AUSTRIAN JOURNAL OF STATISTICS
Volume 31 (2002), Number 2&3, 141-156

Central Regions for Bivariate Distributions

Jose María Fernández-Ponce[1] and Alfonso Suárez-Lloréns[2]
[1] University Sevilla, Tarfia s/n, Sevilla, Spain
[2] University Cadiz, Duque de Nájera 8, 11002 Cádiz, Spain

Abstract: For a one-dimensional probability distribution, the classical concept of central region as a real interquantile interval arises in all applied sciences. We can find applications, for instance, with dispersion, skewness and detection of outliers. All authors agree with the main problem in a multivariate generalization: there does not exist a natural ordering in n-dimensions, $n > 1$. Because of this reason, the great majority of these generalizations depend on their use. We can say that is common to generalize the concept of central region under the definition of the well known concept of spatial median. In our work, we develop an intuitive concept which can be interpreted as level curves for distribution functions and this one provides a trimmed region. Properties referred to dispersion and probability are also studied and some considerations on more than two dimensions are also considered. Furthermore, several estimations for bivariate data based on conditional quantiles are discussed.

Keywords: Spatial Quantile, Central Region, Conditional Quantile.

1 Introduction

For a one-dimensional probability distribution, the classical concept of central region as a real interquantile interval, $IQ(p) = (F^{-1}(1-p), F^{-1}(p))$, for $1/2 < p < 1$, arises in all applied sciences. We can find applications, for instance, with dispersion, skewness and detection of outliers. In the multivariate finite dimensional case, several central regions have been studied. All authors agree with the main problem in a multivariate generalization; there does not exist a natural ordering in n-dimensions, $n > 1$. From this, it has always been a serious obstacle to the development of statistical methods based on order statistics. In a classical paper, Barnett (1976) provides several possible methods for ordering multivariate data. Since 1976 we have in the literature various attempts that all are valuable contributions toward both multidimensional generalization of univariate quantiles and the generalization of the real interquantile interval. The authors usually provide a generalization of the univariate quantiles based on properties in one dimension and after that they use this one to define a new concept of multivariate central region. For instance, we can see the concept of depth function, see Tukey (1975) and the notion of quantiles based on Oja's criterion function that arises in the definition of Oja's simplex median, Oja (1983). Another way is to generalize directly the real interquantile interval, that is the case of the minimum volume ellipsoid with fixed probability, see Rousseeuw and Leroy (1987). Other examples are provided by the notion of trimmed region, defined as the intersection of all half-planes whose μ-probability measure is at least equal to p , the p-trimmed regions are known as peeling procedures, see Nolan (1992) and Massé and

Theodorescu (1994). To study thoroughly the concept of multivariate quantiles is interesting to see the generalization provided in Abdous and Theodorescu (1992) and Chaudhuri (1996), this concept is based on a generalization of the spatial median. Anyway, there are several approaches based on different point of views. Because of this reason the great majority of these generalizations depend on their use. In a recent paper, Averous and Meste (1997) provided an interesting introduction of various classical concepts of central regions and a brief explanation about the problems which we can observe in each one. Basically, the problems come from the shape of these regions which is a priori chosen, so it is difficult to interpret the majority of these regions for distributions non-symmetrical. Another problem is referred to the probability accumulated which is less obvious than in the real case, that is to say $\Pr\{X \in IQ(p)\} = 2p - 1$.

In spite of all problems we can find because of the multivariate nature, there is a point of view which is frequently accepted in the literature; to generalize the concept of central region through the notion of spatial median, for instance the Median Balls, see Averous and Meste (1997). This way has been extensively studied and the central regions are defined through a proximity criterion to the spatial median, it seems to be the most interesting one to generalize the real interquantile interval. Let $\mathbf{X}$ be a n-dimensional random vector with distribution $P(.)$, the spatial median is defined as

$$M = \arg\min_{c} \int_{\mathbf{X}} \parallel \mathbf{x} - \mathbf{c} \parallel - \parallel \mathbf{x} \parallel \, dP(\mathbf{x}).$$

In depends on the used norm, but in the $\mathbf{L}_1$ sense, the city block norm, we obtain the vector of marginal medians for each components of $\mathbf{X}$.

In our work, we develop an intuitive concept of central region for bivariate distributions. We propose a trimmed region which is centered around the spatial median. We also show the shape is not a priori chosen thus it can be used to study symmetrical and non-symmetrical distributions. Properties related to dispersion and probability are studied and some considerations on more than two dimensions are also considered. Finally, the introduced notion is illustrated for particular cloud data and properties of estimation are discussed.

2 Multivariate Quantiles as Level Curves

One of the most difficult problems when we broach the multivariate case is the notation. From now on, we represent in bold both variables and points with more than one dimension. Let V_n^2 denote the set of variations of two elements with n repetitions and let $\delta_{k_1,\cdots,k_n} \in V_n^2$ denote the variation $k_1, \cdots, k_n$, where $k_j = \{0 \text{ or } 1\}$, for $j = 0, \cdots, n$. Let Π_n denote the set of permutations of n elements, $\{1, \cdots, n\}$, where $\pi_{i_1,\cdots,i_n}$ represents the permutation $i_1, \cdots, i_n$. The symbols Δ_0 and Δ_1 represent the inequalities "$\leq$" and "$\geq$" respectively. Furthermore α_0 and α_1 represent the symbols "$+\infty$" and "$-\infty$" respectively. For an easier notation, let $\mathbf{x}$ and $\mathbf{y}$ be two points in $\mathbb{R}^n$ and let $\delta_{k_1,\cdots,k_n} \in V_n^2$, we denote $\mathbf{x}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{y}$ if and only if $x_i\Delta_{k_i}y_i$, for all $i = 1, \cdots, n$. Let X be a univariate random variable and let $p \in (0, 1)$ then $Q_X(p)$ denotes the p-th univariate quantile, that is to say

$$Q_X(p) = F_X^{-1}(p) = \inf\{x : F(x) > p\}.$$

Although we study the bivariate case, we introduce the following concepts for n dimensions.

