 1$ and $0 < v < 1$. Hence, by replacing the unknown θ with $\hat{\theta} = 1.029376$ in these curves, we obtain the estimated curves (dashed lines) plotted jointly with the theoretical ones (continuous lines) in Figure 12, left. Note that the estimations for the curves are very good. The estimations for the quartiles of U (vertical dashed lines) are not so good. However, if we know that the data come from a copula, they can be replaced with the exact values (continuous vertical lines). The estimations in Figure 12, right, for the exponential model are not so good due to the use of the empirical estimators of F_1 and F_2^{-1} in eq. (2.2) (they are especially bad in the top red curve). If we want continuous lines, we can use kernel estimators for F_1 and F_2^{-1}. Remember that the vertical dashed lines correspond to the (empirical) classical univariate box plot of X and that the right limit can be replaced with the maximum value of X which in this sample is $X_{100:100} = 5.008685$ (here it is also considered as an outlier of X).

Moreover, note that the BBP regions can be used to perform fit-tests to confirm the assumed model for the copula. There, if we estimate a dependence parameter by using maximum likelihood, then we should reduce the degrees of freedom in one unit in the (approximate) chi-squared distribution of T.

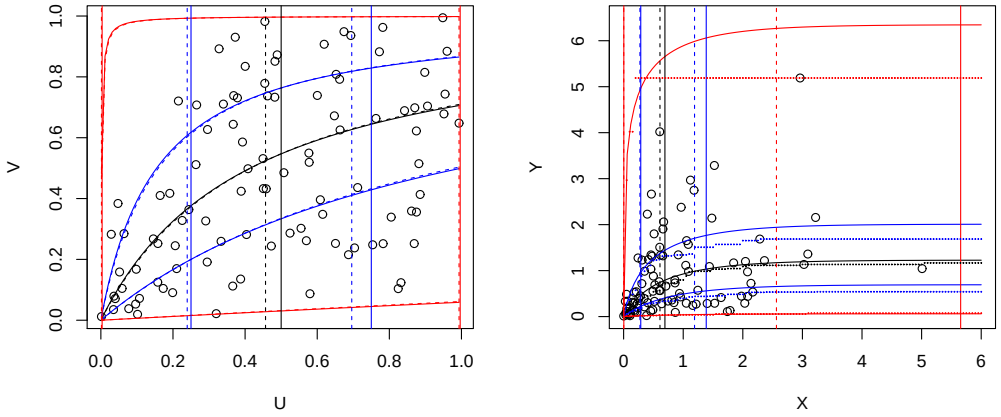

Figure 12: BBP in option 1.*b* for the data in Example 2.2 when we use a parametric estimation for the Clayton copula based on the Kendall's tau coefficient. The dashed lines represent the estimations and the continuous lines the theoretical BBP.

3.5 Non-parametric estimation

If we do not have a parametric model for the copula (as in the preceding subsection), then we have to estimate the copula C and its partial derivative $\partial_1 C$. C can be estimated by using the empirical copula but, this estimator cannot be used to estimate the partial derivative. To this purpose it is better to use a kernel type estimator for C. A survey on the application of this kind of estimators to copula can be seen in [21]. The main problem is that the support of the copula is included in $[0,1]^2$ while the kernel estimators do not have this property. To skip this problem, Nagler [16] proposed to transform the data from (U,V) included in $[0,1]^2$ to new data with support in $\mathbb{R}^2$. This procedure was called the *transform method*. To do that, we can, for example, apply the quantile function of a standard normal univariate distribution. We shall use the following notation. The pdf of this distribution will be denoted by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

and its distribution function by $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$. Hence its quantile (inverse) function is denoted by Φ^{-1}. By applying this transformation to the sample $(U_1,V_1),\dots,(U_n,V_n)$, we obtain a new sample $(X_1,Y_1),\dots,(X_n,Y_n)$ from (X,Y) where $X = \Phi^{-1}(U)$ and $Y = \Phi^{-1}(V)$. Of course, (X,Y) has support $\mathbb{R}^2$, copula C and standard normal marginals. Then we use the new sample to get a kernel estimator for the joint distribution $\mathbf{F}$ of (X,Y). To this end we propose to use a bivariate normal kernel with independent components, i.e.

