Statistical Papers (2023) 64:2183–2205
https://doi.org/10.1007/s00362-022-01380-2

REGULAR ARTICLE

# The zonoid region parameter depth

**Ignacio Cascos**[1] · **Giuseppe Pandolfo**[2] · **Beatriz Sinova**[3]

Received: 8 November 2021 / Revised: 4 November 2022 / Accepted: 25 November 2022 /
Published online: 13 December 2022


**Abstract**
A new concept of depth for central regions is introduced. The proposed depth notion assesses how well an interval fits a given univariate distribution as its zonoid region of level 1/2, and it is extended to the multivariate setting by means of a projection argument. Since central regions capture information about location, scatter, and dependency among several variables, the new depth evaluated on an empirical zonoid region quantifies the degree of similarity (in terms of the features captured by central regions) of the corresponding sample with respect to some reference distribution. Applications to statistical process control and the joint monitoring of multivariate and interval-valued data in terms of location and scale are presented.

**Keywords** Data depth · Parameter depth · Random interval · Zonoid depth

## 1 Introduction

In multivariate statistics, the term *depth*, or more specifically *data depth*, refers to the centrality of an observation with respect to a probability distribution or a data cloud (see Liu 1990; Liu et al. 1999; Mosler 2002; Tukey 1975; Zuo and Serfling 2000). For a given $p$-dimensional distribution $F$, a data depth function assigns a scalar value in the unit interval to each point $\boldsymbol{x}$ in $\mathbb{R}^p$. Such a value measures the centrality of $\boldsymbol{x}$ with respect to $F$. The points that are more central with regard to the distribution assume

✉ Ignacio Cascos
ignacio.cascos@uc3m.es

Giuseppe Pandolfo
giuseppe.pandolfo@unina.it

Beatriz Sinova
sinovabeatriz@uniovi.es

[1] Department of Statistics, Universidad Carlos III de Madrid, 28911 Madrid, Spain

[2] Department of Economics and Statistics, University of Naples Federico II, 80125 Naples, Italy

[3] Department of Statistics, OR, and DM, University of Oviedo, 33071 Oviedo, Spain

high depth values, while peripheral points (again with respect to the distribution) have low depths. The level sets of a data depth function, that is, the sets of points whose depth at least matches some given number, are known as *central* or *depth regions*. It is clear that a depth region conveys valuable information about a distribution: its centremost point can be used as a location parameter, its volume (or surface area, or mean-width...) as a scatter parameter, and it also provides information about the dependency among components or the skewness of the distribution (Liu et al. 1999; Mosler 2013).

On the other hand, a *parameter depth* quantifies how well an element of a parameter space fits a given distribution as its parameter of some given kind, see (Rousseeuw and Hubert 1999) for the regression depth, Mizera and Müller (2004) for the location-scale depth, or Chen et al. (2018) and Paindaveine and Van Bever (2018) for the (scatter) matrix depth. Notice that notions of data depth arise when location parameters are considered in this setting.

In the present manuscript, we consider a central region as parameter and evaluate how well a set adjusts to a reference distribution as its central region. This way we obtain a new notion of depth with respect to either univariate or multivariate distributions whose argument is an interval or a set, respectively. The evaluation of such a depth on the central region of a sample (or distribution) allows us to assess the degree of similarity of this sample (or distribution) with respect to some other reference distribution. This proposal opens a new room to compare distributions in terms of those features captured by their central regions. The closer the central regions of both distributions are, the higher the depth becomes, but notice that a maximal depth value (depth equal to 1) does not guarantee that the distributions are identical.

In the case of interval-valued data, we consider its bivariate representation in terms of endpoints and assess the fit of an observation by means of the zonoid data depth (Koshevoy and Mosler 1997). Further, taking advantage of the new construction, we evaluate how well a sample of intervals adjusts a (random interval) reference distribution by assessing the fit of a central region of the bivariate representation of the sample of intervals with respect to the given reference distribution. Alternative notions of depth for interval-valued data and more general set-valued data are described in Cascos et al. (2021) and to fuzzy data in González-De la Fuente et al. (2022) and Sinova (2022).

Data depth-based nonparametric multivariate analysis techniques have been found to be attractive for building control charts. These latter are a graphical tool commonly employed in statistical quality control to monitor the evolution of a process by means of samples of a quality characteristic. The monitoring is based on the comparison of some statistic that captures the most relevant features of the characteristic with a prescribed control limit (see the monographs Montgomery 2013; Ryan 2011). We illustrate the relevance of the newly introduced depth notion by using it as the charting statistic of several control charts. Specifically, we present Phase I applications, whose goal is to detect anomalies over a set of trial samples by declaring as out-of-control all those samples whose associated depth is below the control limit. Once the out-of-control trial samples are deleted, a polished historical dataset is formed out of the remaining ones and, in Phase II, ongoing monitoring is performed by assessing the

depth of newly taken samples with respect to the historical dataset and comparing it with the control limit.

The rest of the manuscript is structured as follows. Section 2 is devoted to some preliminary notions and results, in particular about the zonoid depth. In Sect. 3 we introduce the (zonoid) interval depth together with its level sets and main properties. Section 4 is devoted to a projection-based extension to the multivariate setting which is presented together with an application in statistical quality control in the form of a control chart, whose performance comparison is analyzed. In Sect. 5 we consider interval-valued data and present a case study for which a control chart is also built. Some concluding remarks are discussed in Sect. 6.

## 2 Preliminaries

Denote by $\mathcal{F}$ the set of cumulative distribution functions (cdfs) of all $p$-dimensional random vectors. For any $F \in \mathcal{F}$ and $d \in (0, 1]$, define the *$d$-trimming of $F$* as

$$F^{(d)} = \{G \in \mathcal{F} : G(\boldsymbol{y}) - G(\boldsymbol{x}) \leq d^{-1}(F(\boldsymbol{y}) - F(\boldsymbol{x})) \text{ for any } \boldsymbol{x} \leq \boldsymbol{y} \in \mathbb{R}^p\},$$

where the '$\leq$' relation in $\mathbb{R}^p$ is understood componentwisely. If $\mathcal{F}$ is restricted to the class of distributions with density, then $F^{(d)}$ is formed by all cdfs whose density functions are upper bounded by $d^{-1}f$, while it generally consists of all distributions whose Radon-Nikodym derivative with respect to $F$ is upper bounded by $d^{-1}$, which clearly becomes larger as $d$ gets smaller.

If $T$ stands for any statistical functional of a probability distribution, that is, $T : \mathcal{F} \mapsto \mathbb{R}^q$, with a natural number $q$ possibly different from $p$, Cascos and López-Díaz (2012) define the *parameter depth region* of level $d \in (0, 1]$ induced by $T$ as

$$\mathrm{D}_T^d(F) = \{T(G) : G \in F^{(d)}\}. \tag{1}$$

Reciprocally, the *parameter depth* of an element $\boldsymbol{\theta} \in \mathbb{R}^q$ is the greatest $d$ such that $\boldsymbol{\theta}$ lies in the parameter depth region of level $d$,

$$\mathrm{D}_T(\boldsymbol{\theta}; F) = \sup\{d \in (0, 1] : \boldsymbol{\theta} \in \mathrm{D}_T^d(F)\}. \tag{2}$$

In plain words, $\boldsymbol{\theta}$ is a candidate to be a parameter of $F$ and the depth $\mathrm{D}_T(\boldsymbol{\theta}; F)$ measures the suitability of such a choice. If the fit is perfect, i.e., $\boldsymbol{\theta} = T(F)$, the parameter depth is 1.

When we consider distributions with finite first moment, $\int_{\mathbb{R}^p} \|\boldsymbol{x}\| \, \mathrm{d}F(\boldsymbol{x}) < \infty$, and $T(F) = \mu(F) = \int_{\mathbb{R}^p} \boldsymbol{x} \, \mathrm{d}F(\boldsymbol{x})$ stands for the mean, the zonoid depth regions and the zonoid depth (denoted ZD) proposed by Koshevoy and Mosler (1997) and thoroughly studied by Mosler (2002) are obtained. Alternatively, the zonoid region of level $d \in (0, 1]$ of a (possibly multivariate) distribution $F$ is the compact and convex set given as

$$\mathrm{ZD}^d(F) = \left\{ \int_{\mathbb{R}^p} \boldsymbol{x} g(\boldsymbol{x}) \mathrm{d}F(\boldsymbol{x}) : g : \mathbb{R}^p \to \left[0, \tfrac{1}{d}\right] \text{measurable}, \int_{\mathbb{R}^p} g(\boldsymbol{x}) \mathrm{d}F(\boldsymbol{x}) = 1 \right\}$$

while the zonoid depth is defined as $\mathrm{ZD}(\boldsymbol{x}; F) = \sup\{d \in (0, 1] : \boldsymbol{x} \in \mathrm{ZD}^d(F)\}$, which meets the definition of the parameter depth in (2). The zonoid depth satisfies the standard properties of a depth function (see Zuo and Serfling 2000): it is affine invariant (i.e., independent of the coordinate system used), attains its maximal value at the mean of a distribution (which is considered as the centre under its perspective), decreases on rays from the deepest point, and vanishes at infinity. Besides, for absolutely continuous distributions, the empirical zonoid depth is a uniformly consistent estimator of the population zonoid depth, see (Cascos and López-Díaz 2016).

If we consider a univariate distribution $F$ (that is, $p = 1$) with finite first moment and denote its associated quantile function by $F^{-1}$, the *zonoid region* of level $d$ is the closed interval

$$\mathrm{ZD}^d(F) = \left[ \frac{1}{d} \int_0^d F^{-1}(t)\, \mathrm{d}t\, ,\ \frac{1}{d} \int_{1-d}^1 F^{-1}(t)\, \mathrm{d}t \right].$$

When $d = 1/2$, the left and right endpoints, respectively denoted by $\underline{\mu}(F)$ and $\overline{\mu}(F)$, represent the gravity centres of the lower and upper halves of the distribution $F$, so $\mathrm{ZD}^{1/2}(F) = \left[\underline{\mu}(F), \overline{\mu}(F)\right]$. Nevertheless, throughout the manuscript, we will use the interval notation rather than the zonoid region notation for this set to emphasize that it is univariate.

For a multivariate distribution $F$ with finite first moment, the central region $\mathrm{ZD}^{1/2}(F)$ is a compact and convex set that is centrally symmetric about the point $\mu(F)$. The zonoid central regions of any level $d$ are affine equivariant and partially capture the dependency structure of the components of $F$, see (Mosler 2013).