Definition 2.1. *Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a n-dimensional random vector. Let $p \in (0,1)$, then define the p-th multivariate quantile set, $Q_{\mathbf{X}}(p; \delta_{k_1,\cdots,k_n})$, under the variation $\delta_{k_1,\cdots,k_n} \in V_n^2$ as*

$$Q_{\mathbf{X}}(p; \delta_{k_1,\cdots,k_n}) = \partial \left\{ \mathbf{x} \in \mathbb{R}^n : Pr\{\mathbf{X}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{x}\} > p \right\}$$

where ∂ represents the topological border.

Note that for fixed $p \in (0,1)$, we have 2^n multivariate quantile sets, one for each variation in $\mathbb{R}^n$. It is easy to show the above definition is a generalization of univariate quantile set. For $n = 1$ it holds

$$Q_X(p; \delta_0) = \inf\{x : F(x) > p\} \text{ and } Q_X(p; \delta_1) = \sup\{x : F(x) < 1 - p\},$$

see Lewis and Thompson (1981) for more details in the definition of univariate quantiles.

The following proposition provides an easy interpretation of the quantile sets as level curves for the distribution function.

Proposition 2.1. *Let $\mathbf{X}$ be a random vector and let $Q_{\mathbf{X}}(p, \delta_{k_1,\cdots,k_n})$.*
Then $\mathbf{x} \in Q_{\mathbf{X}}(p, \delta_{k_1,\cdots,k_n})$ if and only if for each $\mathbf{y_1}$ and $\mathbf{y_2}$ such that

$$\mathbf{y_1}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{x}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{y_2}$$

strictly in all its components, it holds

$$Pr\{\mathbf{X}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{y_1}\} \leq p < Pr\{\mathbf{X}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{y_2}\}.$$

Proof. We consider the sets

$$A = \left\{ \mathbf{x} \in \mathbb{R}^n : \quad Pr\{\mathbf{X}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{x}\} > p \right\},$$

and

$$A^c = \left\{ \mathbf{x} \in \mathbb{R}^n : \quad Pr\{\mathbf{X}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{x}\} \leq p \right\}.$$

Observe that A and A^c are increasing and decreasing under the variation $\delta_{k_1,\cdots,k_n} \in \mathbb{V}_n^2$ respectively. That is to say, for $\mathbf{x} \in A$ and $\mathbf{y}$ such that $\mathbf{x}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{y}$ then it holds $\mathbf{y} \in A$. In the same way, for $\mathbf{x} \in A^c$ and $\mathbf{y}$ such that $\mathbf{y}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{x}$ then it holds $\mathbf{y} \in A^c$.

For the necessary condition, let $\mathbf{x} \in \partial A$ and let $\mathbf{y_1}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{x}$ strictly in all their components. Suppose that $\Pr\{\mathbf{X}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{y_1}\} > p$, then $\mathbf{y_1} \in A$. Because A is increasing under the variation $\delta_{k_1,\cdots,k_n}$, there exists $\epsilon > 0$ such that the Euclidean ball centered in $\mathbf{x}$, $B(x, \epsilon)$, is included in A, hence $\mathbf{x} \notin \partial A$ which is a contradiction. Analogously for $\mathbf{x}\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{y_2}$ and using A^c. The sufficiency of the condition is trivial. □

Note that the definition of multivariate quantile set is based on the probabilities of being in the 2^n orthants in $\mathbb{R}^n$. If $\mathbf{X}$ is a random vector with continuous distribution function then for each $\mathbf{x}$ that belongs to the p-th multivariate quantile set it holds

$$\Pr\{X\Delta_{\delta_{k_1,\cdots,k_n}}\mathbf{x}\} = p.$$

At this point, it is interesting a particular study about flat zones for the distribution function, see Lewis and Thompson (1981).

Proposition 2.2. *Let $\mathbf{X}$ be a random vector with a continuous distribution function and let $Q_{\mathbf{X}}(p;\delta_{k_1,\cdots,k_n})$ for all $\delta_{k_1,\cdots,k_n} \in V_n^2$ be the family of the 2^n quantile sets then*

1. *If $p > \frac{1}{2}$ the intersection of all quantile sets is empty.*
2. *If $\frac{1}{k+1} < p \leq \frac{1}{2}$ the intersection of whatever $k+1$ quantile sets is empty for $2 \leq k \leq 2^n - 1$.*

Proof. Trivial □

It is also easy to show the equivariance under both location and any homogeneous scale transformations. Note that in some situations, it is necessary to standardize the coordinate variables, for instance when the units of measurements for different coordinate variables happen to be different. The following step is to study the relation between marginal quantiles and the multivariate quantile set.

Definition 2.2. *Let $\mathbf{X}$ be a random vector and let $Q_{\mathbf{X}}(p,\delta_{k_1,\cdots,k_n})$ be the multivariate quantile set, then define the p-th multivariate marginal quantile vector under the variation $\delta_{k_1,\cdots,k_n}$, denoted by $\mathbf{q}(p,\delta_{k_1,\cdots,k_n})$ as the vector with i-th component*

$$[\mathbf{q}(p,\delta_{k_1,\cdots,k_n})]_i = Q_{X_i}(p)^{1-k_i}Q_{X_i}(1-p)^{k_i},$$

for all $i = 1,\cdots,n$.

Example 2.1. *Let $\mathbf{X}$ be a bivariate random vector and let $p \in (0,1)$ then there are four multivariate marginal quantile vectors, that is to say, $\mathbf{q}(p,\delta_{0,0}) = (Q_{X_1}(p), Q_{X_2}(p))$, $\mathbf{q}(p,\delta_{1,0}) = (Q_{X_1}(1-p), Q_{X_2}(p))$, $\mathbf{q}(p,\delta_{1,1}) = (Q_{X_1}(1-p), Q_{X_2}(1-p))$ and $\mathbf{q}(p,\delta_{0,1}) = (Q_{X_1}(p), Q_{X_2}(1-p))$.*

Theorem 2.1. *Let $\mathbf{X}$ be a random vector with continuous distribution function strictly increasing in all its components, then it holds*

$$\lim_{\substack{x_i \to \alpha_{k_i} \\ i \neq j \\ \mathbf{x} \in Q_{\mathbf{X}}(p,\delta_{k_1,\cdots,k_n})}} \Pi_j(\mathbf{x}) = [\mathbf{q}(p,\delta_{k_1,\cdots,k_n})]_j \tag{1}$$

where Π_j is the function which maps the j-th component.

