UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 06-69
Statistic and Econometric Series 19
December 2006

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249849

THE EXPECTED CONVEX HULL TRIMMED REGIONS OF A SAMPLE*

Ignacio Cascos [1]

Abstract

Given a data set in the multivariate Euclidean space, we study regions of central points built by averaging all their subsets with a fixed number of elements. The averaging of these sets is performed by appropriately scaling the Minkowski or elementwise summation of their convex hulls. The volume of such central regions is proposed as a multivariate scatter estimate and a circular sequence algorithm to compute the central regions of a bivariate data set is described.

Keywords: convex hull; depth function; depth-trimmed regions; Minkowski sum; multivariate scatter estimate.

* Work partially supported by the Spanish Ministry of Education and Science under grant MTM2005-02254.

[1] Department of Statistics, Universidad Carlos III de Madrid, Av. Universidad 30, E-28911 Leganés (Madrid), Spain, e-mail: ignacio.cascos@uc3m.es.

The expected convex hull trimmed regions of a sample*

Ignacio Cascos†
Department of Statistics, Universidad Carlos III de Madrid,
Av. Universidad 30, E-28911 Leganés (Madrid), Spain
Tel: +34-916248750, Fax: +34-916249430,
E-mail: ignacio.cascos@uc3m.es

Abstract

Given a data set in the multivariate Euclidean space, we study regions of central points built by averaging all their subsets with a fixed number of elements. The averaging of these sets is performed by appropriately scaling the Minkowski or elementwise summation of their convex hulls. The volume of such central regions is proposed as a multivariate scatter estimate and a circular sequence algorithm to compute the central regions of a bivariate data set is described.

Keywords: convex hull, depth function, depth-trimmed regions, Minkowski sum, multivariate scatter estimate

1 Introduction

The lack of a natural order in the multivariate Euclidean space is an obstacle to translate a number of univariate data analytic tools to the multivariate setting. Several authors have overcome this problem introducing a center-outward ordering with respect to a multivariate probability distribution or a data cloud. Depth functions are a key issue in this approach, they assign a point its degree of centrality with respect to a probability distribution or a data cloud. They were first proposed by Tukey (1975) and have been extensively studied in the last years, see Liu (1990) for a depth proposal based on random simplices, Liu et al. (1999) for statistical applications of depth functions and Zuo and Serfling (2000a) for a comprehensive treatment of several notions of data depth.

The level sets of a depth function constitute the so-called depth-trimmed regions. They are nothing but sets of central points with respect to a data cloud or a probability distribution. Among others, Massé and Theodorescu (1994), Koshevoy and Mosler (1997) and

*Extended version of the conference paper Cascos (2006) presented at the 17th Conference of the European Regional Section of the International Association for Statistical Computing (COMPSTAT 2006, Rome, Italy, August 28 – September 1, 2006).

†Supported by the Spanish Ministry Education and Science under grant MTM2005-02254.

Cascos and López-Díaz (2005) have proposed and studied particular instances of families of central regions. Meanwhile, Zuo and Serfling (2000b) systematically obtain properties of several families of depth-trimmed regions from the ones of their generating depth functions.

We will propose a new family of central regions. Given a sample $\boldsymbol{x}_1, \dots \boldsymbol{x}_n \in \mathbf{R}^d$, its mean value $\overline{\boldsymbol{x}}$ is the sum of all the points from the sample multiplied by n^{-1}. Imagine we want to obtain a set that contains $\overline{\boldsymbol{x}}$ and is central with respect to the data cloud. We can build a natural candidate taking all the segments that join two distinct points from the sample and averaging them. The way we average them is by taking their Minkowski (or elementwise) sum weighted by the inverse of the number of segments, that is, we multiply the sum by $\binom{n}{2}^{-1}$. Such a procedure can be generalized from the set of all pairs of points to the set of all k-tuples for any $1 \leq k \leq n$.

In Section 2, we will introduce the expected convex hull trimmed regions of a sample, describe its main properties and characterize its extreme points. In Section 3, the volumes of the expected convex hull regions are proposed as multivariate scatter estimates. We use the classical method of the circular sequence to develop an algorithm of complexity $O(n^2 \log n)$ to compute arbitrary expected convex hull trimmed regions of a bivariate data set in Section 4. Some concluding remarks, including alternative central regions, are presented in Section 5.

2 Expected convex hull trimming

Before introducing the new central regions, we will briefly recall some definitions that are needed as well as some notation.

