Data analysis and classification with the zonoid depth

Karl Mosler and Richard Hoberg

September 14, 2004

Abstract

The zonoid depth is useful in multivariate data analysis in order to describe an empirical distribution by trimmed regions, so-called zonoid regions. The zonoid regions range from the convex hull of the data to their mean and characterize the distribution in a unique way. This paper introduces into some principal data analytic applications of the zonoid depth, including recent developments in classification.

Multivariate data are often asymmetrically distributed so that they cannot be modelled by a normal or elliptical probability distribution. Various notions of data depth have been proposed to analyze such data in a nonparametric way. This paper surveys the zonoid depth and some recent developments of its use in data analysis.

The zonoid depth is useful in multivariate data analysis in order to describe an empirical distribution by trimmed regions, so-called zonoid regions. These regions range from the convex hull of the data to their mean and characterize the distribution in a unique way.

The deepest point, which equals the mean, depicts the location of the data. The volumes of the regions measure dispersion and dependence; their shape reflects dependence and asymmetry. Both the zonoid regions and the zonoid depth have pleasant analytic and computational properties. Many descriptive and inferential methods are based on the zonoid depth and profit from these properties.

In the sequel we introduce into some principal data analytic applications of the zonoid depth. We start with the notion of zonoid depth and relate it to a framework of affine-invariant, convex data depths. The principal properties of the zonoid depth are surveyed, in particular, uniqueness, the projection property and various continuity properties of the depth and the trimmed regions. Then computability issues and the use of zonoid depth in describing and comparing location, dispersion and shape are covered. In the last part, recent applications to the classification problem are presented.

1 Zonoid regions and zonoid depth

Given a d-variate probability distribution function F, a family $\{D_\alpha\}$ of nested sets in d-space is considered. The sets are called the *zonoid regions* of F and defined as follows: Let $D_0(F) = \mathbb{R}^d$ and for $\alpha \in]0,1]$

$$D_\alpha(F) \quad = \quad \left\{ \int_{\mathbb{R}^d} \boldsymbol{x}\, g(\boldsymbol{x})\, dF(\boldsymbol{x}) : 0 \le g \le \frac{1}{\alpha}, \quad \int_{\mathbb{R}^d} g(\boldsymbol{x})\, dF(\boldsymbol{x}) = 1 \right\}. \tag{1.1}$$

These regions exist if and only if F has a finite expectation vector $\mu_F = \int_{\mathbb{R}^d} \boldsymbol{x} dF(\boldsymbol{x})$.

Especially, if F is an empirical distribution, with equal mass on (not necessarily different) points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, then (1.1) becomes

$$\begin{aligned} D_\alpha(F) &= D_\alpha(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \\ &= \left\{ \sum_{i=1}^n \lambda_i \boldsymbol{x}_i : \sum_{i=1}^n \lambda_i = 1,\ 0 \le \lambda_i,\ \lambda_i \le \frac{1}{\alpha} \text{ for all } i \right\}, \end{aligned} \tag{1.2}$$

which holds for $0 \le \alpha \le 1$. It is easily seen from the definition (1.1) that the zonoid regions are nested, the smallest region being the singleton set $D_1(F) = \{\mu_F\}$. If F is an empirical distribution, the zonoid regions range from the mean to the convex hull of the data; the latter arises when α is close enough to zero, $0 < \alpha \le \frac{1}{n}$.

The zonoid regions satisfy the following general properties. For every $\alpha \in]0,1]$, D_α is

R1 **affine equivariant**: $D_\alpha(F_{\boldsymbol{X}\boldsymbol{A}+\boldsymbol{c}}) = D_\alpha(F_X)\boldsymbol{A} + \boldsymbol{c}$ for any $d \times d$ matrix $\boldsymbol{A}$ having full rank and $\boldsymbol{c} \in \mathbb{R}^d$.

Further, for every F having finite first moment, $D_\alpha(F)$ is

R2 **bounded**,

R3 **closed**, and

R4 **convex**.

Of these properties R1, R2, and R4 are obvious from the definition (1.1); for a proof of R3, see (Koshevoy and Mosler, 1997, Th. 5.4(i)).