$$\hat{\mathbf{F}}(x,y) := \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{x-X_i}{h_n}\right)\Phi\left(\frac{y-Y_i}{h_n}\right)$$

with a common bandwidth $h_n > 0$. Then, as $\mathbf{F}(x,y) = C(\Phi(x),\Phi(y))$, the estimator for C is

$$\hat{C}(u,v) := \hat{\mathbf{F}}(\Phi^{-1}(u),\Phi^{-1}(v)) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{\Phi^{-1}(u)-\Phi^{-1}(U_i)}{h_n}\right)\Phi\left(\frac{\Phi^{-1}(v)-\Phi^{-1}(V_i)}{h_n}\right)$$

for $u,v \in (0,1)^2$. Note that $\hat{C}$ is an absolutely continuous bivariate distribution function with support included in $[0,1]^2$ but it is not a copula. Then we can use

$$\hat{d}_u(v) := \partial_1 \hat{C}(u,v) = \frac{1}{nh_n\phi\left(\Phi^{-1}(u)\right)}\sum_{i=1}^{n} \phi\left(\frac{\Phi^{-1}(u)-\Phi^{-1}(U_i)}{h_n}\right)\Phi\left(\frac{\Phi^{-1}(v)-\Phi^{-1}(V_i)}{h_n}\right) \tag{3.2}$$

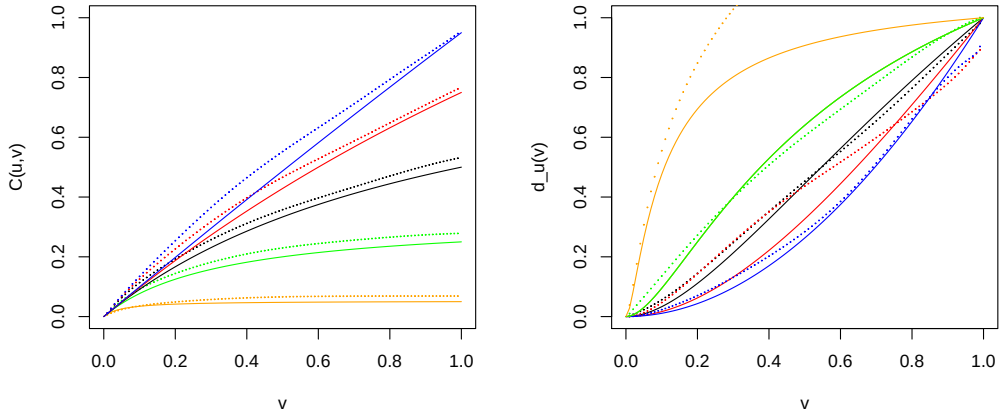

Figure 13: Kernel estimations (left) of the copula $C(u, v)$ in Example 2.2 for $u = 0.05, 0.25, 0.5, 0.75, 0.95$ (dotted lines from the bottom). The respective estimations for $d_u(v)$ obtained from eq. (3.2) are potted in the right. The continuous lines represent the exact values.

as an estimator for $d_u(v)$. Here we can use a numerical method to estimate its inverse function $d_u^{-1}(v)$ and to get the estimations for the quantile curves from eq. (2.2).

The estimations obtained of the copula $C(u, v)$ for $u = 0.05, 0.25, 0.5, 0.75, 0.95$ from the data in Example 2.2 are plotted in Figure 13, left. We have taken the bandwidth $h_n = n^{-1/5} = 0.3981072$ but we have seen that the estimations are very similar for other typical choices of h_n. The respective estimations for $d_u(v)$ obtained from eq. (3.2) are potted in Figure 13, right. The worse estimations are obtained when $u = 0.75$ (red line) and $u = 0.05$ (orange line).