Points in $\mathbf{R}^d$ will be denoted with boldface characters and square brackets are used to specify their components, i.e., $\boldsymbol{x} \in \mathbf{R}^d$ is alternatively given by $\boldsymbol{x} = (\boldsymbol{x}[1], \dots, \boldsymbol{x}[d])$.

Given a set of points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbf{R}^d$, we denote its *convex hull*, the smallest convex set that contains them, by $\mathrm{co}\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$. For two sets $F, G \subset \mathbf{R}^d$, its *Minkowski sum* or *elementwise sum* is given by

$$F \oplus G := \{\boldsymbol{x} + \boldsymbol{y} : \boldsymbol{x} \in F, \boldsymbol{y} \in G\}.$$

The Minkowski sum of two segments centred at the origin is plotted in Figure 1 below.

For $\lambda \in \mathbf{R}$, we define $\lambda F := \{\lambda \boldsymbol{x} : \boldsymbol{x} \in F\}$. Finally, relation $\leq$ is understood componentwisely, i.e. $\boldsymbol{x} \leq \boldsymbol{y}$ if $\boldsymbol{x}[i] \leq \boldsymbol{y}[i]$ for all $1 \leq i \leq d$.

Definition 1. *Given $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbf{R}^d$ and $1 \leq k \leq n$, their* expected convex hull trimmed region of level k *is given by*

$$\mathrm{CD}^k(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) := \binom{n}{k}^{-1} \bigoplus_{1 \leq i_1 < \dots < i_k \leq n} \mathrm{co}\{\boldsymbol{x}_{i_1}, \dots, \boldsymbol{x}_{i_k}\},$$

that is, the Minkowski summation of the convex hulls of all the subsets of k elements of $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ multiplied by the inverse of the number of all such subsets.

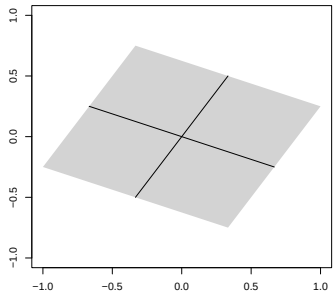

Figure 1: Minkowski sum (shaded area) of two segments centred at the origin.

The reason we name these sets *expected convex hull trimmed regions* is that through the Minkowski sum and the multiplication by a scalar, we are averaging (finding expectations of) convex hulls. Whenever it is possible, we will write CD^k shortly instead of $\mathrm{CD}^k(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

In Figure 2, we have plotted the expected convex hull region of a data set of level $k = 2$. In three different frames, we have respectively depicted $n = 5$ points in the plane, the 10 segments joining them and CD^2.

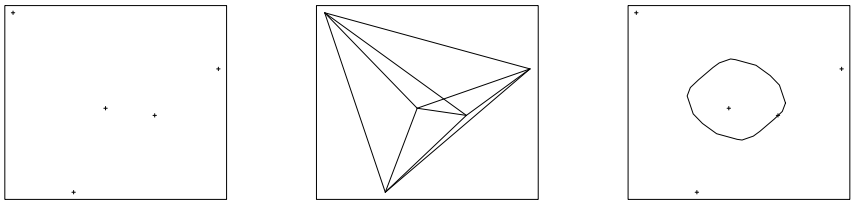

Figure 2: Building CD^2.

2.1 Main properties of the expected convex hull trimmed regions

We will briefly summarize the main properties of the expected convex hull trimmed regions.

Theorem 2. *Given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbf{R}^d$, the expected convex hull regions satisfy*

(i) affine equivariance: *for any nonsingular matrix $A \in \mathbf{R}^{d\times d}$ and any point $\boldsymbol{b} \in \mathbf{R}^d$, we have $\mathrm{CD}^k(A\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, A\boldsymbol{x}_n + \boldsymbol{b}) = A\mathrm{CD}^k(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) + \boldsymbol{b}$;*

(ii) nested: *if $k_1 \leq k_2$, then $\mathrm{CD}^{k_1} \subset \mathrm{CD}^{k_2}$;*

(iii) convex: *CD^k is convex;*

(iv) compact: CD^k *is compact;*

(v) if $k = 1$: $\mathrm{CD}^1 = \{\overline{\boldsymbol{x}}\}$ *;*

(vi) if $k = n$: $\mathrm{CD}^n = \mathrm{co}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.