Now we turn to the definition of the zonoid depth. Given a point $\boldsymbol{y} \in \mathbb{R}^d$, its *zonoid depth* of $\boldsymbol{y}$ with respect to F is defined by

$$d(\boldsymbol{y}; F) = \max\{\alpha : \boldsymbol{y} \in D_\alpha(F)\}\,. \tag{1.3}$$

In other words, the zonoid region $D_\alpha(F)$ is the upper level set of the zonoid depth, with level α.

Given the distribution function $F_{\boldsymbol{X}}$ of a random vector $\boldsymbol{X}$, the zonoid depth is

D1 **affine invariant**: $d(\boldsymbol{y}\boldsymbol{A}+\boldsymbol{b}|F_{\boldsymbol{X}\boldsymbol{A}+\boldsymbol{b}}) = d(\boldsymbol{y}|F_{\boldsymbol{X}})$ for any $d \times d$ matrix $\boldsymbol{A}$ having full rank and $\boldsymbol{b} \in \mathbb{R}^d$,

D2 **vanishing at infinity**: $\lim_{k\to\infty} d(\boldsymbol{y}_k|F_{\boldsymbol{X}}) = 0$ for every sequence $(\boldsymbol{y}_k)_{k\in\mathbb{N}}$ with $\lim_{k\to\infty} ||\boldsymbol{y}_k|| = \infty$,

D3 **upper semi-continuous**: the set $\{\boldsymbol{y} \in \mathbb{R}^d \,|\, d(\boldsymbol{y}|F_{\boldsymbol{X}}) \geq \alpha\}$ is closed for every α,

D4 **quasi-concave**: $d(\lambda\boldsymbol{y} + (1-\lambda)\boldsymbol{z}|F_{\boldsymbol{X}}) \leq \max\{d(\boldsymbol{y}|F_{\boldsymbol{X}}), d(\boldsymbol{z}|F_{\boldsymbol{X}})\}$ for any $\lambda \in [0,1]$ and $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^d$.

Each of these restrictions follows immediately from the definition (1.3) and the affine equivariance, boundedness, closedness and convexity, respectively, of the zonoid regions.

A function $\boldsymbol{y} \mapsto g(\boldsymbol{y}|F)$ which satisfies these restrictions is called an *affine invariant, convex depth function.* The smallest region is mentioned as the median set w.r.t. to the depth. The zonoid median set is the singleton containing the mean of the distribution. Further examples of affine invariant, convex depths are the *halfspace depth* d_H,

$$d_H(\boldsymbol{y}|F) = \inf\left\{\int_H dF(\boldsymbol{x}) : H \text{ a closed halfspace}, \quad \boldsymbol{y} \in H\right\}, \tag{1.4}$$

and the *Mahalanobis depth* d_{Mah},

$$d_{Mah}(\boldsymbol{y}|F) = \left(1 + (\boldsymbol{y} - \mu_F)\Sigma_F^{-1}(\boldsymbol{y} - \mu_F)'\right)^{-1}, \tag{1.5}$$

where Σ_F denotes the covariance matrix of F.

For a set of 15 data points, Figure 1 exhibits several (a) zonoid regions, and regions based on (b) halfspace and (c) Mahalanobis depth.

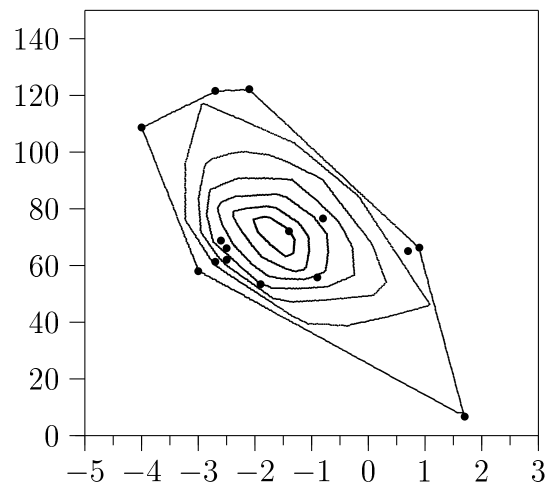
Other depth functions proposed in the literature satisfy only weaker sets of restrictions. For a general investigation into depth functions see Zuo and Serfling (2000a). E.g. the simplicial depth (Liu (1990)), restricted to L-continuous distributions, has level sets which are star-shaped, but in general not convex. The $\mathbb{L}_1$-*depth* (Vardi and Zhang (2000)) and the *Euclidean depth* d_E,