Proof. Let $\mathbf{x} \in Q_{\mathbf{X}}(p, \delta_{k_1,\cdots,k_n})$ and

$$\mathbf{y} = (\alpha_{k_1}, \cdots, y_j, \cdots, \alpha_{k_n}),$$

where $x_j \Delta_{k_j} y_j$ strictly. By Proposition 2.1 it holds

$$\Pr\{\mathbf{X} \Delta_{\delta_{k_1,\cdots,k_n}} \mathbf{y}\} = \Pr\{X_j \Delta_{\delta_{k_j}} y_j\} > p,$$

thus

$$\begin{cases} Q_{X_j}(p) \leq x_j & \text{if } k_j = 0, \\ Q_{X_j}(1-p) \geq x_j & \text{if } k_j = 1. \end{cases}$$

Hence under the assumptions of the distribution function it holds 1. □

Note that the p-th multivariate marginal quantile vector, $\mathbf{q}(p, \delta_{k_1,\cdots,k_n})$, is ordered respect to quantile set under variation $\delta_{k_1,\cdots,k_n}$ that is to say

$$\forall \mathbf{x} \in Q_{\mathbf{X}}(p, \delta_{k_1,\cdots,k_n}) \Longrightarrow \mathbf{q}(p, \delta_{k_1,\cdots,k_n}) \Delta_{\delta_{k_1,\cdots,k_n}} \mathbf{x}.$$

The one-dimensional projections of the multivariate quantile set tend to the univariate quantiles for each component. This last asymptotic property is referred to definition of spatial quantile given by Abdous and Theodorescu (1992) and Chaudhuri (1996) for the city block norm as follows

$$\mathbf{q}(p, \delta_{k_1,\cdots,k_n}) = \arg \inf_{\mathbf{Q} \in \mathbb{R}^n} E\{\Phi(\mathbf{u}, \mathbf{X} - \mathbf{Q}) - \Phi(\mathbf{u}, \mathbf{X})\},$$

where $\mathbf{u}$ is defined as

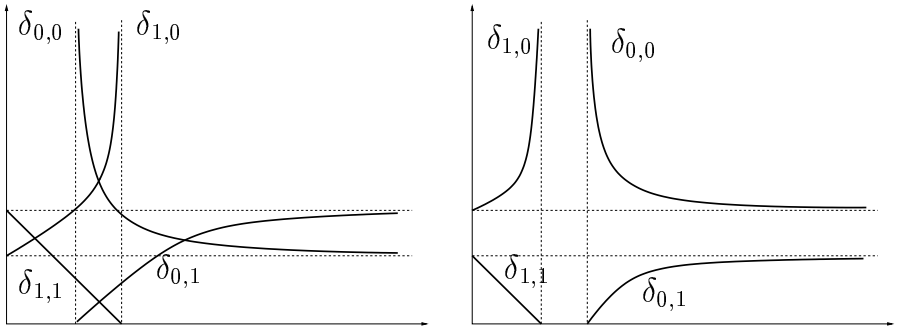
$$u_i = 2\left(p^{1-k_i}(1-p)^{k_i}\right) - 1, \quad \forall i = 1, \cdots, n,$$

and $\Phi(\mathbf{u}, \mathbf{t}) = \| \mathbf{t} \|_1 + < \mathbf{u}, \mathbf{t} >$. Note that when $p = 1/2$ it holds the spatial median.

Example 2.2. *Let $\mathbf{X}$ be a bivariate random vector with independent components which come from an exponential distribution with parameter $\lambda = 1$ and let p_1, p_2 such that $p_1 < 1/2$ and $p_2 > 1/2$. We show in the Figure 1 (a) and (b) all bivariate quantile sets for p_1 and p_2 respectively. For instance, the points which belong to $Q_{\mathbf{X}}(p, \delta_{1,0})$, denoted as $\delta_{1,0}$ in the picture, they accumulate the same probability in the fourth orthant sense, that is to say, the probability $Pr\{\mathbf{X} \Delta_{1,0} \mathbf{x}\} = p$ for all $\mathbf{x}$ that belong to this level curve. Note that the level curves tend to the multivariate quantile marginal vector. The separation between each curve and the asymptotic lines is referred to a well known multivariate property denoted as concordance, see Shaked and Shanthikumar (1994).*

To conclude this section we provide a characterization of the multivariate quantile sets through the univariate conditional distributions.

Definition 2.3. *Let $\mathbf{X}$ be a multivariate random vector and let $\mathbf{u}$ a multivariate vector with components $u_i \in (0, 1)$, for $i = 1, \cdots, n$. Then define the quantile curve under the*



(a) $p_1 < \frac{1}{2}$. (b) $p_2 > \frac{1}{2}$.

Figure 1: Bivariate Quantile Sets.

variation $\delta_{k_1,\cdots k_n} \in \mathbb{V}_n^2$ *and under the permutation* $\pi_{i_1,\cdots i_n} \in \Pi_n$ *denoted* $\overset{\wedge}{\mathbf{q}}(\delta, \pi, \mathbf{u})$ *as the vector* $(\hat{q}_1\ (u_1), \cdots, \hat{q}_n\ (u_n))$ *where each component is given as*

$$\begin{aligned}
\hat{q}_{i_1}\ (u_{i_1}) &= Q_{X_{i_1}}(u_{i_1}), \\
&\vdots \\
\hat{q}_{i_n}\ (u_{i_n}) &= Q_{X_{i_n}|_{\bigcap_{j=1}^{n-1}(X_{i_j}\Delta_{k_j}\hat{q}_{i_j}(u_{i_j}))}}(u_{i_n}).
\end{aligned}$$

Note that $\hat{q}_{i_j}\ (u_{i_j})$ corresponds to the i_j-th component of the vector $\overset{\wedge}{\mathbf{q}}(\delta, \pi, \mathbf{u})$. The above definition is referred to the univariate quantiles for conditional distributions.