Note that some of the functions $\hat{d}_u$ in Figure 13, right, are not distribution functions. This is due to the fact that $\hat{C}$ is not a copula. To avoid this problem the estimator in eq. (3.2) can be replaced with

$$\hat{d}_u^{\star}(v) = \hat{d}_{\hat{C}_1(u)}(\hat{C}_2(v)), \tag{3.3}$$

where

$$\hat{C}_1(u) = \hat{C}(u, 1) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{\Phi^{-1}(u) - \Phi^{-1}(U_i)}{h_n}\right)$$

and

$$\hat{C}_2(v) = \hat{C}(1, v) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{\Phi^{-1}(v) - \Phi^{-1}(V_i)}{h_n}\right)$$

are the marginal distributions of $\hat{C}$. Hence, from eq. (2.2), $\hat{d}_u^{\star}(v)$ are the conditional distribution functions of $\hat{C}$. Unfortunately, the estimations obtained from eq. (3.3), plotted in Figure 14, do not improve that obtained in Figure 13, right. They are not distribution functions due to the numerical errors in the computer calculations.

Finally, we use eq. (3.2) to approximate the quantile regression curves plotted in Figure 14, right, for the model (U, V) in Example 2.2 with $q = 0.5$ (black), $q = 0.25, 0.75$ (blue) and $q = 0.05, 0.95$ (red). These curves can be used to obtain (to estimate) the BBP (option 3) defined in Section 2.

4 Multivariate box plots

The regions defined in Section 2 can be extended to a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ with $n > 2$ to get the multivariate box plots (MBP). For example, if $n = 3$, then we can consider the regions defined as $R =$

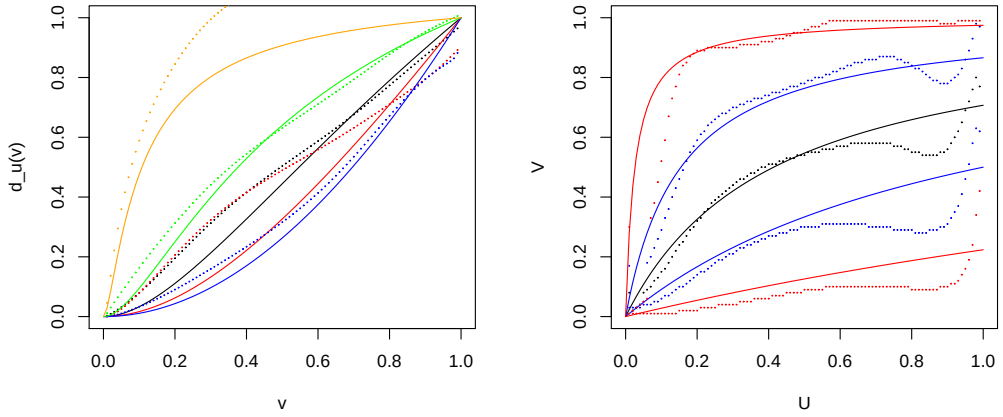

Figure 14: Kernel estimations (left) obtained from eq. (3.3) for $d_u(v)$ of the model (U, V) in Example 2.2 for $u = 0.05, 0.25, 0.5, 0.75, 0.95$ (dotted lines from the top). Quantile regression curves (right) for $q = 0.5$ (black), $q = 0.25, 0.75$ (blue) and $q = 0.05, 0.95$ (red). The continuous lines represent the exact values.

$[a, b] \times [\alpha_2, \beta_2] \times [\alpha_3, \beta_3]$ where $a, b \in \mathbb{R}$, $\alpha_2, \beta_2 : [a, b] \to \mathbb{R}$ and $\alpha_3, \beta_3 : [a, b] \times \mathbb{R} \to \mathbb{R}$. As in the bivariate case, if we use quantile maps, we have the following property.

Proposition 4.1. *With the above notation, if α_2, β_2 are u_2, v_2-quantile curves of $X_2|X_1$ and α_3, β_3 are u_3, v_3-quantile maps of $X_3|X_1, X_2$, then*

$$\Pr(\mathbf{X} \in R) = p_1 p_2 p_3,$$

where $p_1 = \Pr(a \le X_1 \le b)$, $p_2 = v_2 - u_2$ and $p_3 = v_3 - u_3$.