$$d_E(\boldsymbol{y}, F) = (1 + ||\boldsymbol{y} - \mu_F||^2)^{-1}, \tag{1.6}$$

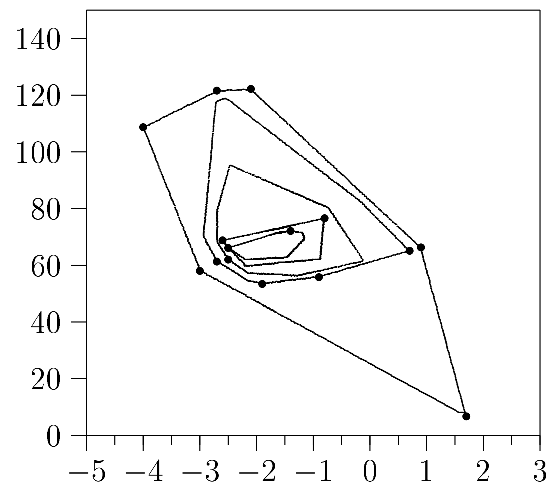
are only spherically invariant.

2 Useful properties of the zonoid depth

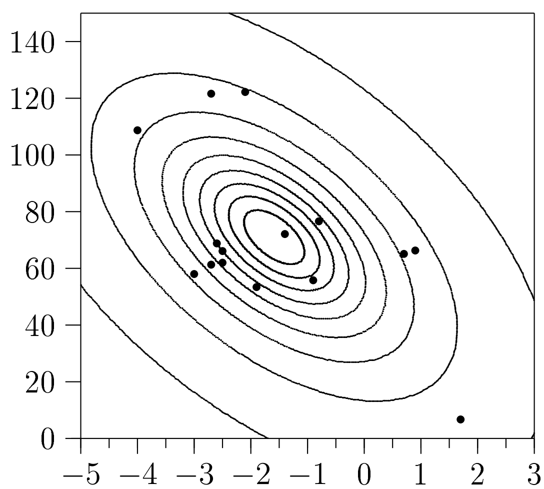
In this section we mention several special properties of the zonoid depth that are particularly useful in statistical applications. They concern uniqeness, projection, continuity, and laws of large numbers.



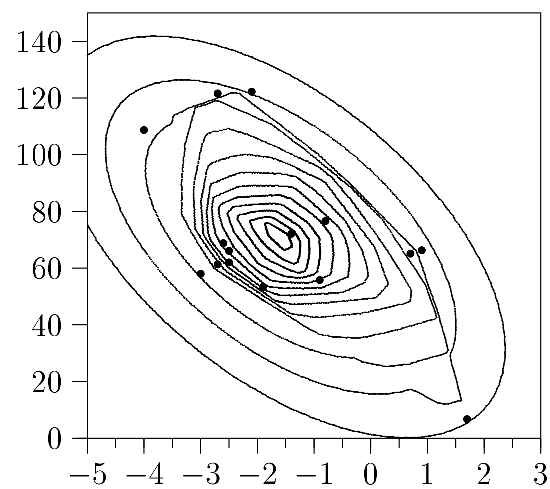
(a) zonoid depth ($\alpha = 1/15,\ 3/15,\ 5/15,\ 7/15,\ 9/15,\ 11/15,\ 13/15$)



(b) halfspace depth ($\alpha = 1/15,\ 2/15,\ 3/15,\ 4/15,\ 5/15$)



(c) Mahalanobis depth ($\alpha = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9$)



(d) zonoid–Mahalanobis depth with $\beta = .33$

Figure 1: Regions based on different depths.

The first property is *uniqueness*: The zonoid depth characterizes the underlying distribution in a unique way. More precisely, for any two d-variate distribution functions F and G that have finite first moments,

$$F = G \qquad \text{if} \qquad d(\boldsymbol{y}|F) = d(\boldsymbol{y}|G) \quad \text{for all } \ \boldsymbol{y} \in \mathbb{R}^d . \tag{2.1}$$

The uniqueness property follows from (Koshevoy and Mosler, 1997, Th. 5.6). It implies that any claim about a distribution F can be equivalently formulated and analyzed as a claim about the zonoid depth w.r.t. F. Note that, e.g., with the Mahalanobis depth we are not able to distinguish between distributions that have the same expectation and covariance. The halfspace depth characterizes the distribution uniquely if either discrete distributions are assumed (Struyf and Rousseeuw (1999), Koshevoy (200x)) or continuous distributions with compact support (Koshevoy (2002)).