Theorem 2.2. *Let* $\mathbf{X}$ *be a multivariate random vector with distribution function absolutely continuous and strictly increasing in all its components. The multivariate quantile set* $Q_{\mathbf{X}}(p, \delta_{k_1,\cdots,k_n})$ *is characterized through the quantile curve under the same variation and whatever chosen permutation, that is to say*

$$\mathbf{x} \in Q_{\mathbf{X}}(p, \delta_{k_1,\cdots,k_n}) \textit{ if and only if } \exists \mathbf{u} \in (0,1)^n, \quad \beta \in (0,1)^n, \textit{ where}$$

$$\mathbf{x} = \overset{\wedge}{\mathbf{q}}(\delta_{k_1,\cdots,k_n}, \pi_{i_1,\cdots,i_n}, \mathbf{u}), \textit{ with}$$

$$u_j = (\beta_j)^{1-k_j}(1-\beta_j)^{k_j}, \quad \forall j = 1, \cdots, n, \textit{ and } \prod_{j=1}^{n} \beta_j = p.$$

Proof. The necessary condition. By the assumptions of the distribution function it holds

$$\forall \mathbf{x} \in Q_{\mathbf{X}}(p, \delta_{k_1,\cdots,k_n}) \Longrightarrow \Pr\{\mathbf{X}\Delta_{\delta_{k_1,\cdots k_n}}\mathbf{x}\} = p,$$

thus

$$p = \Pr\{X_{i_1}\Delta_{k_{i_1}}x_{i_1}\} \cdots \Pr\{X_{i_n}\ |_{\bigcap_{j=1}^{n-1}(X_{i_j}\Delta_{k_{i_j}}x_{i_j})}\ \Delta_{k_{i_n}}x_{i_n}\},$$

hence

$$\beta_{i_l} = \Pr\{X_{i_l} \mid_{\bigcap_{j=1}^{l-1}(X_{i_j}\Delta_{k_{i_j}}x_{i_j})} \Delta_{k_{i_l}} x_{i_l}\}, \quad \forall l = 1, \cdots n.$$

Note there do not exist flat zones for the distribution function. We only have to take $\mathbf{u} \in (0,1)^n$ as

$$u_{i_l} = (\beta_{i_l})^{1-k_{i_l}}(1-\beta_{i_l})^{k_{i_l}},$$

then

$$\begin{aligned} \hat{q}_{i_1}(u_{i_1}) &= Q_{X_{i_1}}(u_{i_1}) = x_{i_1}, \\ &\vdots \\ \hat{q}_{i_n}(u_{i_n}) &= Q_{X_{i_n}|_{\bigcap_{j=1}^{n-1}(X_{i_j}\Delta_{k_{i_j}}\hat{q}_{i_j}(u_{i_j}))}}(u_{i_n}) = x_{i_n}. \end{aligned}$$

The sufficient condition is easy to prove under the properties of the distribution function. □

Note that in the above conditions, the vector $\mathbf{u}$ is given through the equation

$$\prod_{i=1}^{n}(u_i)^{1-k_i}(1-u_i)^{k_i} = p,$$

with $u_i \in (0,1)$, for $i = 1, \cdots n$. The characterization depends on the chosen permutation, so for each variation there exist $n!$ different ways to describe the same multivariate quantile set. This property can be useful for distributions which have different relevant components.

Example 2.3. *Let* $\mathbf{X}$ *be a bivariate random vector and let* $\delta_{0,0} \in \mathbb{V}_2^2$*. By Theorem 2.2 we obtain the level curve for the distribution function through the conditional distributions. We have two characterizations, one for each permutation. For the permutation* $\pi_{1,2}$ *it holds*

$$Q_{\mathbf{X}}(p, \delta_{0,0}) = \{(Q_{X_1}(u), Q_{X_2|X_1 \leq Q_{X_1}(u)}(p/u)) : \quad \forall u > p\}$$

and for the permutation $\pi_{2,1}$

$$Q_{\mathbf{X}}(p, \delta_{0,0}) = \{(Q_{X_1|X_2 \leq Q_{X_2}(u)}(p/u), Q_{X_2}(u)) : \quad \forall u > p\}$$

Note that for random vectors with independent components $Q_{\mathbf{X}}(p, \delta_{k_1,\cdots,k_n})$ is characterized through

$$(Q_{X_1}(u_1), \cdots, Q_{X_n}(u_n)), \quad 0 < u_i < 1 \quad i = 1, \cdots n,$$

where $\prod_{i=1}^{n}(u_i)^{1-k_i}(1-u_i)^{k_i} = p$.

3 The Bivariate Central Region

The definition of multivariate quantile sets lead us to define a new concept of central region. In the bivariate case, the four level curves provide five regions in the plane as a generalization of the three regions for the univariate interquantile interval. Obviously, we are interested in the region among all level curves.

Definition 3.1. *Let* $\mathbf{X} = (X_1, X_2)$ *be a bivariate random vector and let* $p \in [1/2, 1]$. *Then define the central region, denoted* $\Omega_{\mathbf{X}}(p)$*, as follows*

$$\Omega_{\mathbf{X}}(p) = \left\{ \mathbf{x} \in \mathbb{R}^2 : Pr\{X\Delta_{\delta_{k_1,k_2}}\mathbf{x}\} < p, \quad \forall \delta_{k_1,k_2} \in \mathbb{V}_2^2 \right\}.$$

Let Z be a real random variable, then $\Omega_Z(p)$ is a generalization of the real interquantile interval in the following way

$$\Omega_Z(p) = \{x : Q_Z(1-p) < x < Q_Z^+(p)\}.$$

where $Q_Z^+(p) = \sup\{x : F(x) < p\}$. It is easy to show that the central regions are ordered by inclusion, that is to say, for p and q such that $0 < p < q < 1$ then $\Omega_{\mathbf{X}}(p) \subset \Omega_{\mathbf{X}}(q)$. Note that the shape of the central region is not a priori chosen thus it can be applied for symmetrical and non-symmetrical distributions. In the general case, $\Omega_{\mathbf{X}}(p)$ is not a bounded region. Otherwise, we will provide some remarks to bound the central region. Now we provide a result concerning the accumulated probability. From now on, we will consider distribution functions absolutely continuous and strictly increasing in all their components, and we call this regularity conditions.