In order to work with a parameter depth notion whose argument is set-valued, the usual set arithmetic will be considered. In particular, if $K$ is a compact and convex subset of $\mathbb{R}^p$, $A$ a nonsingular $p \times p$ matrix, and $\boldsymbol{b} \in \mathbb{R}^p$, the affine transformation given by $AK + \boldsymbol{b} = \{A\boldsymbol{x} + \boldsymbol{b} : \boldsymbol{x} \in K\}$ is the compact and convex subset obtained after multiplying each element of $K$ times $A$ and adding $\boldsymbol{b}$. Therefore, if $\boldsymbol{X}$ is a $p$-variate random vector with finite first moment, the affine equivariance of the zonoid central regions can be expressed as $\mathrm{ZD}^d(F_{A\boldsymbol{X}+\boldsymbol{b}}) = A\,\mathrm{ZD}^d(F_{\boldsymbol{X}}) + \boldsymbol{b}$. If $p = 1$, the matrix is replaced by any scalar $a \neq 0$, $b \in \mathbb{R}$, and $K = [x_l, x_u]$ is a nonempty compact interval, while $a[x_l, x_u] + b$ coincides with the interval whose endpoints are obtained after multiplying the endpoints of $[x_l, x_u]$ times $a$ and adding $b$ (if $a < 0$, the endpoints are reversed).

Any nonempty compact and convex subset of $\mathbb{R}^p$ is characterized by its support function evaluated on $\mathbb{S}^{p-1}$ (the unit sphere in $\mathbb{R}^p$), see (Schneider 1993). The *support function* of any nonempty compact and convex subset of $\mathbb{R}^p$, $K$, evaluated on $\boldsymbol{u} \in \mathbb{R}^p$ is a homogeneous and subadditive function given by $h_K(\boldsymbol{u}) = \sup\{\langle \boldsymbol{x}, \boldsymbol{u} \rangle : \boldsymbol{x} \in K\}$, where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^p$. The compact and convex set $K$ can be recovered from its support function as $K = \bigcap_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \{\boldsymbol{x} \in \mathbb{R}^p : -h_K(-\boldsymbol{u}) \leq \langle \boldsymbol{x}, \boldsymbol{u} \rangle \leq h_K(\boldsymbol{u})\}$. Further, two simple properties will be relevant later on. The support

function of the affine transformation of $K$ given by a nonsingular $p \times p$ matrix $A$, and $\boldsymbol{b} \in \mathbb{R}^p$ is $h_{AK+\boldsymbol{b}}(\boldsymbol{u}) = h_K(A^\top \boldsymbol{u}) + \langle \boldsymbol{b}, \boldsymbol{u} \rangle$, where $A^\top$ is the transpose of $A$; and given two compact convex subsets $K_1$, $K_2$ of $\mathbb{R}^p$, the inclusion relation $K_1 \subseteq K_2$ holds if and only if $h_{K_1}(\boldsymbol{u}) \leq h_{K_2}(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathbb{R}^p$.

Concerning the metrics involved in, for instance, the convergence for compact convex sets, the Hausdorff distance will be used along this manuscript. The *Hausdorff distance* between two compact convex sets $K_1, K_2 \subset \mathbb{R}^p$ is

$$d_H(K_1, K_2) = \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} |h_{K_1}(\boldsymbol{u}) - h_{K_2}(\boldsymbol{u})| \, .$$

In the univariate case, if $p = 1$, $d_H([x_l, x_u], [x'_l, x'_u]) = \max\{|x_u - x'_u|, |x_l - x'_l|\}$ and the convergence is satisfied whenever both sequences of endpoints converge to the corresponding endpoint of the limit interval.

Taking advantage of the scalar product and the unit sphere in $\mathbb{R}^p$ that have been just introduced, the *weak projection property* of the zonoid depth is presented as $\mathrm{ZD}(\boldsymbol{x}; F_{\boldsymbol{X}}) = \inf_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathrm{ZD}(\langle \boldsymbol{x}, \boldsymbol{u} \rangle; F_{\langle \boldsymbol{X}, \boldsymbol{u} \rangle})$, see (Dyckerhoff 2004). Later on, a similar projection argument will be used to extend the (univariate) interval depth to the (multivariate) region depth. The zonoid depth also satisfies the *strong projection property*, which is better characterized in terms of the central regions as

$$\mathrm{ZD}^d(F_{\langle \boldsymbol{X}, \boldsymbol{u} \rangle}) = \left[ -h_{\mathrm{ZD}^d(F_{\boldsymbol{X}})}(-\boldsymbol{u}) \, , \; h_{\mathrm{ZD}^d(F_{\boldsymbol{X}})}(\boldsymbol{u}) \right] . \tag{3}$$

Apart from the zonoid depth, the other classical data depth notion that appears in this manuscript is the halfspace one. The *halfspace depth* of $\boldsymbol{x}$ with respect to $F_{\boldsymbol{X}}$ is given by $\mathrm{HD}(\boldsymbol{x}; F_{\boldsymbol{X}}) = \inf_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \Pr(\langle \boldsymbol{X}, \boldsymbol{u} \rangle \geq \langle \boldsymbol{x}, \boldsymbol{u} \rangle)$, see (Tukey 1975), which corresponds to the infimum of the probabilities of all closed halfspaces containing $\boldsymbol{x}$ in their boundaries.

## 3 Zonoid interval depth

For any $m_1 \leq m_2 \in \mathbb{R}$, the (zonoid) interval depth assesses the fit of the interval $[m_1, m_2]$ to a univariate distribution as its zonoid region of level $1/2$. This is done by identifying $[m_1, m_2]$ with the pair given by its endpoints $(m_1, m_2)$ and applying (2) to the bivariate functional $T = (\underline{\mu}, \overline{\mu})$ that determines $\mathrm{ZD}^{1/2}$.

**Definition 1** Take $m_1 \leq m_2 \in \mathbb{R}$, the *interval depth* of $[m_1, m_2]$ with respect to a univariate distribution $F$ is defined as

$$\mathrm{ID}([m_1, m_2]; F) = \sup\{d \in (0, 1] : \underline{\mu}(G) = m_1 \text{ and } \overline{\mu}(G) = m_2 \text{ for } G \in F^{(d)}\} \, . \tag{4}$$

Equivalenty, the interval depth can be expressed in terms of the interval-valued functional $\mathrm{ZD}^{1/2}$ as $\mathrm{ID}([m_1, m_2]; F) = \sup\{d \in (0, 1] : [m_1, m_2] = \mathrm{ZD}^{1/2}(G) \text{ for } G \in F^{(d)}\}$.

The following result provides another alternative expression for the interval depth, which will be of much practical use and explicitly shows its relationship with the zonoid depth of the endpoints of the considered interval.

**Lemma 1** *The interval depth of* $[m_1, m_2]$ *with respect to a distribution* $F$ *with finite first moment satisfies*

$$\mathrm{ID}([m_1, m_2]; F) = 2 \sup_{0<\lambda<1} \min\{\lambda \mathrm{ZD}(m_1; F_{\lambda-}), (1-\lambda)\mathrm{ZD}(m_2; F_{\lambda+})\}, \quad (5)$$

*where* $F_{\lambda-}$ *is the cdf defined as* $F_{\lambda-}(x) = F(x)/\lambda$ *if* $F(x) \leq \lambda$ *and* 1 *otherwise, while* $F_{\lambda+}(x) = (F(x) - \lambda)/(1-\lambda)$ *if* $F(x) \geq \lambda$ *and* 0 *otherwise.*

***Proof*** Consider $m_1 \leq m_2$ with strictly positive zonoid depths with respect to $F$ since otherwise both expressions in (5) are equal to 0. For the same reason, consider $m_1 \neq m_2$ unless $F$ has an atom at $m_1 = m_2$.

In order to show the '$\geq$' inequality, consider any $0 < \lambda < 1$ and, for values $d_1 = \mathrm{ZD}(m_1; F_{\lambda-})$ and $d_2 = \mathrm{ZD}(m_2; F_{\lambda+})$, take distributions $G_1 \in F_{\lambda-}^{(d_1)}$ with $m_1 = \mu(G_1)$ and $G_2 \in F_{\lambda+}^{(d_2)}$ with $m_2 = \mu(G_2)$. Their mixture $G = 0.5(G_1 + G_2)$ satisfies $(\underline{\mu}, \overline{\mu})(G) = (m_1, m_2)$ and $G \in F^{(d)}$, where $d = 2\min\{\lambda d_1, (1-\lambda)d_2\}$, and thus it holds that $\mathrm{ID}([m_1, m_2]; F) \geq d$.

With regard to the '$\leq$' inequality, take any $0 < d \leq \mathrm{ID}([m_1, m_2]; F)$ with some $G \in F^{(d)}$ such that $(\underline{\mu}, \overline{\mu})(G) = (m_1, m_2)$. Denote the smallest median of $G$ by $x_0$, that is, $G(x_0) \geq 1/2$ and $G(x) < 1/2$ for any $x < x_0$ and let $\lambda_0$ be such that $d/2 + F(x_0-) - dG(x_0-) \leq \lambda_0 \leq d/2 + F(x_0) - dG(x_0)$, where $F(x_0-)$ and $G(x_0-)$ are the left limits of $F$ and $G$ at $x_0$. Finally, $G_{1/2-} \in F_{\lambda_0-}^{(d/(2\lambda_0))}$ with $\mu(G_{1/2-}) = m_1$, so $\mathrm{ZD}(m_1; F_{\lambda_0-}) \geq d/(2\lambda_0)$, and $G_{1/2+} \in F_{\lambda_0+}^{(d/(2(1-\lambda_0)))}$ with $\mu(G_{1/2+}) = m_2$, so $\mathrm{ZD}(m_2; F_{\lambda_0+}) \geq d/(2(1-\lambda_0))$, and the inequality must hold for all such values $d$ and thus for their supremum. □

If $F$ is continuous, we can replace $\lambda$ by $F(x)$ in (5). At the same time, we restrict to values of $x$ inside the interval $[m_1, m_2]$ since some of the considered zonoid depths vanish out of it. It finally renders

$$\mathrm{ID}([m_1, m_2]; F) = 2 \sup_{m_1 \leq x \leq m_2} \min\{F(x)\mathrm{ZD}(m_1; F_{F(x)-}), (1-F(x))\mathrm{ZD}(m_2; F_{F(x)+})\}.$$

***Example 1*** In the special case that $H$ is the cdf of the uniform distribution on the unit interval,

$$\mathrm{ID}([m_1, m_2]; H) = \begin{cases} 4\min\{m_1, 1-m_2, (m_2-m_1)/2\} & \text{if } 0 \leq m_1 \leq m_2 \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

which can be easily deduced from the zonoid depth with respect to $H$, and this results to be $\mathrm{ZD}(x; H) = 2\min\{x, 1-x\}$ if $0 \leq x \leq 1$ and $\mathrm{ZD}(x; H) = 0$ otherwise.