The second property is the *projection property* (Dyckerhoff (2002); see (Mosler, 2002, Th. 4.7)). It says that the zonoid depth at some $\boldsymbol{y} \in \mathbb{R}^d$ equals the infimum of the zonoid depths of all univariate projections,

$$d_Z(\boldsymbol{y}|F_{\boldsymbol{X}}) = \inf_{\boldsymbol{p}\in S^{d-1}} d_Z(\boldsymbol{p}'\boldsymbol{y}|F_{\boldsymbol{p}'\boldsymbol{X}}), \quad \boldsymbol{y} \in \mathbb{R}^d . \tag{2.2}$$

Due to the projection property, every value of the zonoid depth can be numerically approximated by calculating the values of the depth in several univariate projections and taking the minimum value. This appears to be particularly useful when d is large. The halfspace depth satisfies another projection property: For every $\alpha \in]0,1]$,

$$d_Z(\boldsymbol{y}|F_{\boldsymbol{X}}) \geq \alpha \qquad \text{iff} \qquad d_Z(\boldsymbol{p}'\boldsymbol{y}|F_{\boldsymbol{p}'\boldsymbol{X}}) \geq \alpha \quad \text{for all } \boldsymbol{p} \in S^{d-1} . \tag{2.3}$$

It can be shown (Dyckerhoff (2002)) that (2.3) is implied by (2.2) but not vice versa.

Thirdly, we mention the *continuity property*: The zonoid depth is continuous on the point and on the distribution: Given a convergent series of points $\boldsymbol{y}_n \to \boldsymbol{y}$ it holds that

$$\lim_{\boldsymbol{y}_n \to \boldsymbol{y}} d_Z(\boldsymbol{y}_n|F) = d_Z(\boldsymbol{y}|F)$$

provided all $\boldsymbol{y}_n$ are in the convex hull of the support of F. Given a weakly convergent series of distributions $F_n \to F$ we obtain that

$$d_Z(\boldsymbol{y}|F_n) \to d_Z(\boldsymbol{y}|F)$$

if the series is uniformly integrable and $\boldsymbol{y}$ is in the relative interior of the convex hull of the support of F. For proofs see (Koshevoy and Mosler, 1997, Th. 7.1(iii)) and (Mosler, 2002, Th. 4.5).

Continuity on the distribution means that the depth varies only slowly with small perturbations of the data. Continuity on the point implies that the depth has a finite maximum and that the trimmed regions are closed. Note that the Mahalanobis depth is continuous in both respects, while the halfspace depth is obviously not.

The fourth special property is a *Law of Large Numbers*. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be an i.i.d. sample, with $\boldsymbol{X}_i \sim F$, and denote the empirical distribution function by $\widetilde{F}_n$. Then almost surely holds

$$D_\alpha(\widetilde{F}_n) \quad \xrightarrow{H} \quad D_\alpha(F) \quad \text{for every } \alpha$$

and

$$d_Z(\boldsymbol{y}|\widetilde{F}_n) \to d_Z(\boldsymbol{y}|F) \quad \text{for every } \ \boldsymbol{y}\,,$$

which follows from (Koshevoy and Mosler, 1997, Cor. 5.1). Here, $\xrightarrow{H}$ means convergence with respect to Hausdorff distance. (The Hausdorff distance of two compacts C and D is the smallest ϵ for which C plus an ϵ-ball includes D and D plus an ϵ-ball includes C as well.)

For the halfspace depth holds a similar law (Donoho and Gasko (1992)), and for the Mahalanobis depth clearly the same.

3 Computational issues

To be practical in data analysis, a depth has to be numerically evaluated at many points. However, the computational load differs considerably among the various depths.

Given a d-variate data set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, the Mahalanobis depth of a point $\boldsymbol{y}$ is easily calculated. Also, determining the Mahalanobis trimmed region D_α at some α needs no more computational effort.