Obviously, the central region corresponds to the points among all level curves. For a more operative definition, we describe the central region as following

$$\Omega_{\mathbf{X}}(p) = \bigcup_{i=1}^{2} \Lambda_{X_i}(p) \bigcup_{\delta_{k_1,k_2}} \Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p) \tag{2}$$

where $\Lambda_{X_i}(p)$ corresponds to the region among the marginal quantiles

$$\Lambda_{X_i}(p) = \{\mathbf{x} \in \mathbb{R}^2 : Q_{X_i}(1-p) < x_i < Q_{X_i}(p)\}$$

and $\Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p)$ corresponds to the region

$$\Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p) = \left\{ \mathbf{x} \in \mathbb{R}^2 : \mathbf{q}(\delta_{k_1,k_2}, p)\Delta_{\delta_{k_1,k_2}}\mathbf{x} \text{ and } \Pr\{\mathbf{X}\Delta_{\delta_{k_1,k_2}}\mathbf{x}\} < p \right\}.$$

Since the Proposition 2.1, it holds that $\Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p)$ is referred to the points between the p-th multivariate marginal quantile vector and the multivariate quantile set

$$\Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p) = \left\{ \mathbf{x} \in \mathbb{R}^2 : \mathbf{q}(\delta_{k_1,k_2}, p)\Delta_{\delta_{k_1,k_2}}\mathbf{x} \text{ and } \exists \mathbf{y} \in Q_{\mathbf{X}}(p, \delta_{k_1,k_2}), \quad \mathbf{x}\Delta_{\delta_{k_1,k_2}}\mathbf{y} \right\},$$

and from the Theorem 2.2, using the permutation $\pi_{1,2}$, we obtain that $\Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p)$ can be characterized as

$$\Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p) = \left\{ \mathbf{x} \in \mathbb{R}^2 : \mathbf{q}(\delta_{k_1,k_2}, p)\Delta_{\delta_{k_1,k_2}}\mathbf{x}, \text{ and } x_2\Delta_{k_2}[\hat{\mathbf{q}}\,(\delta_{k_1,k_2}, \pi_{1,2}, \mathbf{u})]_2 \right\}. \tag{3}$$

where $u_1 = F_{X_1}(x_1)$ and $\beta_1 = \Pr\{X_1 \Delta_{k_1} x_1\}$, so $u_2 = (p/\beta_1)^{1-k_2}(1-p/\beta_1)^{k_2}$.

Note that the Equation 2 it is easy to show by inclusion. Observe that

$$\Lambda_{X_i}(p) \bigcap \Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p) = \emptyset,$$

for $i = 1, 2$ and for all $\delta_{k_1,k_2} \in \mathbb{V}_2^2$. It holds too that

$$\Upsilon_{\mathbf{X}}(\delta_{k_1',k_2'}, p) \bigcap \Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p) = \emptyset$$

for all $(k_1', k_2') \neq (k_1, k_2)$.

In respect to the spatial quantiles, the p-th multivariate marginal quantile vector belongs to the central region for all variation, hence the central region is centered around the spatial median.

Connecting with other definitions, it is easy to show the interquantile ball provided by Chaudhuri (1996) for $\| \, . \, \|_1$ as

$$\{(Q_{X_1}(r_1), Q_{X_2}(r_2)) : \mid 2r_1 - 1 \mid + \mid 2r_2 - 1 \mid < r\},$$

it is included in $\Omega_{\mathbf{X}}(p)$ for $r = 2p - 1$. On the other hand, the notion of trimmed region, defined as the intersection of all half-planes which μ-probability measure is at least equal to p, Nolan (1992), it is obviously included in $\Omega_{\mathbf{X}}(p)$. Finally, certain properties referred to median balls defined by Averous and Meste (1997), are also possible for symmetrical distributions, but in general the relation in this sense it is not so clear.

Theorem 3.1. *Let* $\mathbf{X}$ *be a bivariate random vector and let* F *be its distribution function under the regularity conditions and let* f *be its density function. Let* $p \in [1/2, 1)$ *then the accumulated probability is*

$$Pr\{\mathbf{X} \in \Omega_{\mathbf{X}}(p)\} = 4p(1 + \ln(\frac{1}{p})) - 3 + R_{\mathbf{X}}(p),$$

where $R_{\mathbf{X}}(p)$

$$\sum_{k_1=0}^{1} (-1)^{k_1} \int_{q_1 \Delta_{k_1} x_1} \int_{\hat{q}_2(1-p/\beta_1)}^{\hat{q}_2(p/\beta_1)} Pr\{X_1 \Delta_{k_1} x_1\} \frac{\partial}{\partial x_1} f_{X_2|_{X_1 \Delta_{k_1} x_1}}(x_2) dx_2 dx_1,$$

with $q_i = [\mathbf{q}(p, \delta_{k_1,k_2})]_i$ *for* $i = 1, 2$ *and*

$$\hat{q}_2(.) = [\hat{\mathbf{q}}\,(\delta_{k_1,k_2}, \pi_{1,2}, \mathbf{u})]_2,$$

for $u_1 = F_{X_1}(x_1)$ *and* $\beta_1 = Pr\{X_1 \Delta_{k_1} x_1\}$.

Proof. Since 2 we can share the study of the probability in different regions. At first we obtain the probabilities of being in $\Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p)$. Since the expression 3 it holds

$$\Pr\{\mathbf{X} \in \Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p)\} = \int \int_{\Upsilon_{\mathbf{X}}(\delta_{k_1,k_2}, p)} dF_{X_1,X_2}(x_1, x_2)$$

$$= \int_{q_1\Delta_{k_1}x_1} \int_{q_2\Delta_{k_2}x_2\Delta_{k_2}\hat{q}_2(u_2)} dF_{X_1,X_2}(x_1,x_2)$$
$$= \underbrace{\int_{q_1\Delta_{k_1}x_1} \int_{x_2\Delta_{k_2}\hat{q}_2(u_2)} dF_{X_1,X_2}(x_1,x_2)}_{[1]}$$
$$- \underbrace{\int_{q_1\Delta_{k_1}x_1} \int_{x_2\Delta_{k_2}q_2} dF_{X_1,X_2}(x_1,x_2).}_{[2]}$$

Where $u_2 = (p/\beta_1)^{1-k_2}(1-p/\beta_1)^{k_2}$. It is easy to show that

$$[2] = \Pr\{q_1\Delta_{k_1}X_1, X_2\Delta_{k_2}q_2\}.$$

Furthermore, no more that represents

$$\mathbb{R}^2 - \Lambda_{X_1}(p) \cup \Lambda_{X_2}(p) = \bigcup_{k_1,k_2} \{\mathbf{x} \in \mathbb{R}^2 : \quad \mathbf{q}(p,\delta_{k_1,k_2})\Delta_{\delta_{k_1,k_n}}\mathbf{x}\},$$

it holds that

$$-\sum_{k_1,k_2} \Pr\{q_1\Delta_{k_1}X_1, X_2\Delta_{k_2}q_2\} + \Pr\{\mathbf{X} \in \Lambda_{X_1}(p) \cup \Lambda_{X_2}(p)\} = 4p - 3.$$