Properties (i)–(iv) in Theorem 2 are similar to those studied for depth-trimmed regions in Theorem 3.1 in Zuo and Serfling (2000b), except for the fact that the inclusion of the nesting property (ii) is here reversed with respect to the order relation in parameter k. Properties (iii) and (iv) follow from the fact that CD^k is the Minkowski sum of a finite number of polytopes and thus, it is a polytope.

Theorem 2 (i) holds true if instead of a square nonsingular matrix, any $A \in \mathbf{R}^{m\times d}$ and $\boldsymbol{b} \in \mathbf{R}^m$ are taken. As a consequence, the projection of the expected convex hull regions on any subset of coordinates equals the expected convex hull regions of the projected data (on the same set of coordinates).

In Figure 3 we have depicted the contour plots of the expected convex hull regions of some real data. The data corresponds to the results of the 30 athletes that competed in the 2004 Olympics Decathlon in long jump (in meters, axis X) and in the 100m race (in seconds, axis Y) and has been obtained from `http://www.athens2004.com`.

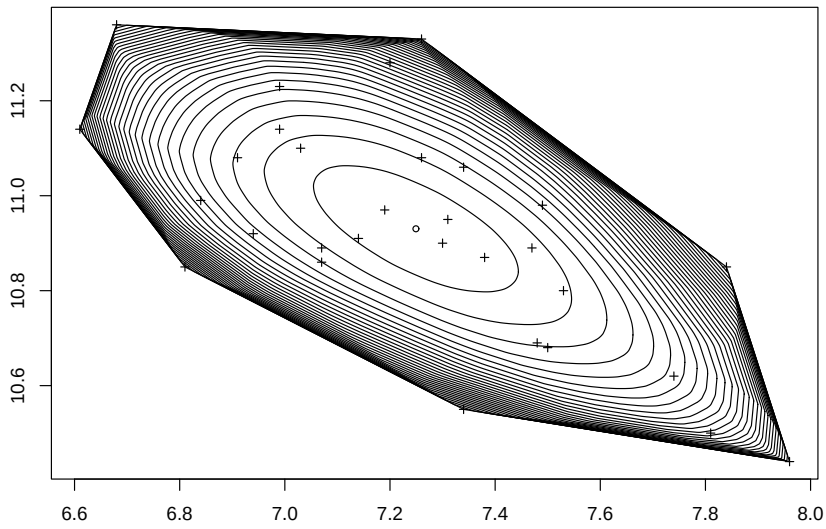

Figure 3: Contour plots of the expected convex hull trimmed regions of the decathlon data set with k ranging from 1 to 30.

Depth function. It is possible to define a *depth function* in terms of the new regions. For a fixed data set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and any $\boldsymbol{x} \in \mathbf{R}^d$, we can define the function

$$\mathrm{CD}(\boldsymbol{x}) := \begin{cases} \min\{k : \boldsymbol{x} \in \mathrm{CD}^k\}^{-1} & \text{if } x \in \mathrm{co}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \\ 0 & \text{otherwise} \end{cases} .$$

The function CD is a bounded, non-negative mapping and from Theorem 2 it follows that it satisfies the properties of *affine invariance*, *maximality at* $\overline{\boldsymbol{x}}$, in fact $\mathrm{CD}(\overline{\boldsymbol{x}}) = 1$, *decreasing in arrays from the mean value* and *vanishing at infinity*. Therefore, CD is a depth function in the sense of Definition 2.1 in Zuo and Serfling (2000a). See Figure 4 for a graphical example of the new depth function.

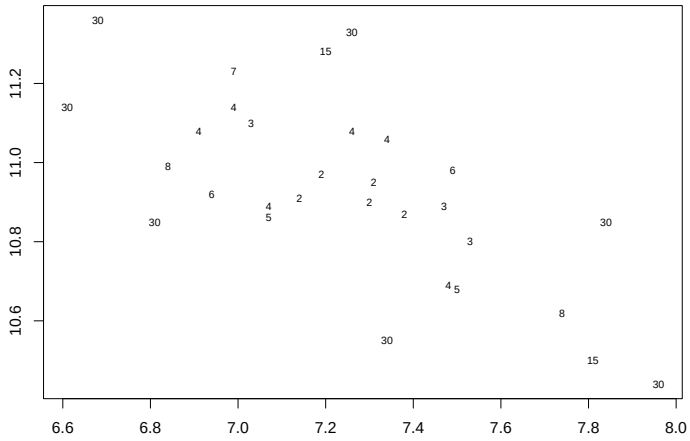

Figure 4: Inverse of the expected convex hull depth for the 30 points in the decathlon data set.