Things are different with the halfspace depth. Observe that, when the points $\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are moved, the value of $d_H(\boldsymbol{y}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ remains unchanged as long as the combinatorial structure of the $n+1$ points is kept, that is, no point reaches or crosses a hyperplane spanned by the others. This property of a depth is called *combinatorial invariance*. Also the simplicial depth, the majority depth and the convex-hull peeling depth are combinatorial invariant, among others; see, e.g. (Mosler, 2002, p 127). Calculating a combinatorial invariant depth is always a heavy task. For the halfspace depth exist algorithms in dimension $d = 2$ (Rousseeuw and Ruts (1996)) and $d \geq 3$ (Rousseeuw and Struyf (1998)) having complexity $O(n^{d-1}\log(n))$.

The zonoid depth is not combinatorial invariant. It relies not only on the combinatorial but also on the metric structure of the data. The zonoid depth can be efficiently calculated by an LP algorithm (Dyckerhoff et al. (1996)).

Table 1 shows the computer time needed to calculate the zonoid, halfspace and Mahalanobis depths of a single point for various dimensions d and sample sizes n. The asterisk $*$ says that time exceeded 24 hours. We see from the table that the halfspace depth, due to the complexity of the exact algorithm, can serve as a practical device only if the data set is small and the dimension low. On the other hand, the zonoid depth can be numerically handled for relatively large data sets and high dimensions.

While this true for the zonoid depth of a single point, the calculation of a zonoid region appears to be a much heavier task. Dyckerhoff (2000) has developed an algorithm to compute the zonoid region $D_\alpha(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ when $d = 2$. It is based on the support function of the zonoid region and has complexity $O(n^2 \log n)$.

	Zoonid			Halfspace			Mahalanobis		
d	$n = 100$	500	1000	$n = 100$	500	1000	$n = 100$	500	1000
2	0.81	6.16	14.41	0.188	1.078	2.176	0.014	0.071	0.141
4	2.21	16.88	39.57	2110	298647	2483118	0.066	0.328	0.647
8	8.08	63.55	151.46	*	*	*	0.16	0.797	1.579

Table 1: Time (in milliseconds) needed to calculate the zonoid, halfspace and Mahalanobis depths of a single point. The asterisk $*$ says that time exceeded 24 hours.

However, for higher dimensions the computational load of an exact calculation of zonoid regions becomes prohibitive.

For approximative calculations of the zonoid and the halfspace depths, the projection properties (2.2) and (2.3) can be usefully employed. For this, see Dyckerhoff (2004).

4 Analysis of location, scale, and dependency

The principle tasks of data analysis are to describe a data set with respect to its location, scale and shape and to compare it with other data sets in these respects.

With any depth, the location of a given distribution is measured by the median set of the depth. For the zonoid depth the median set is a singleton containing the mean vector of the distribution.

The scale of a given distribution F can be measured by the volume of one or several trimmed regions. Especially with the zonoid depth either the volume $S_\alpha(F) = \mathrm{vol}(D_\alpha(F))$ for some selected $\alpha \in]0,1[$, say $\alpha = \frac{1}{2}$, can be used as a scale index, or the weighted integral over all these volumes

$$S(F) = \int_0^1 \alpha^d S_\alpha(F) d\alpha \,,$$

which amounts to the volume of the lift zonoid of F; see Mosler (2002). These volumes do not vanish if and only if the convex hull of the support of F has a non-empty interior.

By these indexes, two given distributions F_X and F_Y can be compared for scale. A partial order of scale is obtained by comparing volumes for every α,

$$F_X \preceq_{scale} F_Y \quad \text{iff} \quad \mathrm{vol}(D_\alpha(F_X)) \leq \mathrm{vol}(D_\alpha(F_Y)) \quad \text{for all} \ \ \alpha \,.$$

This ordering has been investigated by Zuo and Serfling (2000b).

A special aspect of shape is dependency. The ordering $\preceq_{scale}$ may be also employed to measure the dependency of a distribution and to compare two distributions regarding their degrees of dependency (Mosler (2003)).

5 Classification

Consider a finite set C of data points in $\mathbb{R}^d$ which is partioned into given classes $C_1, \ldots, C_k$. An additional data point y has to be assigned to the class to which, in some sense, it 'fits best'. In other words, a new 'object' is assigned to one of several given classes of 'objects'.

To solve this *classification problem*, many rules have been proposed in the literature and successfully used in applications (e.g. Hand, 1981). They differ in their notion of 'best fit' and in the structure imposed on the data.