The interval depth with respect to the empirical probability associated with a univariate sample with ordered version $x_{(1)} \leq \ldots \leq x_{(n)}$, denoted by $\mathrm{ID}_n$, is computed by replacing the cdf $F$ in (5) by an empirical distribution in order to obtain

$$\mathrm{ID}_n([m_1, m_2]) = \frac{2}{n} \sup_{0<\lambda<1} \min\{(r+\lambda)\mathrm{ZD}_{1:r,\lambda}(m_1), (n-r-\lambda)\mathrm{ZD}_{1-\lambda,r:n}(m_2)\}, \quad (6)$$

where $r = \lfloor \lambda n \rfloor + 1$, with $\lfloor \lambda n \rfloor$ the greatest integer less than or equal to $\lambda n$, $\mathrm{ZD}_{1:r,\lambda}(x)$ is the zonoid depth with respect to the discrete distribution on the points $x_{(1)} \leq \ldots \leq x_{(r)}$ such that the probability of $x_{(r)}$ is $(\lambda n - \lfloor \lambda n \rfloor)/(\lambda n)$ and the remaining $r-1$ points have the same probability $1/(\lambda n)$, and $\mathrm{ZD}_{1-\lambda,r:n}(x)$ is the zonoid depth with respect to the discrete distribution on $x_{(r)} \leq \ldots \leq x_{(n)}$ such that the probability of $x_{(r)}$ is $(\lfloor \lambda n \rfloor + 1 - \lambda n)/((1-\lambda)n)$ and the remaining $n-r$ points have the same probability $1/((1-\lambda)n)$.

A natural use of the interval depth is to evaluate how well a sample (or distribution) fits some reference distribution (which can be given in terms of a dataset) with respect to the features captured by central regions. In the rest of the paper, the assessment of the fit of that specific sample (or distribution) with respect to another reference distribution will always refer to those features. In such a case, the first sample is summarized in terms of its zonoid interval of level $1/2$ and the interval depth of such a zonoid region with respect to the reference distribution is computed. The R source code to compute the interval depth as in (6) is available on the GitHub repository https://github.com/icascos/intervaldepth. Two functions can be found there, `m1m2depth` for the computation of the interval depth of an interval with respect to some reference sample, and `sm1m2depth` for the computation of the interval depth of the zonoid interval of level $1/2$ of a sample with respect to another reference sample. Algorithms for the approximate computation of the multivariate extension introduced in Sect. 4, named region depth, are also available on the repository.

***Example 2*** We have drawn a sample of 100 observations from a standard normal distribution and 4 more samples of size 10 from further normal distributions. The zonoid interval of level $d = 1/2$ of each of the samples was computed, and its depth with respect to the standard normal sample of size 100 is presented in Fig. 1.

As mentioned in the introduction, depth regions provide interesting information about the distribution. The following lemma specifies how the level sets of the interval depth are built.

**Lemma 2** *The level set at $d \in (0, 1]$ of the interval depth of a distribution $F$ with finite first moment renders*

$$\mathrm{ID}^d(F) = \left\{ [m_1, m_2] : \begin{array}{c} \int_0^{\frac{d}{2}} F^{-1}(t)\,\mathrm{d}t \leq \frac{d}{2} m_1 \leq \int_{s-\frac{d}{2}}^{s} F^{-1}(t)\,\mathrm{d}t \\ \int_s^{s+\frac{d}{2}} F^{-1}(t)\,\mathrm{d}t \leq \frac{d}{2} m_2 \leq \int_{1-\frac{d}{2}}^{1} F^{-1}(t)\,\mathrm{d}t \end{array}, \frac{d}{2} \leq s \leq 1 - \frac{d}{2} \right\}.$$

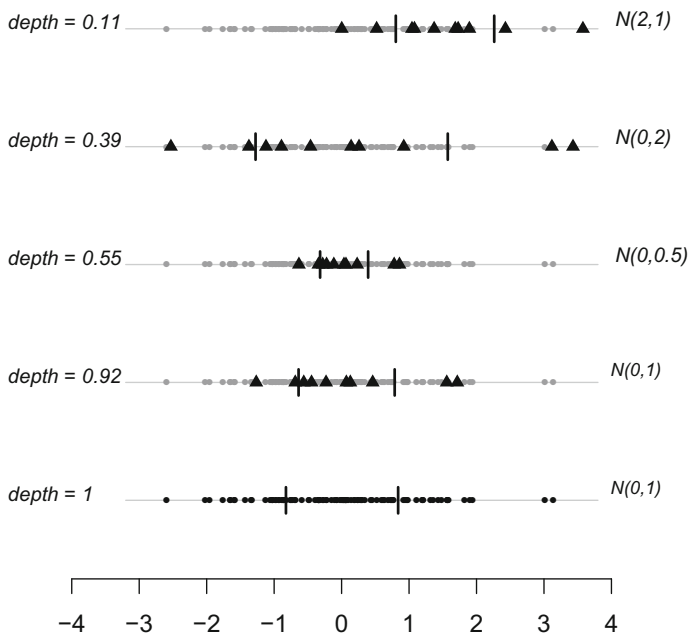


**Fig. 1** Interval depth of the zonoid region of level 1/2 of 4 samples of size 10 of normal distributions (in black triangles) with respect to a sample of size 100 of a standard normal (in grey circles). This last sample, together with its zonoid interval of level 1/2, is represented at the bottom row (in black circles)

***Proof*** These inequalities derive from considering the relationship between the quantile functions of $G \in F^{(d)}$ and $F$, and applying convenient changes of variables. Specifically, the right-to-left inclusion follows from the fact that, for every $d/2 \leq s \leq 1-d/2$, the integrals correspond to the endpoints of the zonoid regions of distributions with median $x_0$ which are formed by restricting $F$ to some part of its support whose probability is $d$, and thus they are bounded above by $d^{-1}F$. In order to show the left-to-right inclusion, for any $G \in F^{(d)}$, whose smallest median is $x_0$, we must take $s = F(x_0)$ so that $\frac{d}{2}\underline{\mu}(G)$ and $\frac{d}{2}\overline{\mu}(G)$ are bounded by the integrals presented in the expression on the right. □

***Example 3*** Figure 2 left shows the contours of some level sets of the interval depth of a standard normal distribution. The centremost point corresponds to $(\mathbb{E}[X|X < 0], \mathbb{E}[X|X > 0]) = \left(-\sqrt{2/\pi}, \sqrt{2/\pi}\right)$, where $\mathbb{E}$ denotes the mathematical expectation and $X$ is a standard normal random variable. On the right, the contours of the level sets of the interval depth of a mixture of two normal distributions are shown. Observe that due to the affine equivariance of the zonoid central regions, the region of level 1/2 of a general univariate normal distribution is the interval $\mathrm{ZD}^{1/2}(F_{\sigma X+\mu}) = \left[\mu - \sigma\sqrt{2/\pi}, \mu + \sigma\sqrt{2/\pi}\right]$, whence for a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma$, it is the convex set $\{\boldsymbol{x} \in \mathbb{R}^p : (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq 2/\pi\}$, see (Koshevoy and Mosler 1997, Section 6).

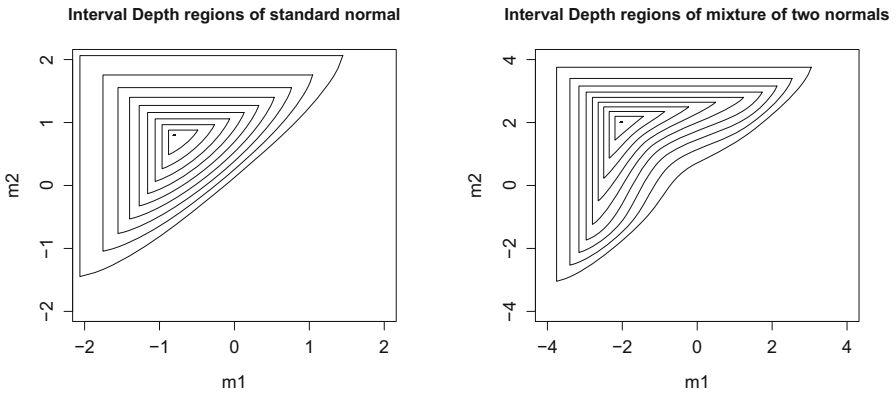


**Fig. 2** Contours of level sets ($d = 1/10, 2/10, \ldots, 10/10$) of the interval depth of N(0, 1) (left) and the mixture $0.5\mathrm{N}(-2, 1) + 0.5\mathrm{N}(2, 1)$ (right)

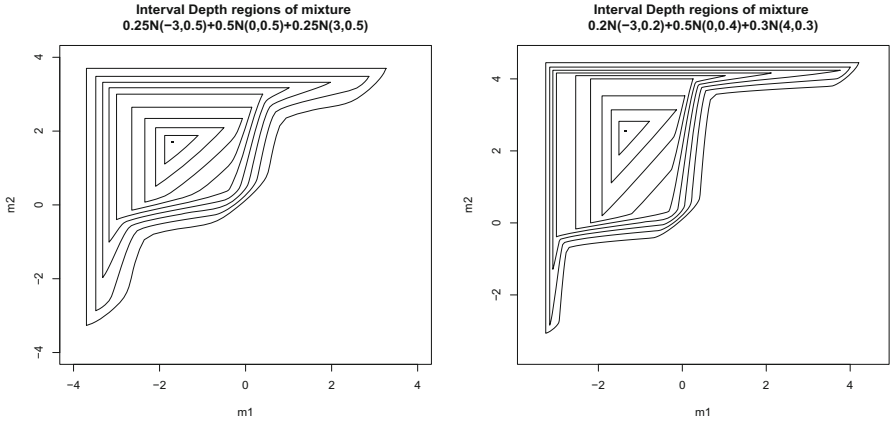


**Fig. 3** Contours of level sets ($d = 1/10, 2/10, \ldots, 10/10$) of the interval depth of the mixture $0.25\mathrm{N}(-3, 0.5) + 0.5\mathrm{N}(0, 0.5) + 0.25\mathrm{N}(3, 0.5)$ (left) and the mixture $0.2\mathrm{N}(-3, 0.2) + 0.5\mathrm{N}(0, 0.4) + 0.3\mathrm{N}(4, 0.3)$ (right)

Figure 3 shows the contours of some level sets of the interval depth of two different mixtures of three normal distributions whose weights and normal parameters are given in the caption.