2.2 Additional properties

Monotonicity. Depth-trimmed regions are set-valued location estimates and thus the affine equivariance considered in Theorem 2 (i) is an important requirement. Among other classical requirements for location estimates, it is also important to analyze their behaviour when considered on two ordered data sets, see condition (i) for a univariate *location parameter* in Bickel and Lehmann (1975).

Let $\mathbf{R}^d_+$ be the positive orthant and $\mathbf{R}^d_-$ the negative orthant. That is, $\boldsymbol{x} \in \mathbf{R}^d_+$ (resp. $\boldsymbol{x} \in \mathbf{R}^d_-$) if $\boldsymbol{x}[i] \geq 0$ (resp. $\boldsymbol{x}[i] \leq 0$) for all $1 \leq i \leq d$.

Proposition 3. *If* $\boldsymbol{x}_i \leq \boldsymbol{y}_i$ *for* $1 \leq i \leq n$*, the following two relations hold for any* $1 \leq k \leq n$*,*

(i) $\mathrm{CD}^k(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \subset \mathrm{CD}^k(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \oplus \mathbf{R}^d_+$ *;*

(ii) $\mathrm{CD}^k(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \subset \mathrm{CD}^k(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n) \oplus \mathbf{R}^d_-$.

From Proposition 3 (i) it follows that given any point $\boldsymbol{y} \in \mathrm{CD}^k(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)$ there is always a point $\boldsymbol{x} \in \mathrm{CD}^k(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ such that $\boldsymbol{x} \leq \boldsymbol{y}$, meanwhile from (ii) it follows that given any point $\boldsymbol{x} \in \mathrm{CD}^k(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ there is always a point $\boldsymbol{y} \in \mathrm{CD}^k(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)$ such that $\boldsymbol{x} \leq \boldsymbol{y}$.

Minkowski subadditivity. The convex hull satisfies a special type of subadditivity. Given two data sets with the same size, $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n \in \mathbf{R}^d$ it holds

$$\mathrm{co}\{\boldsymbol{x}_1+\boldsymbol{y}_1,\ldots,\boldsymbol{x}_n+\boldsymbol{y}_n\} \subset \mathrm{co}\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\} \oplus \mathrm{co}\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n\}\,.$$

An immediate consequence of such subadditivity is that.

Proposition 4. *For any* $1 \leq k \leq n$,

$$\mathrm{CD}^k(\boldsymbol{x}_1+\boldsymbol{y}_1,\ldots,\boldsymbol{x}_n+\boldsymbol{y}_n) \subset \mathrm{CD}^k(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \oplus \mathrm{CD}^k(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)\,.$$

In order to obtain a graphical explanation of Proposition 4, we have simulated a sample $(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_{20},\boldsymbol{y}_{20}) \in \mathbf{R}^4$ and present two frames in Figure 5. On the frame on the left, we have plotted two data sets $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{20}$ (as $+$) and $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{20}$ (as $\triangle$) together with the contours of their respective expected convex hull regions of level $k = 10$. On the frame on the right, we have plotted the data set $\boldsymbol{x}_1+\boldsymbol{y}_1,\ldots,\boldsymbol{x}_{20}+\boldsymbol{y}_{20}$ together with the contour of $\mathrm{CD}^{10}(\boldsymbol{x}_1+\boldsymbol{y}_1,\ldots,\boldsymbol{x}_{20}+\boldsymbol{y}_{20})$ and a shaded region that corresponds to $\mathrm{CD}^{10}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{20}) \oplus \mathrm{CD}^{10}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{20})$.

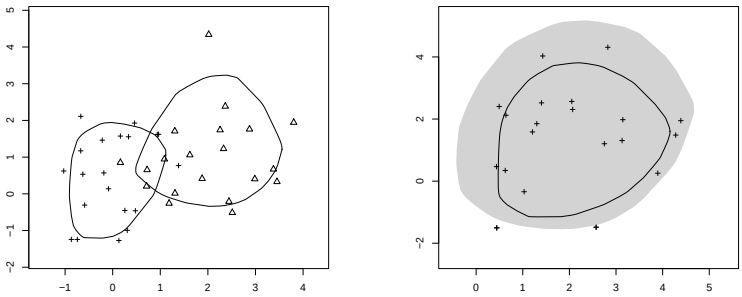

Figure 5: Central region of the sum of two samples and Minkowski sum of two central regions.