On the other hand, since the following expressions

$$f_{X_2|X_1=x_1}(x_2) = \frac{f(x_1,x_2)}{f_{X_1}(x_1)}, \quad f_{X_2|X_1\Delta_{k_1}x_1}(x_2) = \frac{\partial/\partial x_2 \Pr\{X_1\Delta_{k_1}x_1, X_2 \le x_2\}}{\Pr\{X_1\Delta_{k_1}x_1\}},$$

it is easy to show that

$$f_{X_2|X_1=x_1}(x_2) = f_{X_2|X_1\Delta_{k_1}x_1}(x_2) + (-1)^{k_1}\frac{\Pr\{X_1\Delta_{k_1}x_1\}}{f_{X_1}(x_1)}\frac{\partial}{\partial x_1}f_{X_2|X_1\Delta_{k_1}x_1}(x_2).$$

So if denote I_1, I_2 the intervals $q_1\Delta_{k_1}x_1$ and $x_2\Delta_{k_2}\hat{q}_2(u_2)$ respectively, it holds

$$[1] \Longrightarrow \int_{I_1} f_{X_1}(x_1) \int_{I_2} f_{X_2|X_1=x_1}(x_2)dx_2dx_1$$
$$= \underbrace{\int_{I_1} f_{X_1}(x_1) \int_{I_2} f_{X_2|X_1\Delta_{k_1}x_1}(x_2)dx_2dx_1}_{[3]}$$
$$+ \underbrace{\int_{I_1}\int_{I_2} (-1)^{k_1}\Pr\{X_1\Delta_{k_1}x_1\}\frac{\partial}{\partial x_1}f_{X_2|X_1\Delta_{k_1}x_1}(x_2)dx_2dx_1.}_{[4]}$$

We obtain that $[3] = p\ln(\frac{1}{p})$, for all $\delta_{k_1,k_2} \in \mathbb{V}_2^2$. In addition,

$$\int_{q_1\Delta_{k_1}x_1} f_{X_1}(x_1) \int_{-\infty}^{\infty} \left(f_{X_2|X_1=x_1}(x_2) - f_{X_2|X_1\Delta_{k_1}x_1}(x_2)\right) dx_1dx_2 = 0,$$

so if we denote $[4]$= RL(δ_{k_1,k_2}, p), it easily holds that

$$\sum_{k_1,k_2} \text{RL}(\delta_{k_1,k_2}, p) = \text{R}_{\mathbf{X}}(p).$$

□

Note that in the general case we always have the following inequality

$$\Pr\{\mathbf{X} \in \Omega_{\mathbf{X}}(p)\} \geq \Pr\{\mathbf{X} \in \Lambda_{X_i}(p)\} = 2p - 1,$$

as a generalization of the accumulated probability in the real interval quantile. On the other hand, when p tends to 1 it holds that $R_{\mathbf{X}}(p)$ tends to 0.

Corollary 3.1. *Let* $\mathbf{X}$ *be a bivariate random vector with independent components, then* $R_{\mathbf{X}}(p) = 0$.

Corollary 3.2. *Let* $\mathbf{X}$ *be a bivariate random vector with independent components and denote* $L_{\mathbf{X}}(p, \delta_{k_1,k_2})$ *the lateral region*

$$L_{\mathbf{X}}(p, \delta_{k_1,k_2}) = \{\mathbf{x} \in \mathbb{R}^2 : \quad Pr\{\mathbf{X}\Delta_{\delta_{k_1,k_2}}\mathbf{x}\} > p\},$$

then

$$Pr\{\mathbf{X} \in L_{\mathbf{x}}(p, \delta_{k_1,k_2})\} = 1 - p(1 + \ln(1/p)). \tag{4}$$

Remark that it would be interesting to study the necessary condition for $R_{\mathbf{X}}(p)$ equal to 0. Several empirical examples show that the rate $R_{\mathbf{X}}(p)$ modifies the probability in the central region and it depends on properties for the distribution function.

Because $\Omega_{\mathbf{X}}(p)$ is not a bounded region we can not directly use it for detecting outliers, in this way some considerations for outliers in both components x_1 and x_2 are possible. It would be interesting to bound the central region through parallel lines to $x_1 = Q_{X_1}(p)$, $x_1 = Q_{X_1}(1-p)$, $x_2 = Q_{X_2}(p)$ and $x_2 = Q_{X_2}(1-p)$. Note that in the last sense we keep the shape of the central region. Alternatively, because the shape is not a priori chosen, discussions related to skewness, dispersion, concordance and dependence are easier to show in a descriptive way.

3.1 A Note on the Dispersion Study

A concept of bivariate dispersion can be generalized as the classical univariate ordering of quantiles more widely separated studied by Lewis and Thompson (1981). Let $\mathbf{X}$ be a bivariate random vector and let $\Omega_{\mathbf{X}}(p)$, $\Omega_{\mathbf{X}}(q)$ the central regions for $1/2 < p < q < 1$. Intuitively, the concept of dispersion is associated to "the distance" between $\Omega_{\mathbf{X}}(p)$ and $\Omega_{\mathbf{X}}(q)$. Note that this can be obtained through the distance among the multivariate quantiles, that is to say, among the different level curves for each variation. Without loss of generality, we consider the variation $\delta_{0.0}$ and let u such that $1/2 < p < q < u < 1$. Then from Theorem 2.2, the points $\hat{\mathbf{q}}(\delta_{0,0}, \pi_{1,2}, (u, p/u))$ and $\hat{\mathbf{q}}(\delta_{0,0}, \pi_{1,2}, (u, q/u))$ belong to $Q_{\mathbf{X}}(p, \delta_{0,0})$ and $Q_{\mathbf{X}}(q, \delta_{0,0})$ respectively. Observe that the distance