In the empirical setting, Lemma 2 for $k \in \{1, 2, \ldots, \lfloor n/2 \rfloor\}$ yields

$$\mathrm{ID}_n^{2k/n} = \left\{ [m_1, m_2] : \begin{array}{c} \sum_{i=1}^{k} x_{(i)} \le k m_1 \le \sum_{i=j-k+1}^{j} x_{(i)} \\ \sum_{i=j+1}^{j+k} x_{(i)} \le k m_2 \le \sum_{i=n-k+1}^{n} x_{(i)} \end{array} \text{ for } k \le j \le n-k \right\}.$$

Some basic properties of the interval depth function are presented below, the penultimate one showing a lower bound with conditions on $[m_1, m_2]$ for a nonzero depth.

**Proposition 3** *For any univariate distribution $F$ associated with a random variable $X$ with finite first moment, $F = F_X$, and any $m_1 \leq m_2 \in \mathbb{R}$, the following properties hold:*

*ID0.* Upper semicontinuity, $\mathrm{ID}([m_1, m_2]; F) \geq \limsup_n \mathrm{ID}([x_n, y_n]; F)$ *if* $\lim_n x_n = m_1$ *and* $\lim_n y_n = m_2$.

*ID1.* Affine invariance, $\mathrm{ID}(a[m_1, m_2] + b; F_{aX+b}) = \mathrm{ID}([m_1, m_2]; F_X)$ *for any* $a \neq 0$, $b \in \mathbb{R}$.

*ID2.* Centre, $\mathrm{ID}\left(\left[\underline{\mu}(F), \overline{\mu}(F)\right]; F\right) = 1$.

*ID3.* Decreases as the centremost interval widens or shrinks, *if* $[m_1, m_2] \supseteq [m_1', m_2'] \supseteq \left[\underline{\mu}(F), \overline{\mu}(F)\right]$ *or* $[m_1, m_2] \subseteq [m_1', m_2'] \subseteq \left[\underline{\mu}(F), \overline{\mu}(F)\right]$, *then* $\mathrm{ID}([m_1, m_2]; F) \leq \mathrm{ID}([m_1', m_2']; F)$.

*ID3'.* Decreases as the centremost interval is shifted in location, *if* $b \in \mathbb{R}$ *and* $0 \leq \delta \leq 1$, $\mathrm{ID}\left(\left[\underline{\mu}(F), \overline{\mu}(F)\right] + b; F\right) \leq \mathrm{ID}\left(\left[\underline{\mu}(F), \overline{\mu}(F)\right] + \delta b; F\right)$.

*ID4.* Vanishes when the interval grows arbitrarily wide or its midpoint tends to $\infty$ or $-\infty$ *since* $\mathrm{ID}([m_1, m_2]; F) \leq 2\min\{\mathrm{ZD}(m_1; F), \mathrm{ZD}(m_2; F)\}$.

*ID4'.* Matches the probability of a specific value if the interval collapses to a singleton, $\mathrm{ID}([m, m]; F) = \Pr(X = m)$.

*ID5.* Lower bound, *if $F$ is a continuous distribution,*

$$\mathrm{ID}([m_1, m_2]; F) \geq 2\min\{F(m_1), (F(m_2) - F(m_1))/2, 1 - F(m_2)\}.$$

*ID5'.* Upper bound, $\mathrm{ID}([m_1, m_2]; F) \leq \mathrm{ZD}((m_1 + m_2)/2; F)$.

***Proof*** *ID0* follows from the level sets being closed, which holds due to the continuity of $\underline{\mu}$ and $\overline{\mu}$ for the weak convergence on the $d$-trimming of a distribution with finite first moment (see Cascos and López-Díaz 2012, Corollary 4.2).

*ID1* follows from the affine equivariance of $\underline{\mu}$ and $\overline{\mu}$ (see Cascos and López-Díaz 2012, Proposition 4.6).

*ID2* is immediate from Definition 1.

*ID3* and *ID3'* can be argued from the level sets presented in Lemma 2.

*ID4* follows from Lemma 1 and the zonoid depth definition.

*ID4'* holds true since $\underline{\mu}$ and $\overline{\mu}$ only match at degenerate distributions.

*ID5'* follows from $\mu(F) = (\underline{\mu}(F) + \overline{\mu}(F))/2$.

In order to show *ID5*, consider $m_1 < m_2$ and let $m_1 < x < m_2$ be such that $F(x) = (F(m_1) + F(m_2))/2$. After Lemma 1, it holds that $\mathrm{ID}([m_1, m_2]; F) \geq 2\min\{F(x)\mathrm{ZD}(m_1; F_{F(x)-}), (1 - F(x))\mathrm{ZD}(m_2; F_{F(x)+})\}$.

Observe now that for any $y \in \mathbb{R}$, the zonoid depth of $y$ with respect to $F$ is never less than the minimum of $F(y)$ and $1 - F(y)$. Then $\mathrm{ZD}(m_1; F_{F(x)-}) \geq \min\{F_{F(x)-}(m_1), 1 - F_{F(x)-}(m_1)\} = \min\{F(m_1)/F(x), 1 - F(m_1)/F(x)\}$, while $\mathrm{ZD}(m_2; F_{F(x)+}) \geq \min\{(F(m_2) - F(x))/(1 - F(x)), (1 - F(m_2))/(1 - F(x))\}$. Finally we conclude that $\mathrm{ID}([m_1, m_2]; F) \geq 2\min\{F(m_1), (F(m_2) - F(m_1))/2, 1 - F(m_2)\}$. □

***Remark 1*** The *zero-depth problem* arises when elements in the parameter space with a depth different from 0 are scarce. Despite it is a common issue in some settings, the lower bound presented in *ID5* evidences that we do not face it here.

With regard to the properties presented in Proposition 3, *ID0* is a basic continuity requirement. Properties *ID1*, *ID2*, *ID3*, *ID3'*, *ID4*, and *ID4'* mimic the standard properties of a data depth function (affine invariance, maximality at the centre, decreasing from deepest point, and vanishing as the norm of a point tends to infinity) first investigated for the simplicial depth by Liu (1990) and later popularized by Zuo and Serfling (2000). Finally, the lower bound presented in *ID5* shows that the interval depth is not affected by the zero-depth problem, while the upper bound given in *ID5'* relates the interval depth with the zonoid depth.

## 4 Multivariate extension

The extension of the zonoid interval depth to the multivariate setting will be tackled now. The region depth assesses the fit of a $p$-dimensional compact convex set to a $p$-dimensional distribution by means of a projection argument applied to the interval depth.

**Definition 2** Consider a $p$-dimensional random vector $\boldsymbol{X}$ with finite first moment and a compact convex set $K \subset \mathbb{R}^p$. The *region depth* of $K$ is the infimum of the interval depths of all the univariate projections of $K$ with respect to the cdf of the projected $\boldsymbol{X}$, that is,

$$\mathrm{RD}(K; F_{\boldsymbol{X}}) = \inf_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathrm{ID}\left([-h_K(-\boldsymbol{u}), h_K(\boldsymbol{u})]; F_{\langle \boldsymbol{X}, \boldsymbol{u} \rangle}\right) .$$

The natural way to assess the fit of a sample (or distribution) with respect to another reference distribution is by means of the region depth of the zonoid region of level $1/2$ of the sample with respect to the reference distribution.

***Remark 2*** The *straightforward* multivariate extension of the interval depth would be to merge Eqs. (1) and (2) in order to define the depth of a $p$-dimensional convex body $K$ with respect to a $p$-dimensional distribution $F$ as the supremum of the levels $d$ such that $K$ matches the zonoid region of level $1/2$ of some distribution in $F^{(d)}$. Unfortunately such a construction suffers from the zero-depth problem. Zonoid regions of level $1/2$ are centrally symmetric about the mean of the reference distribution, so only centrally symmetric sets would attain a depth strictly greater than zero. Further, the zonoid regions of an empirical distribution are always polytopes, so any set with a smooth boundary would always have depth zero if the reference distribution is an empirical one. For these reasons, we have introduced the region depth as presented in Definition 2. It is lower bounded by the construction described in this remark, and also by the expression presented in Proposition 4, *RD5*, given in terms of the probability content of the set whose depth is evaluated and the halfspace depth of its boundary points.

The result below mimics the properties of the interval depth presented in Proposition 3 in the multivariate setting. It collects those properties that we expect for a functional that assesses the fit of a compact convex set to a multivariate distribution in

a depth function fashion and describes further upper and lower bounds for our specific proposal.

**Proposition 4** *For any multivariate distribution $F$ associated with a $p$-dimensional random vector with finite first moment, $F = F_{\boldsymbol{X}}$, and any compact convex sets $K_1, \ldots, K_n, K \subset \mathbb{R}^p$, the following properties hold:*

*RD0.* Upper semicontinuity, $\mathrm{RD}(K; F) \geq \limsup_n \mathrm{RD}(K_n; F)$ *if* $\lim_n K_n = K$, *where the convergence of compact convex sets is in the Hausdorff sense.*

*RD1.* Affine invariance, $\mathrm{RD}(AK + \boldsymbol{b}; F_{A\boldsymbol{X}+\boldsymbol{b}}) = \mathrm{RD}(K; F_{\boldsymbol{X}})$ *for any nonsingular matrix* $A \in \mathbb{R}^{p\times p}$ *and* $\boldsymbol{b} \in \mathbb{R}^p$.

*RD2.* Centre, $\mathrm{RD}(\mathrm{ZD}^{1/2}(F); F) = 1$.

*RD3.* Decreases as the centremost set widens or shrinks, *if* $K \supseteq K' \supseteq \mathrm{ZD}^{1/2}(F)$ *or* $K \subseteq K' \subseteq \mathrm{ZD}^{1/2}(F)$, *then* $\mathrm{RD}(K; F) \leq \mathrm{RD}(K'; F)$.