The subadditivity in the Minkowski sense has important implications when using depth-trimmed regions to quantify the financial risk of a vector portfolio, see Cascos and Molchanov (2006).

Characterization result.

Proposition 5. *The family of sets* CD^k *for* $1 \leq k \leq n$ *determine the set of points* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

In the univariate case, $d = 1$, the set CD^n is an interval whose upper extreme is the greatest point from the sample. In general, the upper extreme of CD^{n-k+1} is a linear combination of the k greatest points from the sample and thus all the points from the sample can be retrieved in a simple way.

In the multivariate case, we have all univariate projections of the points from the sample and thus it is possible to retrieve the original data.

2.3 Extreme points

Since the set CD^k is a polytope, it is the convex hull of a finite number of its boundary points. The minimal set of such points is constituted by its extreme points. These can be expressed in terms of summations of the extreme points of each of the Minkowski summands $\mathrm{co}\{\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_k}\}$.

For any set $F \subset \mathbf{R}^d$, its support function is defined as $h(F, \boldsymbol{u}) := \sup\{\langle \boldsymbol{x}, \boldsymbol{u}\rangle : \boldsymbol{x} \in F\}$ for any $\boldsymbol{u} \in \mathbf{R}^d$. Since the support function is linear on the set-valued argument and $h(\mathrm{co}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}, \boldsymbol{u}) = \max\{\langle \boldsymbol{x}_1, \boldsymbol{u}\rangle, \ldots, \langle \boldsymbol{x}_n, \boldsymbol{u}\rangle\}$, we obtain

$$h(\mathrm{CD}^k, \boldsymbol{u}) = \binom{n}{k}^{-1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \max\{\langle \boldsymbol{x}_{i_1}, \boldsymbol{u}\rangle, \ldots, \langle \boldsymbol{x}_{i_k}, \boldsymbol{u}\rangle\}\,.$$

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be pairwise distinct.

Directions in the d-dimensional Euclidean space can be interpreted as elements of S^{d-1}, the unit sphere in $\mathbf{R}^d$. If $\boldsymbol{u} \in S^{d-1} \setminus L$ where L is the set of directions that are orthogonal to the line through any two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with $i \neq j$, then there is $\pi_{\boldsymbol{u}}$ a permutation of $\{1, \ldots, n\}$ with

$$\langle \boldsymbol{x}_{\pi_{\boldsymbol{u}}(1)}, \boldsymbol{u}\rangle < \langle \boldsymbol{x}_{\pi_{\boldsymbol{u}}(2)}, \boldsymbol{u}\rangle < \ldots < \langle \boldsymbol{x}_{\pi_{\boldsymbol{u}}(n)}, \boldsymbol{u}\rangle\,.$$

We will characterize the extreme point of CD^k in the direction of $\boldsymbol{u}$. Given $j \geq k$, the number of sets of k points such that $\langle \boldsymbol{x}_{\pi_{\boldsymbol{u}}(j)}, \boldsymbol{u}\rangle$ is greater than any other $\langle \boldsymbol{x}_i, \boldsymbol{u}\rangle$ is $\binom{j}{k} - \binom{j-1}{k}$ where $\binom{k-1}{k} = 0$. Since

$$\binom{j}{k} - \binom{j-1}{k} = \binom{j-1}{k-1}\,,$$

the point

$$\boldsymbol{x}(\boldsymbol{u}) = \binom{n}{k}^{-1} \sum_{j=k}^{n} \binom{j-1}{k-1} x_{\pi_{\boldsymbol{u}}(j)}$$

is the extreme point of CD^k such that $h(\mathrm{CD}^k, \boldsymbol{u}) = \langle \boldsymbol{x}(\boldsymbol{u}), \boldsymbol{u}\rangle$. It can alternatively be written as

$$\boldsymbol{x}(\boldsymbol{u}) = \frac{k}{n^{(k)}} \sum_{j=k}^{n} (j-1)^{(k-1)} \boldsymbol{x}_{\pi_{\boldsymbol{u}}(j)}\,, \tag{1}$$

where $n^{(k)} = \prod_{j=0}^{k-1}(n-j)$ if $n < k$ and $k^{(k)} = k!$. Notice that the same notation is used in $(j-1)^{(k-1)}$.