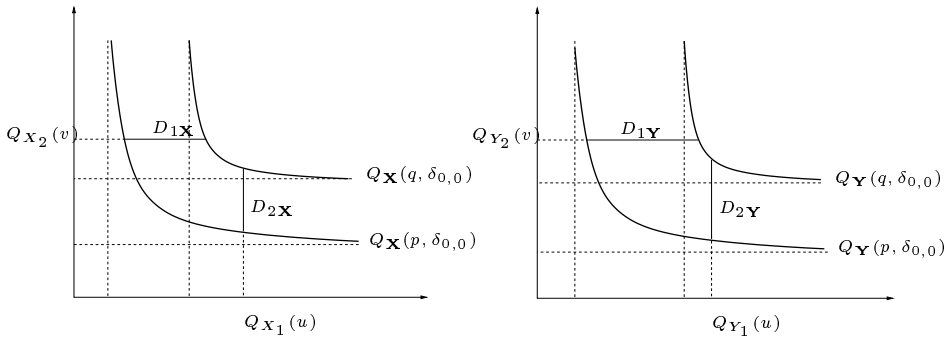
$$\| \hat{\mathbf{q}}(\delta_{0,0}, \pi_{1,2}, (u, p/u)) - \hat{\mathbf{q}}(\delta_{0,0}, \pi_{1,2}, (u, q/u)) \|_2,$$

denoted by $D_{2\mathbf{X}}$, represents the separation between quantiles for the conditional distribution $X_2|_{X_1 \leq Q_{X_1}(u)}$. Otherwise, using the permutation $\pi_{2,1}$ to describe the level curves and let v such that $1/2 < p < q < v < 1$, it holds that

$$D_{1\mathbf{X}} = \mid Q_{X_1|_{X_2 \leq Q_{X_2}(v)}}(q/v) - Q_{X_1|_{X_2 \leq Q_{X_2}(v)}}(p/v) \mid .$$

In the same way, $D_{1\mathbf{X}}$ represents the separation among quantiles for the conditional distribution $X_1|_{X_2 \leq Q_{X_2}(v)}$.

Definitively, if we have two bivariate random vectors $\mathbf{X}$ and $\mathbf{Y}$, we compare quantiles more widely separated for the conditional variables. This one can be interpreted as dispersion in the orthant defined under the variation $\delta_{k_1,k_2} \in \mathbb{V}_2^2$. We see in the Figure 2 that the multivariate quantiles under the variation $\delta_{0,0}$, are more separated for the distribution $\mathbf{Y}$ than for $\mathbf{X}$, that is to say, $D_{1\mathbf{Y}} > D_{1\mathbf{X}}$ and $D_{2\mathbf{Y}} > D_{2\mathbf{X}}$.



(a) Dispersion for $\mathbf{X}$. (b) Dispersion for Y.

Figure 2: Quantiles more widely separated.

4 An Example of Estimation

Let $\mathbf{X_i}$, $i = 1, \cdots, n$ be n independent and identically distributed copies of $\mathbf{X} = (X_1, X_2)$ and denote the empirical distribution function of $\{\mathbf{X_i} : \quad i = 1, \cdots, n\}$ by $F^{(n)}$ and the corresponding $i - th$ marginal distribution by $F_i^{(n)}$, $i = 1, 2$. From Theorem 2.2, using the permutation $\pi_{1,2}$, it holds

$$\begin{aligned}
Q_{\mathbf{X}}(p, \delta_{0,0}) &= \left\{ (Q_{X_1}(u), Q_{X_2|_{X_1 \leq Q_{X_1}(u)}}(p/u)) : \quad \forall u > p \right\}, \\
Q_{\mathbf{X}}(p, \delta_{1,0}) &= \left\{ (Q_{X_1}(u), Q_{X_2|_{X_1 \geq Q_{X_1}(u)}}(p/(1-u))) : \quad \forall u < 1-p \right\}, \\
Q_{\mathbf{X}}(p, \delta_{0,1}) &= \left\{ (Q_{X_1}(u), Q_{X_2|_{X_1 \leq Q_{X_1}(u)}}(1-p/u)) : \quad \forall u > p \right\}, \\
Q_{\mathbf{X}}(p, \delta_{1,1}) &= \left\{ (Q_{X_1}(u), Q_{X_2|_{X_1 \geq Q_{X_1}(u)}}(1-p/(1-u))) : \quad \forall u < 1-p \right\}.
\end{aligned}$$

Obviously, the estimating method will be based on the estimation of the conditional cumulative distribution function and of the conditional quantiles. In this paper is not our purpose to show all asymptotic properties related to estimate the conditional quantiles. From

this one, we will use the well known estimating method through the empirical distribution. Without loss of generality we consider the variation $\delta_{0,0} \in \mathbb{V}_2^2$. Let $\{x_{1i} : \quad i = 1, \cdots, m\}$ be a set of real values which belong to the support of the marginal distribution X_1 and let $u_i = F_1^{(n)}(x_{1i})$, for $i = 1, \cdots, m$, the estimation of the accumulated probability. We denote $\hat{Q}_{X_2|_{X_1 \leq x_{1i}}}(.)$ the estimator of the conditional quantile defined in terms of $F^{(n)}$ and $F_1^{(n)}$. Then, the set

$$\hat{Q}_{\mathbf{X}}(p, \delta_{0,0}) = \left\{ \mathbf{p}_i = (x_{1i}, \hat{Q}_{X_2|_{X_1 \leq x_{1i}}}(p/u_i)) : \quad u_i > p \right\}$$

provides the estimation of a family of points in $Q_{\mathbf{X}}(p, \delta_{0,0})$. Note that to estimate the support of X_1, we take the points $x_{1i} = x_{1(i)}$ as the i-th ordered statistic of the marginal distribution X_1, that is to say $x_{1(1)} \leq \cdots \leq x_{1(n)}$. 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 easier the different regions.

To illustrate the estimation with a real data cloud, we provide an example for two classical variables associated to applied sciences in the environment. It is widely studied the atmospheric concentration levels in an urban area. For this purpose, we consider the sulfur oxide (SO_2) and nitrogen oxide (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. Although there are a lot of factors they should be considered as the wind speed or rain fall, our purpose is to obtain a collection of points which represent the population. We show in the Figure 3 the dispersion diagram where the sulfur oxide is in the horizontal axis and the nitrogen oxide is in the vertical axis.