*RD3'.* Decreases as the centremost set is shifted in location, *for any* $\boldsymbol{b} \in \mathbb{R}^p$ *and* $0 \leq \delta \leq 1$, $\mathrm{RD}\left(\mathrm{ZD}^{1/2}(F) + \boldsymbol{b}; F\right) \leq \mathrm{RD}\left(\mathrm{ZD}^{1/2}(F) + \delta\boldsymbol{b}; F\right)$.

*RD4.* Vanishes when the set grows arbitrarily wide or the norm of any of its points tends to $\infty$ *since* $\mathrm{RD}(K; F) \leq 2 \inf_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \{\mathrm{ZD}(h_K(\boldsymbol{u}); F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle})\}$.

*RD4'.* Matches the probability of a specific value when the set collapses to a singleton, $\mathrm{RD}(\{\boldsymbol{m}\}; F) = \Pr(\boldsymbol{X} = \boldsymbol{m})$.

*RD5.* Lower bound, *if* $\boldsymbol{X}$ *is a continuous random vector,*

$$\mathrm{RD}(K; F_{\boldsymbol{X}}) \geq \min\{\Pr(\boldsymbol{X} \in K), 2 \inf_{\boldsymbol{x}\in\partial K} \mathrm{HD}(\boldsymbol{x}; F_{\boldsymbol{X}})\},$$

*where $\partial K$ is the boundary of $K$ and we recall that* $\mathrm{HD}(\boldsymbol{x}; F_{\boldsymbol{X}})$ *is the halfspace depth of* $\boldsymbol{x}$ *with respect to* $F_{\boldsymbol{X}}$.

*RD5'.* Upper bound, *if $K$ is centrally symmetric about* $\boldsymbol{m} \in \mathbb{R}^p$, *then* $\mathrm{RD}(K; F) \leq \mathrm{ZD}(\boldsymbol{m}; F)$.

***Proof*** Each of the properties of the region depth is a consequence of the corresponding property of the interval depth in Proposition 3.

*RD0* holds since the region depth is the infimum of a collection of upper semicontinuous functions.

*RD1.* Consider a nonsingular matrix $A \in \mathbb{R}^{p\times p}$, $\boldsymbol{b} \in \mathbb{R}^p$, and $\boldsymbol{u} \in \mathbb{S}^{p-1}$. In order to obtain the identities below, we successively use the expression of the support function of the affine transformation of a compact and convex subset of $\mathbb{R}^p$, *ID1*, the homogeneity of the support function, and *ID1*

$$\begin{aligned}
&\mathrm{ID}\left([-h_{AK+\boldsymbol{b}}(-\boldsymbol{u}), h_{AK+\boldsymbol{b}}(\boldsymbol{u})]; F_{\langle A\boldsymbol{X}+\boldsymbol{b},\boldsymbol{u}\rangle}\right)\\
&\quad = \mathrm{ID}\left([-h_K(-A^\top\boldsymbol{u}), h_K(A^\top\boldsymbol{u})] + \langle\boldsymbol{b},\boldsymbol{u}\rangle; F_{\langle \boldsymbol{X},A^\top\boldsymbol{u}\rangle+\langle\boldsymbol{b},\boldsymbol{u}\rangle}\right)\\
&\quad = \mathrm{ID}\left([-h_K(-A^\top\boldsymbol{u}), h_K(A^\top\boldsymbol{u})]; F_{\langle \boldsymbol{X},A^\top\boldsymbol{u}\rangle}\right)\\
&\quad = \mathrm{ID}\left(\|A^\top\boldsymbol{u}\|[-h_K(-A^\top\boldsymbol{u}/\|A^\top\boldsymbol{u}\|), h_K(A^\top\boldsymbol{u}/\|A^\top\boldsymbol{u}\|)]; F_{\langle \boldsymbol{X},A^\top\boldsymbol{u}\rangle}\right)\\
&\quad = \mathrm{ID}\left([-h_K(-A^\top\boldsymbol{u}/\|A^\top\boldsymbol{u}\|), h_K(A^\top\boldsymbol{u}/\|A^\top\boldsymbol{u}\|)]; F_{\langle \boldsymbol{X},A^\top\boldsymbol{u}/\|A^\top\boldsymbol{u}\|\rangle}\right).
\end{aligned}$$

Finally, the nonsingularity of matrix $A$ guarantees that any element in $\mathbb{S}^{p-1}$ can be expressed as $A^\top \boldsymbol{u}/\|A^\top \boldsymbol{u}\|$ for some $\boldsymbol{u} \in \mathbb{S}^{p-1}$, so taking the infimum of the expresion above on such $\boldsymbol{u} \in \mathbb{S}^{p-1}$, the identity would start with $\mathrm{RD}(AK + \boldsymbol{b}; F_{A\boldsymbol{X}+\boldsymbol{b}})$ and end with $\mathrm{RD}(K; F_{\boldsymbol{X}})$.

*RD2.* By the strong projection property of the zonoid depth, see (3), for any $\boldsymbol{u} \in \mathbb{S}^{p-1}$, the projection of the zonoid region of level $1/2$ of $\boldsymbol{X}$ in the direction of $\boldsymbol{u}$ is the zonoid region of level $1/2$ of $\langle \boldsymbol{X}, \boldsymbol{u}\rangle$, whose interval depth with respect to $F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle}$ is 1 by *ID2*.

*RD3.* Take $K, K'$ compact and convex subsets of $\mathbb{R}^p$ such that $K \supseteq K' \supseteq \mathrm{ZD}^{1/2}(F)$. This inclusion guarantees that for any $\boldsymbol{u} \in \mathbb{S}^{p-1}$,

$$[-h_K(-\boldsymbol{u}), h_K(\boldsymbol{u})] \supseteq [-h_{K'}(-\boldsymbol{u}), h_{K'}(\boldsymbol{u})] \supseteq [-h_{\mathrm{ZD}^{1/2}(F)}(-\boldsymbol{u}), h_{\mathrm{ZD}^{1/2}(F)}(\boldsymbol{u})].$$

Further, by the strong projection property of the zonoid depth, see (3), $\underline{\mu}(F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle}) = -h_{\mathrm{ZD}^{1/2}(F)}(-\boldsymbol{u})$ and $\overline{\mu}(F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle}) = h_{\mathrm{ZD}^{1/2}(F)}(\boldsymbol{u})$. Finally, we can use *ID3* to conclude $\mathrm{ID}([-h_K(-\boldsymbol{u}), h_K(\boldsymbol{u})]; F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle}) \leq \mathrm{ID}([-h_{K'}(-\boldsymbol{u}), h_{K'}(\boldsymbol{u})]; F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle})$, and the inequality must be also satisfied for the infimum over all $\boldsymbol{u} \in \mathbb{S}^{p-1}$ which constitutes the region depth. If all the inclusions are reversed the property is kept, also by *ID3*.

*RD3'.* Take $\boldsymbol{b} \in \mathbb{R}^p$, $0 \leq \delta \leq 1$, and $\boldsymbol{u} \in \mathbb{S}^{p-1}$. We write the support function over the zonoid region of level $1/2$ in terms of the endpoints of the zonoid interval of $\langle \boldsymbol{X}, \boldsymbol{u}\rangle$, and after *ID3'* obtain that

$$\begin{aligned}&\mathrm{ID}([\underline{\mu}(F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle}), \overline{\mu}(F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle})] + \langle \boldsymbol{b}, \boldsymbol{u}\rangle; F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle})\\ &\quad \leq \mathrm{ID}([\underline{\mu}(F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle}), \overline{\mu}(F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle})] + \delta\langle \boldsymbol{b}, \boldsymbol{u}\rangle; F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle})\,.\end{aligned}$$

Since the inequlity holds for all $\boldsymbol{u} \in \mathbb{S}^{p-1}$, we confirm the desired result for the region depth.

*RD4.* Observe that in the presented upper bound, it is enough to evaluate the zonoid depth in one of the endpoints of the interval $[-h_K(-\boldsymbol{u}), h_K(\boldsymbol{u})]$. The reason is that, after the affine equivariance of the zonoid depth, for any $\boldsymbol{u} \in \mathbb{S}^{p-1}$ it holds that $\mathrm{ZD}(-h_K(-\boldsymbol{u}); F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle}) = \mathrm{ZD}(h_K(-\boldsymbol{u}); F_{\langle \boldsymbol{X},-\boldsymbol{u}\rangle})$, which is attained when $-\boldsymbol{u}$ is considered.

*RD4'.* Take any $\boldsymbol{m} \in \mathbb{R}^p$,

$$\begin{aligned}\mathrm{RD}(\{\boldsymbol{m}\}; F_{\boldsymbol{X}}) &= \inf_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \mathrm{ID}\left([\langle \boldsymbol{m}, \boldsymbol{u}\rangle, \langle \boldsymbol{m}, \boldsymbol{u}\rangle]; F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle}\right)\\ &= \inf_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \Pr\left(\langle \boldsymbol{X}, \boldsymbol{u}\rangle = \langle \boldsymbol{m}, \boldsymbol{u}\rangle\right) = \Pr\left(\boldsymbol{X} = \boldsymbol{m}\right)\,,\end{aligned}$$

where the first equality is the definition of the region depth and the second follows from *ID4'*.

*RD5.* For any $\boldsymbol{u} \in \mathbb{S}^{p-1}$, use the lower bound for $\mathrm{ID}([-h_K(-\boldsymbol{u}), h_K(\boldsymbol{u})]; F_{\langle \boldsymbol{X},\boldsymbol{u}\rangle})$ given in *ID5*. Since $\Pr\left(\langle \boldsymbol{X}, \boldsymbol{u}\rangle \leq -h_K(-\boldsymbol{u})\right)$ is the probability that the random vector $\boldsymbol{X}$ lies in a halfspace whose boundary is a supporting hyperplane of $K$ (with normal $\boldsymbol{u}$), we have $\Pr\left(\langle \boldsymbol{X}, \boldsymbol{u}\rangle \leq -h_K(-\boldsymbol{u})\right) \geq \inf_{\boldsymbol{x}\in\partial K} \mathrm{HD}(\boldsymbol{x}; F_{\boldsymbol{X}})$, while the same bound is attained for $\Pr\left(\langle \boldsymbol{X}, \boldsymbol{u}\rangle \geq h_K(\boldsymbol{u})\right)$. Further, $\Pr\left(-h_K(-\boldsymbol{u}) \leq \langle \boldsymbol{X}, \boldsymbol{u}\rangle \leq h_K(\boldsymbol{u})\right)$ is the

probability that $\boldsymbol{X}$ lies between the two supporting hyperplanes of $K$ with a fixed normal $\boldsymbol{u}$, which is lower bounded by the probability that $\boldsymbol{X}$ lies in $K$.