Theorem 6. *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbf{R}^d$ be pairwise distinct, the set of extreme points of CD^k is $\{\boldsymbol{x}(\boldsymbol{u}) : \boldsymbol{u} \in S^{d-1} \setminus L\}$.*

By considering all possible orderings of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, we obtain

$$\mathrm{CD}^k = \mathrm{co}\left\{ \frac{k}{n^{(k)}} \sum_{j=k}^{n} (j-1)^{(k-1)} \boldsymbol{x}_{\pi(j)} : \pi \in \mathbf{P}_n \right\}, \tag{2}$$

where $\mathbf{P}_n$ is the set of permutations of $\{1, \ldots, n\}$. Unfortunately formula (2) is not efficient to compute the extreme points of CD^k since many of the points described are not extreme points.

In $\mathbf{R}^2$, the number of directions involved in Theorem 6 is at most $n(n-1)$ and the circular sequence algorithm built in Section 4 will take advantage of this.

3 Multivariate scatter estimates

The expected convex hull regions provide us with information about the scatter of the data. The bigger the regions are, the more scattered the sample is. Consequently, their volume can be used to define *multivariate scatter estimates*. Although the volume of the central region of any level is valid as a scatter estimate, we will specialize our discussion to the central regions of level $k = 2$ because of the ease of the computation of their volume. We will use notation V for such scatter estimate,

$$\mathrm{V}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) := \mathrm{vol}_d\big(\mathrm{CD}^2\big) .$$

The set CD^2, see Definition 1, is a Minkowski sum of segments, i.e., a zonotope. Zonotopes have already been studied in connection with data analysis by Koshevoy and Mosler (1997) and Mosler (2002). In the construction of CD^2 it is possible to perform the Minkowski summation over all $i \neq j$, in such a case the overall sum must be rescaled and we obtain

$$\mathrm{CD}^2 = \frac{1}{n(n-1)} \bigoplus_{i \neq j} \mathrm{co}\{\boldsymbol{x}_i, \boldsymbol{x}_j\} . \tag{3}$$

Notice that the possibility that $i = j$ in equation (3) would not change the volume of the resulting set. Finally, the d-dimensional volume of the set CD^2, is derived from (3) and the formula of the volume of a zonotope, see Mosler (2002),

$$\mathrm{V}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{1}{d! n^d (n-1)^d} \sum_{i_1=1}^{n} \sum_{j_1=1}^{n} \cdots \sum_{i_d=1}^{n} \sum_{j_d=1}^{n} \left| \det(\boldsymbol{x}_{i_1} - \boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{i_d} - \boldsymbol{x}_{j_d}) \right| , \tag{4}$$

where $\det(\boldsymbol{x}_{i_1} - \boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{i_d} - \boldsymbol{x}_{j_d})$ is the determinant of the matrix whose columns are $\boldsymbol{x}_{i_1} - \boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{i_d} - \boldsymbol{x}_{j_d}$.

Let us now see how the scatter estimate changes if a new system of coordinates is taken.

Proposition 7. *For any matrix* $A \in \mathbf{R}^{d\times d}$ *and any point* $\boldsymbol{b} \in \mathbf{R}^d$,

$$\mathrm{V}(A\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, A\boldsymbol{x}_n + \boldsymbol{b}) = \det(A)\mathrm{V}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\,.$$

In particular, for $a \in \mathbf{R}$ we have that $\mathrm{V}(a\boldsymbol{x}_1+\boldsymbol{b}, \ldots, a\boldsymbol{x}_n+\boldsymbol{b}) = |a|^d\mathrm{V}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, which, in the univariate case, means that V is a *measure of dispersion* in the sense of Bickel and Lehmann (1976).

Proposition 7 plays also an important role when comparing the scatter of normalized samples. Such is the case for two data sets in $\mathbf{R}^d_+$ whose coordinates are measured in different units. It is usually more reasonable to normalize each data set dividing each coordinate by its mean value. In such a case, A would be a diagonal matrix whose non-trivial elements are $(1/\overline{\boldsymbol{x}}[1], \ldots, 1/\overline{\boldsymbol{x}}[d])$. The comparison would be possible using formula (4) and multiplying each volume by the determinant of the corresponding matrix.

Remark 8. Proposition 7 holds true if the central regions of any other level $k > 2$ are used in order to quantify the scatter.

Remark 9. A consequence of the expected convex hull trimmed regions of level 2 being zonotopes, see equation (3), is that they will always be centrally symmetric. Nevertheless, we understand that CD^2 has interest of its own as a descriptive statistic although we are working with skewed data. Notice that, in such a case, the central regions of higher levels will not be symmetric.