Proof. If the region R is defined as above, then

$$\begin{aligned}\Pr(\mathbf{X} \in R) &= \int_a^b \int_{\alpha_2(x_1)}^{\beta_2(x_1)} \int_{\alpha_3(x_1,x_2)}^{\beta_3(x_1,x_2)} \mathbf{f}(x_1, x_2, x_3) dx_3 dx_2 dx_1 \\ &= \int_a^b \int_{\alpha_2(x_1)}^{\beta_2(x_1)} \Pr(\alpha_3(x_1, x_2) \le X_3 \le \beta_3(x_1, x_2) | X_1 = x_1, X_2 = x_2) \mathbf{f}_{2|1}(x_2|x_1) dx_2 f_1(x_1) dx_1.\end{aligned}$$

Thus, if α_3 and β_3 are u_3, v_3-quantile maps of $(X_3|X_1 = x_1, X_2 = x_2)$, then

$$\Pr(\alpha_3(x_1, x_2) \le X_3 \le \beta_3(x_1, x_2) | X_1 = x_1, X_2 = x_2) = p_3$$

for $p_3 = v_3 - u_3$. Then, we get

$$\Pr(\mathbf{X} \in R) = p_3 \int_a^b \Pr(\alpha_2(x_1) \le X_2 \le \beta_2(x_1) | X_1 = x_1) f_1(x_1) dx_1.$$

Finally, if α_2 and β_2 are also u_2, v_2-quantile curves of $(X_2|X_1 = x_1)$, then

$$\Pr(\alpha_2(x_1) \le X_2 \le \beta_2(x_1) | X_1 = x_1) = p_2$$

for $p_2 = v_2 - u_2$ and so we get the stated result. □

The expression obtained in the preceding proposition can be used to define regions with specific probabilities. For example, if we use the first and third quartile curves to define the central region R_{ccc}, then we get $p_1 = p_2 = p_3 = 1/2$ and $\Pr(\mathbf{X} \in R_{ccc}) = 1/8$. The other regions can be defined similarly. As in Section 2, other options can be considered as well. The same technique can be used for $n > 3$.

Unfortunately, the preceding approach does not provide useful plots. We could just have 3D plots when $n = 3$. In this case we obtain 21 regions (obtaining plots similar to Rubik's cube). In the other cases we just might use the regions (but we cannot plot them). So we need to consider other options.

A typical one, is to provide in a common figure all the BBP obtained from pairs of random variables from $\mathbf{X}$ (see, e.g., Figure 7 in [20]). Moreover, note that here the BBP for (X_1, X_2) and (X_2, X_1) can be different (i.e. they are not just a symmetric transposition). To show this approach we plot in Figure 15 the BBP obtained by using linear regression quantile curves for the R-data called `stackloss`, which contains operational data of a plant for the oxidation of ammonia to get nitric acid (see [4] or the help included in R about this data set). The plots in the main diagonal represent the univariate box plots. The "linear regression quantile" procedure used in this figure can be replaced with any of the other procedures proposed in Section 3.

To conclude this section we propose a different approach also based in quantile regression. In all the paper we are assuming that we have a response variable Y that should be approximated when we know the value of an explanatory variable X. Actually, this is what happen in this data set where a response variable Y (called "stack loss") should be estimated from the variables X_1, X_2, X_3 plotted in Figure 15 (note that they are not "independent"). In linear quantile regression this estimation will be provided by the estimated median regression line given in this case by the formula

$$Y \approx m(x_1, x_2, x_3) := -39.68985507 + 0.83188406x_1 + 0.57391304x_2 - 0.06086957x_3. \tag{4.1}$$