*RD5'*. If $K$ is centrally symmetric about $\boldsymbol{m} \in \mathbb{R}^p$, then for any $\boldsymbol{u} \in \mathbb{S}^{p-1}$, $\langle \boldsymbol{m}, \boldsymbol{u} \rangle$ is the midpoint of the interval $[-h_K(-\boldsymbol{u}), h_K(\boldsymbol{u})]$. Together with the weak projection property of the zonoid depth, this proves the result. □

### 4.1 Monitoring multivariate processes

In the following, we consider Phase I control charts for multivariate processes. Specifically, we introduce a nonparametric control chart for global monitoring of subgrouped data whose charting statistic is the region depth in the manner suggested in Cascos and López-Díaz (2018). This chart is compared with two other charts that monitor the location, the Hotelling $T^2$ (see Montgomery 2013, Section 11.3.1) and the zonoid depth chart.

The goal in Phase I applications is to detect samples with anomalous observations among a set of available (trial) ones. In the charts built for depth notions, there is a unique lower control limit and any sample whose associated depth is below it is declared as out-of-control, while conversely, the Hotelling $T^2$ chart has a unique upper control limit and any sample whose $t^2$ statistic exceeds it is declared as out-of-control. These control limits depend on the sample size, the distribution of the reference dataset, the nominal false alarm (type I error) probability (probability that the statistic of a sample that follows the reference distribution lies in the out-of-control region), and on the number of trial samples.

Consider a reference dataset formed by $k$ trial samples of a given size $n$ each. In the Hotelling $T^2$ chart, the sample mean of each trial sample is taken and the charting statistic is its distance to the grand mean (average of sample means) measured as a quadratic form that involves the pooled covariance matrix built out of the covariance matrices of all trial samples. In the charts based on a notion of depth, a statistic is taken for each trial sample (sample mean for the zonoid depth chart and zonoid region of level $1/2$ for the interval depth chart), and the charting statistic is the (zonoid or interval) depth of such statistic with respect to the pooled or reference dataset formed by merging all trial samples.

An alternative approach for statistical process control with data depth involves assessing the depth of each individual observation, converting such depths into ranks and finally averaging all the ranks of the observations in the same sample in order to combine the provided information and achieve a prompt detection of the anomaly, as suggested in Liu (1995). In this case, the monitoring is always done in terms of location, which is the feature captured by data depths.

#### 4.1.1 Location monitoring with the Hotelling $T^2$ and the zonoid depth charts

The dataset given in Ryan (2011, Table 9.2) contains $k = 20$ bivariate trial samples of size $n = 4$ each. It was originally presented to illustrate the Hotelling $T^2$ chart for subgrouped data, whose aim is to detect shifts in location at specific samples for normal processes. If we set the nominal false alarm probability to 0.025, the control

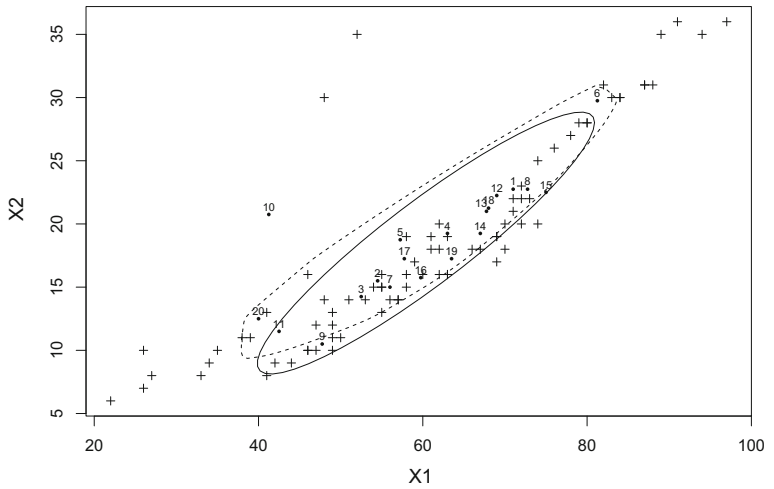


**Fig. 4** Scatterplot of the $80 = 4 \times 20$ observations in Ryan's dataset, identification number of each of the 20 samples located at their sample means, ellipse with solid contour containing all in-control sample means in the Hotelling $T^2$ chart and zonoid region with dashed contour containing all in-control sample means in terms of the zonoid depth

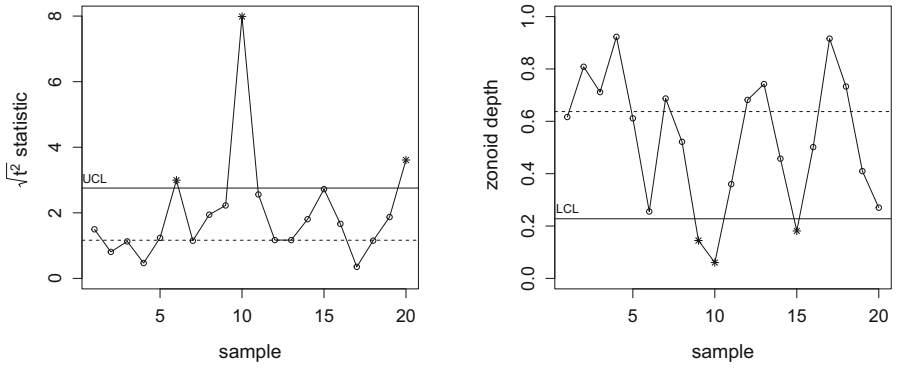


**Fig. 5** Evolution of the square root of the Hotelling $t^2$ statistic of the samples with respect to the pooled dataset (left) and evolution of the zonoid depth of each sample mean with respect to the pooled dataset (right)

ellipse plotted with a solid contour in Fig. 4 contains all sample means (located at the black bullets with the respective sample number on top for identification) except those of samples #6, #10, and #20. In Fig. 5 (left) we have represented the evolution of the square root of the $t^2$ statistic, and only the values for the three previous samples lie above the upper control limit, which is plotted as a solid line. The dashed line represents the median of the statistic in the in-control state.

The evolution of the zonoid depth of the sample means of the trial samples with respect to the pooled dataset is presented in Fig. 5 (right). The lower control limit (solid horizontal line) was obtained after Monte Carlo simulations for a bivariate normal distribution, so that the probability that the zonoid depth of the mean of one

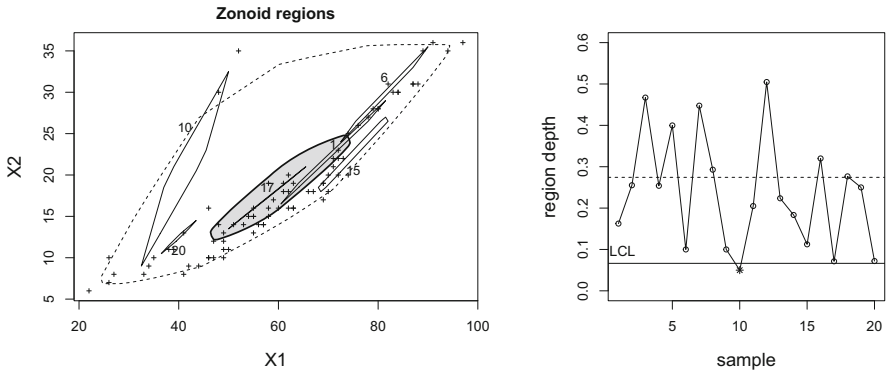


**Fig. 6** Zonoid regions of level 1/2 of Ryan's pooled dataset, in grey, and samples #1, #6, #10, #15, #17, and #20 together with dashed contour of the zonoid region of the pooled dataset that identifies location outliers (left) and evolution of the region depth of the samples with respect to the pooled dataset (right)

given sample of $n = 4$ observations out of the $k = 20$ simulated ones lies below the control limit is 0.025. The dashed horizontal line points out the median depth of the Monte Carlo samples. Anomalies are here detected in samples #9, #10, and #15. The dashed curve in Fig. 4 is the zonoid region of the pooled dataset whose level is the lower control limit. Please observe that the three previously mentioned sample means lie out of it.

### 4.1.2 Monitoring with the region depth

The region depth chart has been built for Ryan's dataset with nominal false alarm probability 0.05. Since the region depth captures not only the location of a sample, but also other features such as the scatter and correlation, the Bonferroni correction suggests to use as nominal false alarm probability the addition of the false alarm probability 0.025 used to monitor the location (see the Hotelling $T^2$ and the zonoid charts) and whatever other false alarm probability (take also 0.025) is used to monitor the remaining features. This particular dataset can be adjusted by a bivariate normal model, which has five parameters. By fixing the same false alarm probability (0.025) to the two means and to the other three parameters (two standard deviations and the correlation), we give slightly more importance to shifts in the location parameters than in the other ones.