4 Algorithm

Given a bivariate data set, it is possible to consider all the orderings of its 1-dimensional projections in an efficient way. This can be done through a circular sequence, see Chapter 2 in Edelsbrunner (1987). Ruts and Rousseeuw (1996) and Dyckerhoff (2000) have used circular sequences to build algorithms to compute the extreme points of the halfspace and zonoid trimmed regions for a bivariate data set. The algorithm we present here is based on that of Dyckerhoff (2000), where we have modified Step 5 to obtain our family of central regions.

Step 1. Store the x-coordinates of the data points in an array X and the y-coordinates in an array Y, so that each point of the data cloud can be written as $(\mathrm{X}[i], \mathrm{Y}[i])$ for certain i. These arrays must be ordered in such a way that $\mathrm{X}[i] \leq \mathrm{X}[i+1]$ and if $\mathrm{X}[i] = \mathrm{X}[i+1]$, then $\mathrm{Y}[i] > \mathrm{Y}[i+1]$. Initialize the arrays ORD and RANK with the values from 1 to n. Initialize the arrays EXT1 and EXT2 to store the extreme points.

Step 2. In an array ANGLE of length $\binom{n}{2}$, store the angle between the first coordinate axis and the normal vector of the line through the points $(\mathrm{X}[i], \mathrm{Y}[i])$ and $(\mathrm{X}[j], \mathrm{Y}[j])$, take always angles in $[0, \pi)$. This array is finally sorted in an increasing way.

Step 3. Find all successive entries of the array ANGLE with the same value, these will correspond to sets of collinear points.

Step 4. Reverse the order of each set of collinear points in the arrays ORD and RANK. If $(\mathrm{X}[i], \mathrm{Y}[i])$ and $(\mathrm{X}[j], \mathrm{Y}[j])$ are collinear, interchange the values of ORD[RANK[i]] and ORD[RANK[j]] and afterwards RANK[i] and RANK[j]. If further (X[p], Y[p]), $(\mathrm{X}[q], \mathrm{Y}[q])$ and $(\mathrm{X}[r], \mathrm{Y}[r])$ with $p < q < r$ are collinear, interchange the values of ORD[RANK[p]], ORD[RANK[q]] and ORD[RANK[r]] with the ones of ORD[RANK[r]], ORD[RANK[q]] and ORD[RANK[p]] respectively and further RANK[p], RANK[q] and RANK[r] with RANK[r], RANK[q] and RANK[p]. Continue so forth with sets of more than three collinear points.

Step 5. If any RANK[i] $\geq k$ has been modified in Step 4, append

$$\sum_{j=k}^{n} (j-1)^{(k-1)} (\mathrm{X}[\mathrm{ORD}[j]], \mathrm{Y}[\mathrm{ORD}[j]]) \tag{5}$$

at the end of the array EXT1.

If any RANK[i] $\leq n-k$ has been modified in Step 4, append

$$\sum_{j=k}^{n} (j-1)^{(k-1)} (\mathrm{X}[\mathrm{ORD}[n+1-j]], \mathrm{Y}[\mathrm{ORD}[n+1-j]]) \tag{6}$$

at the end of the array EXT2.

Notice that each new entry of EXT1 or EXT2 can be computed in terms of the previous one and the entries modified in the array ORD in Step 4 and thus it is not necessary to perform the summation of $n-k+1$ terms.

Step 6. If there is any angle left in the array ANGLE continue with Step 3.

Step 7. Multiply the entries of EXT1 and EXT2 by $k/n^{(k)}$ and concatenate both arrays to obtain the sequence of extreme points in counterclockwise order.

The sorting of the array ANGLE in Step 2 has complexity $O(n^2 \log n)$ and it determines the overall complexity of the algorithm.