In order to get a bivariate plot of this relationship between four variables we propose to create the artificial variable

$$Z = -39.68985507 + 0.83188406X_1 + 0.57391304X_2 - 0.06086957X_3 \tag{4.2}$$

and use it to get the BBP of (Z, Y) (plotted in Figure 16). Note that the linear regression curves included in this plot can be used to obtain confidence bands for the estimation of Y from the function m in Eq. (4.1). However, note that these confidence bands can be different from that obtained from a linear quantile regression with X_1, X_2, X_3. For example, the limits for 50% confidence band obtained from the first and third quartile lines in (Z, Y) (red lines) are $0.1917811 + 0.9266720z$ and $-0.8828384 + 1.1609055z$, that is,

$$q_1(x_1, x_2, x_3) = -36.5877 + 0.7708837x_1 + 0.5318291x_2 - 0.05640613x_3$$

and

$$q_3(x_1, x_2, x_3) = -46.95901 + 0.9657388x_1 + 0.6662588x_2 - 0.07066382x_3,$$

while the corresponding ones obtained from (X_1, X_2, X_3, Y) are

$$\tilde{q}_1(x_1, x_2, x_3) = -36 + 0.5x_1 + x_2 - 4.57967 \cdot 10^{-16}x_3$$

and

$$\tilde{q}_3(x_1, x_2, x_3) = -54.18966 + 0.8706897x_1 + 0.9827586x_2 + 2.677979 \cdot 10^{-16}x_3.$$

Note that $\tilde{q}_1$ and $\tilde{q}_3$ cannot be written in terms of Z and so they cannot be included in Figure 16. In this sense the artificial random variable Z can be seen as a kind of "first principal component" when we want to approximate Y from a median regression line based on X_1, X_2, X_3 since Z has all the information to get this estimation. However, Z does not contain the information needed to compute the limits $\tilde{q}_1$ and $\tilde{q}_3$. Instead we should use q_1 and q_3. The same can be applied to the bands based on other quantile lines or to other approaches based on non-linear quantile regression (see Subsection 3.2).