Figure 6(left) shows a scatterplot of the pooled Ryan's dataset comprising all 80 observations and their grey-coloured zonoid region of level 1/2 together with other six zonoid regions of level 1/2 whose respective sample numbers (#1, #6, #10, #15, #17, and #20) are displayed close to each one of them. The dashed line is the contour of the zonoid region of the pooled dataset with level half of the lower control limit of the nearby control chart. After $RD4$ in Proposition 4, any sample whose zonoid region of level 1/2 does not lie inside that dashed contour should be declared as out-of-control. Nevertheless, it is also possible to detect anomalies in the samples whose $\mathrm{ZD}^{1/2}$ statistic lies inside the dashed contour by means of the region depth. Figure 6 (right) shows a control chart for the evolution of the region depth with respect to the

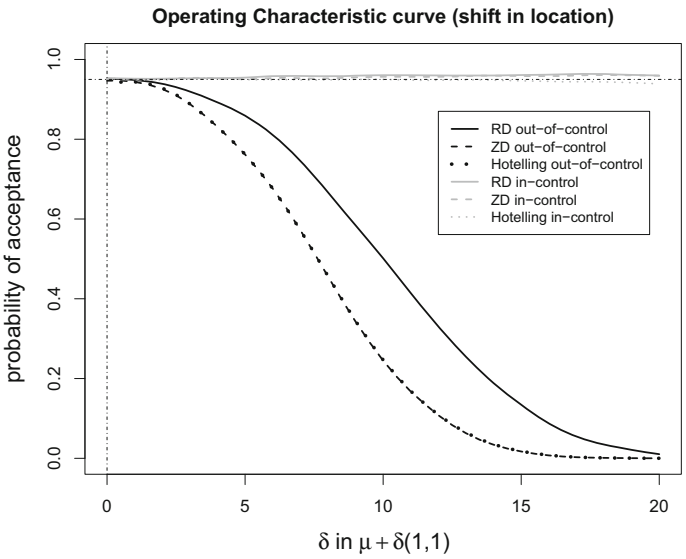


**Fig. 7** OC curves of the region depth (solid), the zonoid depth (dashed), and the Hotelling $T^2$ chart (dotted) for nominal false alarm probability 0.05. Axis X represents shifts in location on both coordinates

pooled dataset through the 20 samples. Only sample #10 is flagged as anomalous, but also samples #17 and #20 have very low depths. Sample #20 lies far for the centre of the pooled dataset and is less scattered than the reference distribution, while the observations of sample #17 are almost aligned, having thus a univariate projection with a very small variability.

### 4.1.3 Performance comparison

In order to compare the performance of the region depth in the detection of shifts in mean, scale, and correlation with the Hotelling $T^2$ chart and the zonoid depth, we obtained 5000 Monte Carlo simulations of 20 samples of $n = 4$ observations of a bivariate normal distribution; 19 of them with the parameters being rounded estimations of those of Ryan's dataset (means $\mu_1 = 60$, $\mu_2 = 18$, standard deviations $\sigma_1 = 17$, $\sigma_2 = 8$, and correlation $\rho = 0.85$), while one sample suffered the shift in mean (Fig. 7), scale (Fig. 8), and correlation (Fig. 9) described on each operating characteristic (OC) curve. In black, we represent the probability that the special sample is not detected in the control chart and is declared as in-control. Under the same circumstances, in grey, we represent the probability that one given sample out of the 19 simulated with the correct parameters is also declared as in-control.

In Figure 7 (black curves), we observe that the Hotelling $T^2$ chart and the zonoid depth chart have the same power to detect shifts in location, while the performance of the region depth is poorer in this respect. The false alarm probability in the presence of one sample with a shift in location (in grey) is similar for the zonoid and region depth charts and below 0.05. Nevertheless, when there is a large shift in location in

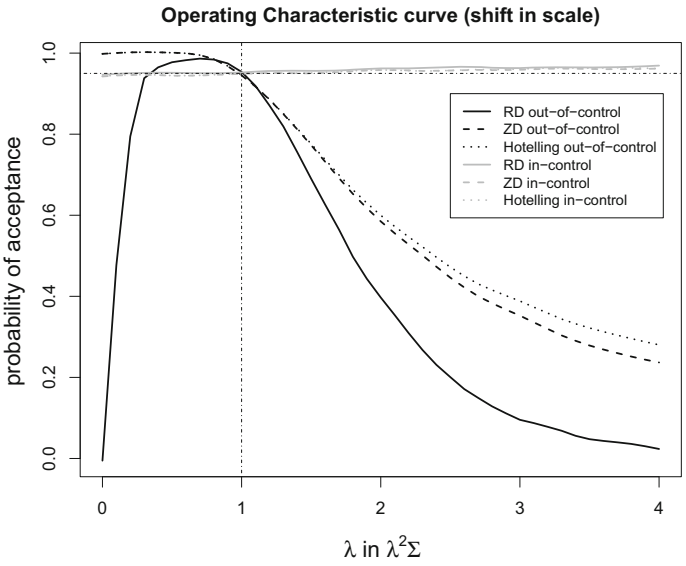


**Fig. 8** OC curves of the region depth (solid), the zonoid depth (dashed), and the Hotelling $T^2$ chart (dotted) for nominal false alarm probability 0.05. Axis X represents shifts in scale

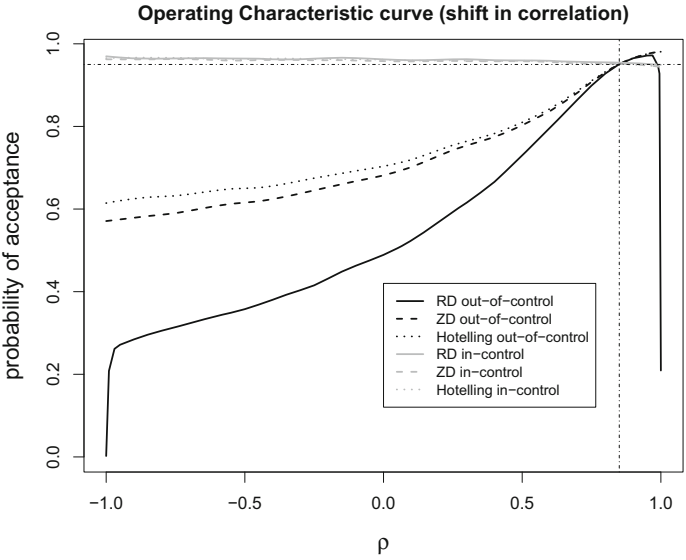


**Fig. 9** OC curves of the region depth (solid), the zonoid depth (dashed), and the Hotelling $T^2$ chart (dotted) for nominal false alarm probability 0.05. Axis X represents the correlation coefficient, with $\rho = 0.85$ the in-control one

one sample, the grand mean is affected and the Hotelling $T^2$ chart wrongly identifies an in-control sample as out-of-control with a probability greater than 0.05.

In Fig. 8 (black curves), we observe that the region depth detects increments in scale much faster than the zonoid depth, whose behaviour in that respect is slightly better than the one of the Hotelling $T^2$ chart. None of the three charts is able to detect slight decrements in scale, but for large decrements, the region depth detects them, while the zonoid depth and Hotelling $T^2$ chart are unable. This makes perfect sense, since both charts are specifically designed to detect shifts in location, and only detect a shift in scale when it, by chance, also affects the location. Instead, the region depth monitors all features captured by the zonoid region of level $1/2$, in particular the scale. The false alarm probability in the presence of one sample with a shift in scale (in grey) is similar for the three charts.

In Fig. 9 (black curves), we observe that the region depth detects shifts in correlation much faster than the zonoid depth, whose behaviour in that respect is slightly better than the one of the Hotelling $T^2$ chart. The false alarm probability in the presence of one sample with a shift in correlation (in grey) is similar for the three charts.

## 5 Depth for interval-valued data

In Sect. 3, we studied a notion of depth for intervals with respect to a univariate distribution. Now we will analyze the situation where not only the argument, but also the available data are interval-valued. In this setting, consider a random interval $\mathsf{X} = [X_l, X_u]$, whose endpoints are the real-valued random variables $X_l$ and $X_u$, which represent the infimum and the supremum of the interval values that $\mathsf{X}$ takes, respectively. The distribution of a random interval can be characterized by that of the bivariate random vector $(X_l, X_u)$ when its components fulfill the order restriction $X_l \leq X_u$.

***Location depth*** The standard notion of expectation for random intervals is the selection or Aumann mean, see e.g., Molchanov (2017), for which the expectation of $\mathsf{X} = [X_l, X_u]$ with both $X_l$ and $X_u$ having finite first moments is the compact interval with endpoints $\mathbb{E}X_l$ and $\mathbb{E}X_u$. We can assess the fit of a deterministic interval $[x_l, x_u]$ with respect to the random interval $\mathsf{X}$ as the largest trimming level $d$ such that $[x_l, x_u]$ is the expectation of a random interval whose endpoints' distribution lies in $F^{(d)}_{X_l,X_u}$, or equivalently as the zonoid depth $\mathrm{ZD}((x_l, x_u); F_{X_l,X_u})$. Notice that $F^{(d)}_{X_l,X_u}$ only involves cdfs of bivariate random vectors that fulfill the order relationship between their components, $X_l \leq X_u$, because the distributions that belong to $F^{(d)}_{X_l,X_u}$ are absolutely continuous with respect to the distribution of the random vector $(X_l, X_u)$. Consequently, their support lies in the halfspace $\{(x, y) \in \mathbb{R}^2 : x \leq y\}$, and all these distributions are possible options for the distribution of a random interval.

In particular, if $[x_l, x_u]$ corresponds to the expectation of another random interval $\mathsf{Y} = [Y_l, Y_u]$, we can compute $\mathrm{ZD}((\mathbb{E}Y_l, \mathbb{E}Y_u); F_{X_l,X_u})$ to evaluate the similarity of the location of both random intervals in terms of their expectations.

***Remark 3*** Random intervals are frequently characterized in terms of their midpoint (or centre) and spread (or radius), instead of the infimum and the supremum, when applying statistical techniques for interval-valued data since the nonnegativity restriction associated with the spread is more easy-to-handle than the order restriction mentioned above. However, any of these two characterizations provide the same results concerning the zonoid depth due to its affine invariance.

***Location and scale depth*** If our goal is to assess the fit of a random interval $\mathsf{Y} = [Y_l, Y_u]$ (maybe through its empirical distribution based on a small sample) with respect to the distribution of some other random interval $\mathsf{X} = [X_l, X_u]$ in terms of both location and scale, we can evaluate the region depth of the zonoid region of level $1/2$ of the bivariate random vector $(Y_l, Y_u)$ with respect to the distribution of $(X_l, X_u)$, that is, $\mathrm{RD}(\mathrm{ZD}^{1/2}(F_{Y_l,Y_u}); F_{X_l,X_u})$.

Concerning the properties this proposal presents, Proposition 4 admits a natural adaptation when the first argument of the region depth is a zonoid region of level $1/2$. Furthermore, thanks to the fact that $\mathrm{ZD}^{1/2}(F_{Y_l,Y_u})$ is centrally symmetric about $(\mathbb{E}Y_l, \mathbb{E}Y_u)$, the application of *RD5'* allows us to consider the upper bound $\mathrm{RD}(\mathrm{ZD}^{1/2}(F_{Y_l,Y_u}); F_{X_l,X_u}) \leq \mathrm{ZD}((\mathbb{E}Y_l, \mathbb{E}Y_u); F_{X_l,X_u})$.