5 Conclusion

5.1 Alternative central regions

The set CD^k is build as a U-statistic, i.e., we consider all subsets from the sample with a fixed size. An alternative family of central regions is obtained if the points from the sample are selected at random and with replacement, that is, for $k \geq 1$ we consider all combinations of k integers $1 \leq i_1, i_2, \ldots, i_k \leq n$,

$$\frac{1}{n^k} \bigoplus_{i_1=1}^{n} \cdots \bigoplus_{i_k=1}^{n} \mathrm{co}\{\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_k}\}\,. \tag{7}$$

For $k = 1$, applying (7), we obtain again the singleton $\{\overline{\boldsymbol{x}}\}$ and for other values of k we obtain central regions that are different to the expected convex hull ones. The extreme points would be now

$$\boldsymbol{x}(\boldsymbol{u}) = \frac{1}{n^k} \sum_{j=1}^{n} \left(j^k - (j-1)^k \right) \boldsymbol{x}_{\pi_{\boldsymbol{u}}(j)} \,. \tag{8}$$

If we replace $(j-1)^{(k-1)}$ with $j^k - (j-1)^k$ in equations (5) and (6) of Step 5 and $k/n^{(k)}$ with n^{-k} in Step 7, the algorithm from Section 4 can be used to compute the extreme points of (7). These changes are due to the differences between the extreme points in equations (1) and (8).

5.2 Financial application

In Section 2.2 it has been mentioned that depth-trimmed regions can be applied to asses the financial risk of a vector portfolio. This is discussed in detail in Cascos and Molchanov (2006). The Minkowski subadditivity plays a crucial role when producing coherent, i.e., subadditive risk measures, see Artzner et al. (1999) for univariate coherent risk measures and Jouini et al. (2004) for coherent risk measures of vector portfolios.

5.3 Final remarks

The aim of this paper is building central regions of a data sample by averaging convex hulls of subsets of its points through Minkowski summation. Our effort has focused on the description of the expected convex hull regions and the development of an algorithm for efficiently finding their extreme points for a bivariate data set. However, the set of extreme points described in Theorem 6 holds in any dimension. It is thus possible to built an approximate algorithm for dimensions greater than 2 considering only a random subset of all directions as Dyckerhoff (2000) briefly discusses for the zonoid trimming. With respect to the population counterparts of these central regions, they can be defined in terms of the selection expectation for random sets, see Molchanov (2005), and they will be the issue of future research.

References

[1] Artzner Ph, Delbaen F, Eber JM, Heath D (1999) Coherent measures of risk. Math Finance 9:203–228.

[2] Bickel PJ, Lehmann EL (1975) Descriptive statistics for nonparametric models. II. Location. Ann Statist 3:1045–1069.

[3] Bickel PJ, Lehmann EL (1976) Descriptive statistics for nonparametric models. III. Dispersion. Ann Statist 4:1139–1158.

[4] Cascos I (2006) Expected convex hull trimming of a data set. In: Rizzi A, Vichi M (eds) COMPSTAT 2006. Proceedings in Computational Statistics. Physica-Verlag, Heidelberg, pp 673–680.

[5] Cascos I, López-Díaz M (2005) Integral trimmed regions. J Multivariate Anal 96:404–424.

[6] Cascos I, Molchanov I (2006) Multivariate risks and depth-trimmed regions. Arxiv:math.PR/0606520.

[7] Dyckerhoff R (2000) Computing zonoid trimmed regions of bivariate data sets. In: Bethlehem J, van der Heijden P (eds) COMPSTAT 2000. Proceedings in Computational Statistics. Physica-Verlag, Heidelberg, pp 295–300.

[8] Edelsbrunner H (1987) Algorithms in Combinatorial Geometry. Springer-Verlag, Berlin.

[9] Jouini E, Meddeb M, Touzi N (2004) Vector-valued coherent risk measures. Finance Stoch 8:531–552.

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

[11] Liu RY (1990) On a notion of data depth based on random simplices. Ann Statist 18:405–414

[12] Liu RY, Parelius JM, Singh K (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann Statist 27:783–858.

[13] Massé JC, Theodorescu R (1994) Halfplane trimming for bivariate distributions. J Multivariate Anal 48:188–202.

[14] Molchanov I (2005) Theory of Random Sets. Springer-Verlag, London.

[15] Mosler K (2002) Multivariate dispersion, central regions and depth. The lift zonoid approach. Lecture Notes in Statistics, 165. Springer-Verlag, Berlin.

[16] Ruts I, Rousseeuw PJ (1996) Computing depth contours of bivariate point clouds. Comput Statist Data Anal 23:153–168.

[17] Tukey JW (1975) Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians. Vancouver, pp 523–531.

[18] Zuo Y, Serfling R (2000a) General notions of statistical depth function. Ann Statist 28:461–482.

[19] Zuo Y, Serfling R (2000b) Structural properties and convergence results for contours of sample statistical depth functions. Ann Statist 28:483–499.