For every depth function d a *depth classification rule*

$$classd(\boldsymbol{y}) = argmax_j d(\boldsymbol{y}|C_j)$$

is defined. The rule assigns $\boldsymbol{y}$ to that class C_j in which $\boldsymbol{y}$ is deepest. Denote $C_j = \{\boldsymbol{x}_{j1}, \ldots, \boldsymbol{x}_{jn_j}\}$. Especially, with the *Euclidean depth*

$$d_{Euc}(\boldsymbol{y}|C_j) = \frac{1}{1 + ||\boldsymbol{y} - \overline{\boldsymbol{x}}_j||^2}\,, \quad \text{where } \overline{\boldsymbol{x}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \boldsymbol{x}_{ji}\,,$$

the classic *centroid classification rule* is obtained, that is, $\boldsymbol{y}$ is assigned to the class C_j that has 'centroid' $\overline{\boldsymbol{x}}_j$ closest to $\boldsymbol{y}$ in Euclidean distance. Alternatively, the *Mahalanobis depth*

$$d_{Mah}(\boldsymbol{y}|C_j) = \frac{1}{1 + (\boldsymbol{y} - \overline{\boldsymbol{x}}_j)'\Sigma_j^{-1}(\boldsymbol{y} - \overline{\boldsymbol{x}}_j)}$$

with Σ_j denoting the covariance matrix of group C_j, yields the well known *Mahalanobis classification rule*. While with the Euclidean depth the points of equal given depth form a sphere around $\overline{\boldsymbol{x}}_j$, with the Mahalanobis depth they form the border of an ellipsoid. Thus, the distance from class C_j is measured in an elliptically symmetric way, which appears to be a natural distance if the data in the class have an elliptically symmetric distribution, like normal data, but not if they are distributed in an asymmetric way.

A drawback of the zonoid depth (and many other depths) is that it vanishes outside the convex hull of the set C_j. By this, a point $\boldsymbol{y}$ lying outside the convex hulls of all classes has zonoid depth 0 with respect to all classes and thus cannot be classified to one of them. Therefore a new depth has been introduced by Hoberg (2003), named *zonoid-Mahalanobis depth*,

$$d_{ZMah}(\boldsymbol{y}|C_j) = \max\{d_Z(\boldsymbol{y}|C_j), \beta \cdot d_{Mah}(\boldsymbol{y}|C_j)\} \quad \textit{with} \quad \beta = \frac{1}{\max_j n_j}\,.$$

This function is an affine-invariant convex depth as well, and it is positive at all $\boldsymbol{y} \in \mathbb{R}^d$. It equals the zonoid depth inside the convex hull of C_j and is a multiple of the Mahalanobis depth outside. Thus it extends the zonoid depth beyond the convex hull. See Figure 1(d) for an example of regions based on the zonoid-Mahalanobis depth. Classification by this depth is called the *zonoid-Mahalanobis rule*.

This new classification rule has been applied to several small benchmark data sets from the literature and compared with known classification rules. Figure 2 shows four artificial data sets, which are taken from the literature. For HARDY1, HARDY2 and HARDY3, see Hardy (1991, 1994, 1996), for RUSPINI, see Ruspini (1970).

Three depth rules have been investigated (Mahalanobis, zonoid, zonoid-Mahalanobis) and compared with seven classical rules, among them four rules based on density estimation with different kernels (histogram, rectangle, Gauss, Epanechnikov), two nearest neighbour rules, and the hypervolume rule; see Hand (1981), Baufays and Rasson (1985).

The classification rules are evaluated with respect to their *apparent error rates* as well as to their *leave-one-out (loo) error rates* ((Lachenbruch, 1975, Ch 2), Lachenbruch (1968)). To determine the apparent error rate of a rule, each point $\boldsymbol{x}_{ji}$ is classified with that rule to one of the given classes $C_1, \dots C_k$; the apparent error rate is then defined as the portion of falsely classified points. To determine the *loo* error rate, the same is done with the 'correct' class C_j substituted by $C_j \setminus \{x_{ji}\}$. Obviously the *loo* error rate is always greater or equal to the apparent error rate.