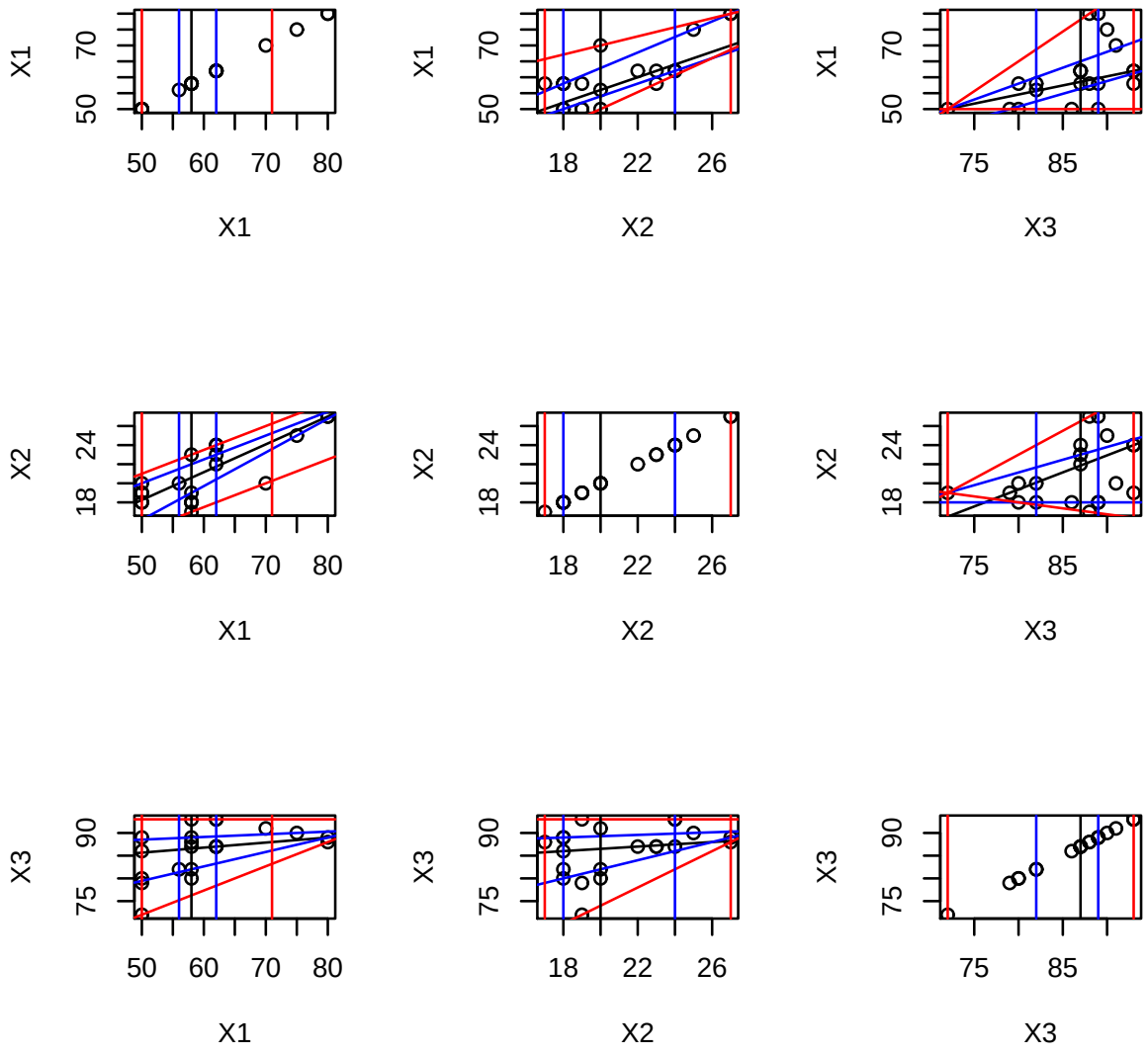

Figure 15: BBP obtained from linear quantile regression for a real data set with three variables. The vertical lines represent the univariate box plots.

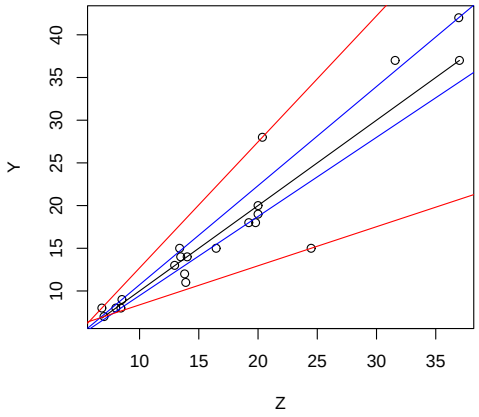

Figure 16: BBP in option 3 obtained from linear quantile regression for a real data set with three explanatory variables X_1, X_2, X_3 and a response variable Y. The variable Z given in (4.2) represents the linear median regression expression obtained from X_1, X_2, X_3.

5 Conclusions and open problems

We have proposed different ways to obtain bivariate box plots (BBP). They are a good alternative to other similar plots. They are based on some regions obtained from univariate box plots and conditional distributions. So they are especially useful when we have a response variable Y and an explanatory variable X.

The main advantage of this approach is that the regions in the theoretical BBP have fixed probabilities. In Section 2 we propose three options to fix these probabilities but the users can consider other options as well. As a consequence of this property, these regions can be used to perform fit tests to study if some data can come from a given bivariate model.

Moreover, the theoretical BBP can be obtained easily from eq. (2.2), the partial derivatives of the copula and the marginal distribution functions. They can also be obtained in a direct way for models with known conditional distributions (as happen for the normal model and all the models proposed in [1]). Two examples illustrate these procedures.

In practice the theoretical BBP should be estimated. Fortunately, we already have several techniques available in the literature to this purpose. Here we have showed some of them. The two first techniques are based on linear and non-linear quantile regression. So they can be computed easily by using the available packages in R. In the third one, we propose new empirical estimators based on concomitants. In the fourth, we assume a copula parametric model for the dependence and then we estimate this parameter from the data. As mentioned above, the assumed model can be confirmed by using a fit test based on the regions in the BBP. We also propose a fifth option based on new non-parametric kernel estimators for the partial derivative of the copula. All these approaches are illustrated with the two proposed examples.