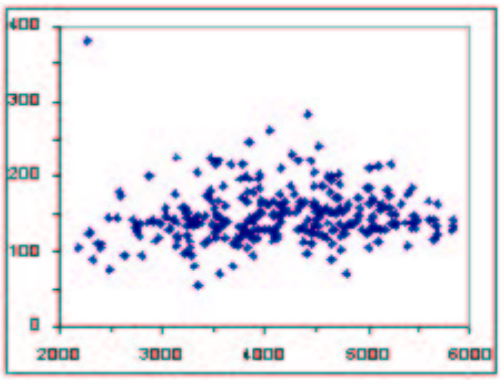
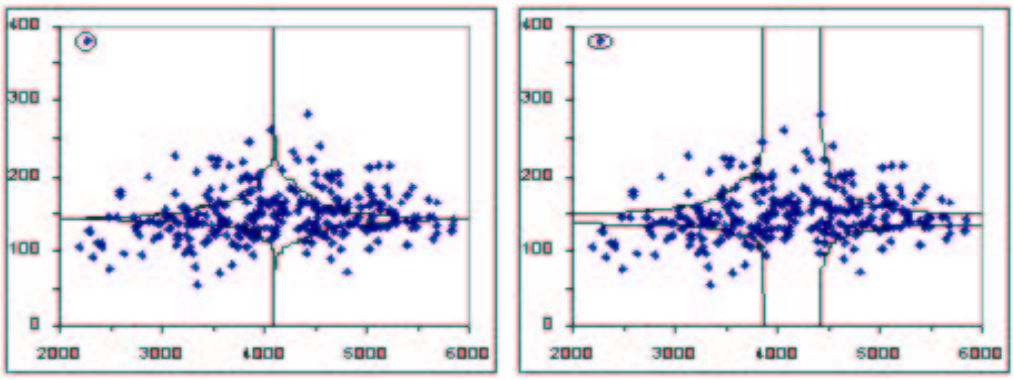


Figure 3: (SO_2, NO_x)

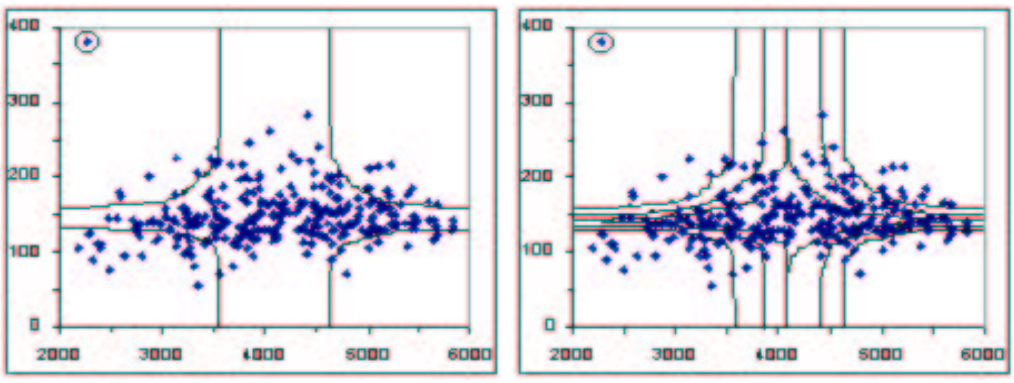
The relationship between the marginal distributions implies the shape of the central region. Figure 4 (a) represents the central region for $p = 0.5$. Observe that there are points in $\Omega_{\mathbf{X}}(0.5)$ with a large NO_x component, so this bivariate distribution is more dispersed in this sense. In Figure 4 (b) and 5 (a) we represent the central regions for $p = 0.6$ and $p = 0.7$. Under the assumption of independent components, using 4, we obtain that

the probabilities of the lateral region, $L_{\mathbf{X}}(p, \delta_{k_1,k_2})$, for $p = 0.5$, $p = 0.6$ and $p = 0.7$ are 15.34%, 9.35% and 5.03% respectively for all variations. Note that these last probabilities do not correspond with the empirical probabilities. The lateral region for the variations $\delta_{1,0}$ and $\delta_{0,1}$ accumulate an inferior percentage than the lateral regions for the variations $\delta_{0,0}$ and $\delta_{1,1}$ -this analysis result of performing a hypothesis test concerning to a simple proportion- thus there exist a relation of dependence between the contaminants with a positive correlation. We also represent in Figure 5 (b) the central region for $p = 0.5$, $p = 0.6$ and $p = 0.7$ simultaneously. We can see that there are directions where the conditional quantiles are more separated than other ones. It is also interesting to show the stability of the central region to existence of outliers. Finally, the central region is centered around the spatial median thus we always obtain a representative data set of the population which can be interpreted as a trimmed region.



(a) $\Omega_{\mathbf{X}}(0.5)$ (b) $\Omega_{\mathbf{X}}(0.6)$

Figure 4: Central Regions.



(a) $\Omega_{\mathbf{X}}(0.7)$ (b) Simultaneously

Figure 5: Central Regions.

Acknowledgments

The authors would like to thank referees for their suggestions on an earlier version of this paper.

References

B. Abdous and R. Theodorescu. Note on the spatial quantile of a random vector. *Statistics and Probability Letters*, 13:333–336, 1992.

J. Averous and M. Meste. Median balls: An extension of the multivariate interquantile intervals to multivariate distributions. *Academic Press*, pages 221–241, 1997.

V. Barnett. The ordering of multivariate data. *J. Roy. Statist. Soc. Ser. A*, 139:318–355, 1976.

P. Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872, 1996.

T. Lewis and J.W. Thompson. Dispersive distributions and the connection between dispersivity and strong unimodality. *Journal Applied Probability*, 18:76–90, 1981.

J.C. Massé and R. Theodorescu. Halfplane trimming for bivariate distributions. *Journal of Multivariate Analysis*, 48:188–202, 1994.

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

H. Oja. Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, 1:327–332, 1983.

P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outliers Detection*. Wiley, New York, 1987.

M. Shaked and J.G. Shanthikumar. *Stochastic Orders and Their Applications*. Academic Press, New York, 1994.

J.W. Tukey. Mathematics and the picturing of data. *In Proceedings International Congress of Mathematicians*, 2:523–531, 1975.

Authors' addresses:

J.M. Fernández-Ponce
University Sevilla
Tarfia s/n
Sevilla, Spain

A. Suárez-Lloréns
University Cadiz
Duque de Nájera 8
11002 Cádiz, Spain

Tel. 956-01-5481
E-mail: alfonso.suarez@uca.es