	RUSPINI	HARDY1	HARDY2	HARDY3
Histogramm	0.00	0.00	0.00	0.00
Rectangle	0.00	0.00	0.00	0.00
Gauss	0.00	0.01	0.00	0.00
Epanechnikov	0.00	0.00	0.00	0.00
Nearest neighbour	0.00	0.01	0.00	0.00
Nearest neighbour (mod.)	0.00	0.01	0.00	0.00
Mahalanobis	0.00	0.00	0.06	0.00
Zonoid	0.00	0.00	0.02	0.00
Hypervolume	0.00	0.00	0.02	0.00
Zonoid-Mahalanobis	0.00	0.00	0.02	0.00

Table 2: Apparent error rates of various classification rules applied to the data in Figure 2.

Table 2 presents the apparent error rates, Table 3 the leaving-one-out error rates. For two data sets, RUSPINI and HARDY3, the apparent error rate is zero with each of the ten rules, while only five of the rules yield a leaving-one-out error rate of zero.

With the zonoid classification rule, the difference between the two error rates stands out. If the classes have disjoint convex hulls, as it is the case in RUSPINI,

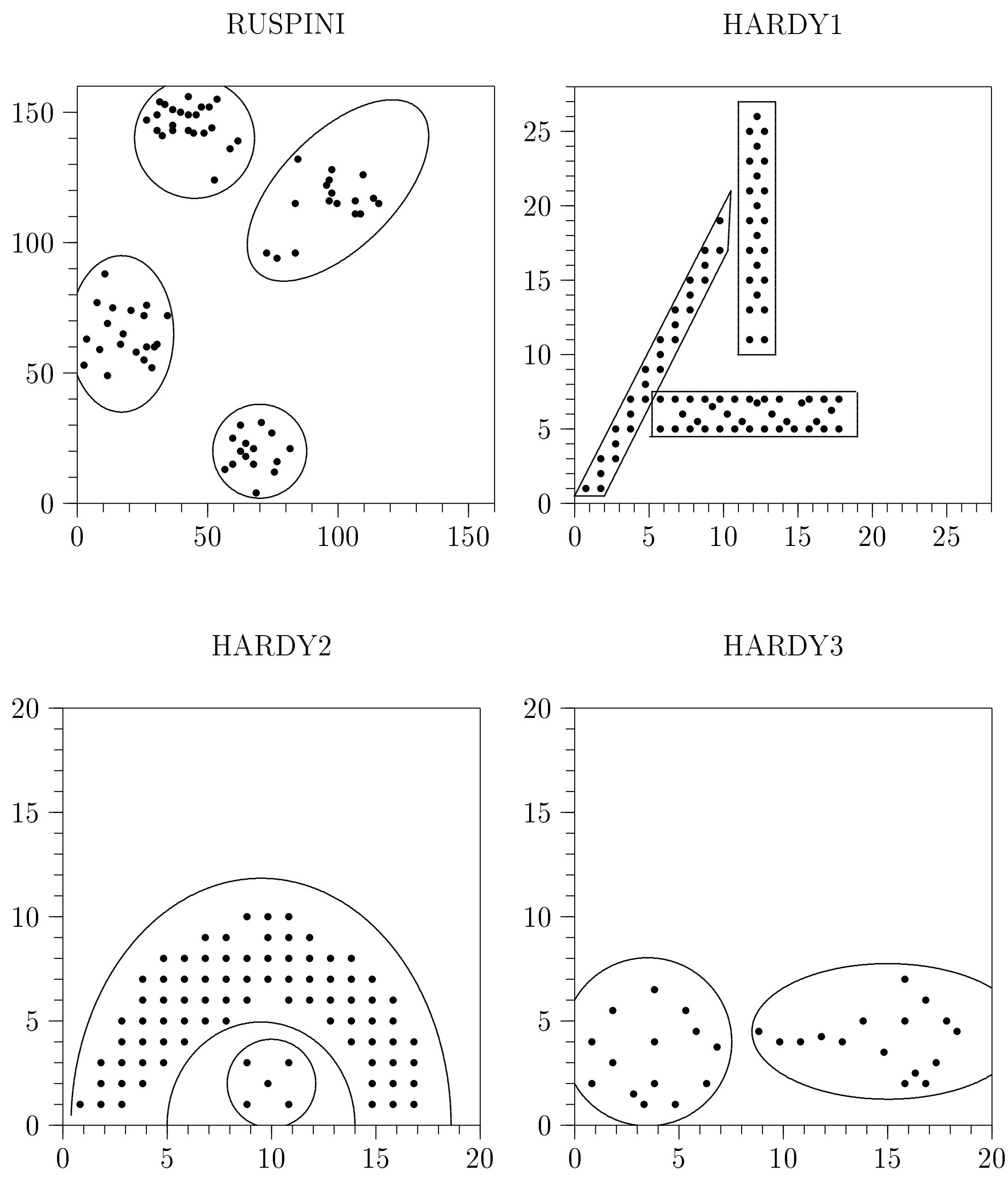


Figure 2: Four artificial data sets with given classes.