Finally, we also consider the multivariate case, suggesting two ways to get BBP when we have more than two variables. A real data set (from R) is used to illustrate these options.

This paper is just a first step. So there are many tasks for future research. Let us mention just some of them. From a theoretical point-of-view, the most difficult one is to decide which option is the more reasonable to define the bivariate (and the univariate) box plots. In practice, we should decide which estimation procedure is the best for our data. There are a lot of results (papers) for the first approaches. However, the last one should be studied in detail. We have proposed two kernel estimators and we do not know which one is the best one. Also, we should study how to determine the optimal bandwidth. Moreover, if we assume a

bivariate copula regression model, one would like to study some of the properties of the residuals resulting from fitting such model. These and other tasks are left for future research projects.

Acknowledgements: JN was supported in part by Ministerio de Ciencia e Innovación of Spain under grant PID2019-103971GB-I00.

References

[1] Arnold, B. C., E. Castillo, and J. M. Sarabia (1999). *Conditional Specification of Statistical Models*. Springer, New York.
[2] Bernard, C. and C. Czado (2015). Conditional quantiles and tail dependence. *J. Multivariate Anal. 138*, 104–126.
[3] Bernardi, M., F. Durante, and P. Jaworski (2017). CoVaR of families of copulas. *Statist. Probab. Lett. 120*, 8–17.
[4] Brownlee, K. A. (1960). *Statistical Theory and Methodology in Science and Engineering*. John Wiley & Sons, New York.
[5] Chen, X. (2011). A new generalization of Chebyshev inequality for random vectors. Available at https://arxiv.org/abs/0707.0805.
[6] Fernández-Ponce, J. M. and A. Suárez-Lloréns (2002). Central regions for bivariate distributions. *Austrian J. Stat. 31*(2&3), 141–156.
[7] Greenwood, P. E. and M. S. Nikulin (1996). *A Guide to Chi-Squared Testing*. John Wiley & Sons, New York.
[8] Hyndman, R. J. and Y. Fan (1996). Sample quantiles in statistical packages. *Amer. Statist. 50*(4), 361–365.
[9] Jaworski, P. (2017). On conditional value at risk (CoVaR) for tail-dependent copulas. *Depend. Model. 5*, 1–19.
[10] Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.
[11] Koenker, R. and G. Bassett Jr. (1978). Regression quantiles. *Econometrica 46*(1), 33–50.
[12] Koenker, R. and B. J. Park (1996). An interior point algorithm for nonlinear quantile regression. *J. Econometrics 71*(1-2), 265–283.
[13] Koenker, R., S. Portnoy, P. Tian, A. Zeileis, P. Grosjean, C. Moler, and B. D. Ripley (2020). `Quantreg`: *Quantile Regression*. R package version 5.55. Available on CRAN.
[14] Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press, London.
[15] Nagler, T. (2014). Kernel Methods for Vine Copula Estimation. Master's thesis, Technische Universität München, Germany.
[16] Navarro, J. (2014). A note on confidence regions based on the bivariate Chebyshev inequality. *İstatistik 7*(1), 1–14.
[17] Navarro, J. (2016). A very simple proof of the multivariate Chebyshev's inequality. *Comm. Statist. Theory Methods 45*(12), 3458–3463.
[18] Navarro, J. and M. A. Sordo (2018). Stochastic comparisons and bounds for conditional distributions by using copula properties. *Depend. Model. 6*, 156–177.
[19] Nelsen, R. B. (2006). *An Introduction to Copulas*. Second edition. Springer, New York.
[20] Rousseeuw, P. J., I. Ruts, and J. W. Tukey (1999). The bagplot: A bivariate boxplot. *Amer. Statist. 53*(4), 382–387.
[21] Sumarjaya, I. W. (2017). A survey of kernel-type estimators for copula and their applications. *J. Phys.: Conf. Ser. 893*, Article ID 012027, 6 pages.