### 5.1 Monitoring processes with interval-valued observations

Hsu et al. (2013) present some interval-valued data on the luminous intensity of LEDs (in cd) taken during their fabrication process. Specifically, their dataset consists of 24 samples of 4 intervals each. In order to perform a Phase I analysis that allows us to detect samples with anomalous observations among the available (trial) ones, we first monitor the sample location with the zonoid depth of each sample average with respect to the pooled dataset and then we use the region depth for the joint monitoring of sample location and scale by evaluating the region depth of each sample zonoid region of level $1/2$ with respect to the pooled dataset.

Despite we do not build a specific peformance comparison for the zonoid depth and the region depth for interval-valued data, we refer to that for multivariate data presented in Sect. 4.1.3. Nevertheless, notice that the Hotelling $T^2$ chart is not appropriate in this setting since the inequality restriction on the endpoints of an interval prevents the bivariate normal distibution from being a suitable model for them. Based on the previous performance comparison, the zonoid depth would allow a faster detection of shifts in location, while the region depth would detect shifts in endpoints correlation and scale faster. The reason is that the information about the scale and correlation of each individual trial sample is lost when they are summarized in the average value used by the zonoid chart, while it is kept when the sample is summarized in terms of the zonoid central region used by the region depth chart.

#### 5.1.1 Monitoring the location

For each of the 24 trial samples, we have obtained the average interval, presented in Fig. 10 (left) as a thick line segment, and computed the zonoid depth of its endpoints

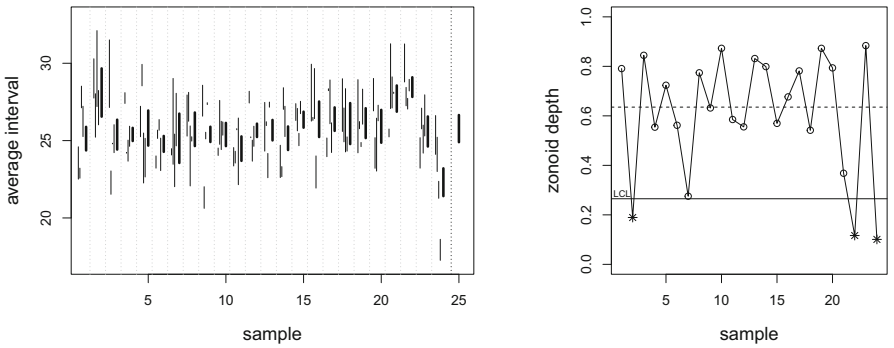


**Fig. 10** Interval-valued observations (thin line segments) together with the average interval for each of the 24 samples (thick line segments) followed by the one of the pooled dataset on the 25th position (left) and evolution of the zonoid depth with respect to the reference distribution (right)

with respect to the $24 \times 4 = 96$ pairs of endpoints of all the observed intervals (presented as thin line segments), which is considered as the reference distribution.

In order to identify anomalous samples, $10^4$ bootstrap samples of 4 intervals each were taken from the sample containing all intervals and its 0.05-quantile was established as control limit (solid horizontal line) in the zonoid depth chart, see Fig. 10 (right). The dashed horizontal line points out the median depth of the bootstrap samples. Samples #24 (smallest lower and upper average endpoints), #22 (greatest lower average endpoint and second greatest upper average endpoint), and #2 (greatest upper average endpoint) are flagged as anomalous.

### 5.1.2 Joint monitoring of location and scale

For each of the 24 trial samples, we have obtained the empirical zonoid region of level $1/2$ of its 4 pairs of endpoints and computed its region depth with respect to the $24 \times 4 = 96$ pairs of endpoints of all the observed intervals, which is considered as the reference distribution. Notice that each zonoid region of level $1/2$ contains the pairs of endpoints of all intervals that can be formed by averaging any subset of half of the intervals from the original dataset. Some of those zonoid regions are presented in Fig. 11 (left) with the respective sample number close to each one for identification. The dashed line is the contour of the zonoid region of the pooled dataset with level half of the lower control limit of the nearby control chart.

In order to identify anomalous samples, $10^4$ bootstrap samples of 4 intervals each were taken from the sample containing all intervals and its 0.05-quantile was established as control limit in the region depth chart, see Fig. 11 (right). Only sample #24, which has the smallest upper and lower average endpoints plus the largest in-sample variability of both endpoints, is flagged as anomalous.

Based on their extensions of Shewhart's $\overline{X}$ and $R$ charts with degrees of uncertainty for interval-valued data, Hsu et al. (2013) flag sample #22 as out-of-control, while they declare samples #21 and #24 to be *rather out-of-control*. Unfortunately, they do not provide a false alarm probability, so we cannot compare our results with theirs.

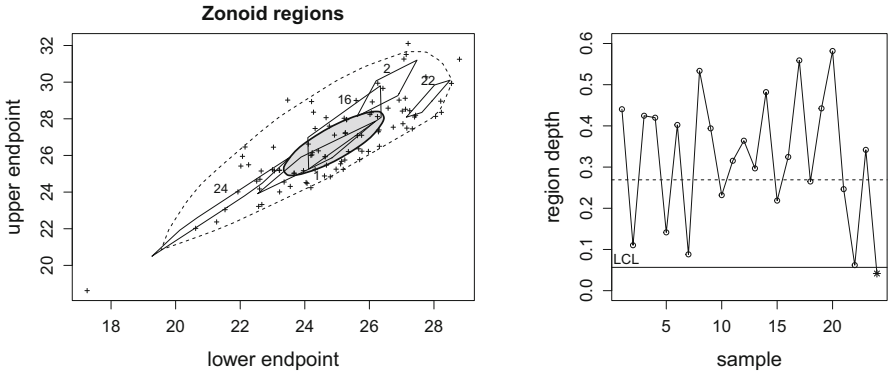


**Fig. 11** Scatterplot of all $24 \times 4 = 96$ pairs of endpoints of the interval observations together with the zonoid regions of level $1/2$ of some of the 24 samples plus the one of the pooled dataset in grey and the dashed contour of the zonoid region of the pooled dataset that identifies location outliers (left) and evolution of the region depth with respect to the reference distribution (right)

## 6 Conclusions

A new notion of parameter depth, called (zonoid) interval depth, has been introduced and extended to the multivariate setting. Some theoretical properties in the fashion of the classical properties of depth functions, but adapted to the situation when the argument is a set have been derived. The new proposal assesses the fit in terms of all those distributional features captured by a central region. Finally, real-life data applications have been presented to illustrate how to use the proposed depth in statistical process control with multivariate observations and interval-valued data. The simulation study conducted for one of these applications shows that the new procedure detects shifts in scale and correlation faster than other classical ones.

In order to build a multivariate extension of the interval depth, a projection argument has been used since, as argued in Remark 2, a straigtforward extension is not feasible. An alternative approach that we aim to pursue in the future is to approximate the set whose depth is to be computed from inside and outside separately, and then define its depth in terms of the trimming levels that allow the most accurate inner and outer approximations.

**Acknowledgements** The authors are grateful to the reviewers for their insightful comments and suggestions. IC acknowledges the support of the Community of Madrid by the V Regional Plan of Scientific Research and Technological Innovation 2016-2020 and that of the Spanish Ministry of Science and Innovation by Grants PID2021-123592OB-I00 and TED2021-129316B-I00; IC and BS acknowledge the support of the Principality of Asturias/FEDER Funds by grants GRUPIN-IDI2018-000132 and SV-PA-21-AYUD/2021/50897; BS acknowledges the support of the Spanish Ministry of Science and Innovation by grants PID2019-104486GB-I00 and MTM2015-63971-P. GP acknowledges the support of the National Operative Program (PON) Ricerca e Innovazione 2014-2020 (PON R&I) - Azione IV.4 - "Dottorati e contratti di ricerca su tematiche dell'innovazione".

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

## References

Cascos I, López-Díaz M (2012) Trimmed regions induced by parameters of a probability. J Multivar Anal 107:306–318
Cascos I, López-Díaz M (2016) On the uniform consistency of the zonoid depth. J Multivar Anal 143:394–397
Cascos I, López-Díaz M (2018) Control charts based on parameter depths. Appl Math Model 53:487–509
Cascos I, Li Q, Molchanov I (2021) Depth and outliers for samples of sets and random sets distributions. Aust N Z J Stat 63:55–82
Chen M, Gao C, Ren Z (2018) Robust covariance and scatter matrix estimation under Huber's contamination model. Ann Stat 46:1932–1960
Dyckerhoff R (2004) Data depths satisfying the projection property. Allg Stat Arch 88:163–190
González-De la Fuente L, Nieto-Reyes A, Terán P (2022) Statistical depth for fuzzy sets. Fuzzy Sets Syst 443:58–86
Hsu B-M, Kung J-Y, Shu M-H (2013) Interval-valued process data monitoring and controlling. Artif Intell Res 2:90–101
Koshevoy G, Mosler K (1997) Zonoid trimming for multivariate distributions. Ann Stat 25:1998–2017
Liu RY (1990) On a notion of data depth based on random simplices. Ann Stat 18:405–414
Liu RY (1995) Control charts for multivariate processes. J Am Stat Assoc 90:1380–1387
Liu RY, Parelius JM, Singh K (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). Ann Stat 27:783–858
Mizera I, Müller CH (2004) Location-scale depth. J Am Stat Assoc 99:949–966
Molchanov I (2017) Theory of random sets, 2nd edn. Springer, London
Montgomery DC (2013) Introduction to statistical quality control, 7th edn. Wiley, Hoboken
Mosler K (2002) Multivariate dispersion, central regions and depth: the lift zonoid approach. Springer, New York
Mosler K (2013) Central regions and dependency. Methodol Comput Appl Probab 5:4–21
Paindaveine D, Van Bever G (2018) Halfspace depths for scatter, concentration and shape matrices. Ann Stat 46:3276–3307
Rousseeuw PJ, Hubert M (1999) Regression depth. J Am Stat Assoc 94:388–402
Ryan TP (2011) Statistical methods for quality improvement, 3rd edn. Wiley, New York
Schneider R (1993) Convex bodies. The Brunn-Minkowski theory. Cambridge University Press, Cambridge
Sinova B (2022) On depth-based fuzzy trimmed means and a notion of depth specifically defined for fuzzy numbers. Fuzzy Sets Syst 443:87–105
Tukey JW (1975) Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, vol 2. Vancouver, pp 523–531
Zuo Y, Serfling R (2000) General notions of statistical depth function. Ann Stat 28:461–482