	RUSPINI	HARDY1	HARDY2	HARDY3
Histogramm	0.04	0.01	0.00	0.03
Rectangle	0.04	0.01	0.00	0.03
Gauss	0.00	0.02	0.00	0.00
Epanechnikov	0.05	0.01	0.00	0.00
Nearest neighbour	0.00	0.01	0.00	0.03
Nearest neighbour (mod.)	0.00	0.01	0.00	0.00
Mahalanobis	0.00	0.00	0.06	0.00
Zonoid	0.27	0.10	0.05	0.2
Hypervolume	0.00	0.00	0.05	0.00
Zonoid-Mahalanobis	0.00	0.00	0.05	0.00

Table 3: Leaving-one-out error rates of various classification rules applied to the data in Figure 2.

HARDY1 and HARDY3, it is clear that the apparent error rate of the zonoid rule amounts to zero, as for each class the zonoid depth vanishes outside the convex hull of the data. However, when calculating the leaving-one-out error rate, due to the same fact each point that is extremal in the convex hull of some class is not assigned to that class.

Thus, by the zonoid classification rule 27 % of the RUSPINI data are misclassified, while all other classification rules obtain *loo* rates below 5 %. Also the rules based on density estimation show relatively high *loo* rates when applied to the RUSPINI data; this may be due to an unfavourable choice of bandwidth. In the data sets HARDY1 and HARDY3 a few misclassifications are obtained. They concern objects which 'connect' two classes.

The data set HARDY2 differs from the other three sets in that one of the two classes is contained in the convex hull of the other class. This classification obviously cannot be identified by a rule which is based on a convex depth. Here, the rules based on density estimation or on nearest neighbours outperform the depth based rules. The latter show *loo* rates of 5 to 6 %.

We conclude that the zonoid-Mahalanobis classification rule appears to be a good alternative to the existing rules, provided the convex hulls of the groups do not intersect.

Compared with the commonly employed rules which are based on density estimators, the new rule avoids the – often problematic – choice of bandwidth.

The zonoid-Mahalanobis classification is practical since the zonoid depth of a point with respect to each of the given classes can be efficiently calculated. This approach is not feasible with the halfspace depth.

The idea to employ the zonoid depth also for dividing a given data set into classes ('clusters') seems to suggest itself. But this involves the calculation of many zonoid regions and/or their volumes, which appears to be too costly in applications.

References

BAUFAYS, P. and RASSON, J. (1985). A new geometric discriminant rule. *Computational Statistical Quarterly* **2**, 15–30.

DONOHO, D. L. and GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics* **20**, 1803–1827.

DYCKERHOFF, R. (2000). Computing zonoid trimmed regions of bivariate data sets. In J. Bethlehem and P. van der Heijden, eds., *COMPSTAT 2000. Proceedings in Computational Statistics*, 295–300. Physica-Verlag, Heidelberg.

DYCKERHOFF, R. (2002). Datentiefe: Begriff, Berechnung, Tests. Mimeo, Fakultät für Wirtschafts-und Sozialwissenschaften, Universität zu Köln.

DYCKERHOFF, R. (2004). Data depths satisfying the projection property. *Allgemeines Statistisches Archiv* **88**, 163–190.

DYCKERHOFF, R., KOSHEVOY, G. and MOSLER, K. (1996). Zonoid data depth: Theory and computation. In A. Pratt, ed., *COMPSTAT 1996. Proceedings in Computational Statistics*, 235–240, Heidelberg. Physica-Verlag.

HAND, D. (1981). *Discrimination and Classification.* Wiley, Chichester.

HARDY, A. (1991). On tests concerning the existence of a classification. *Belgian Journal of Operations Research, Statistics and Computer Sciences* **31**, 111–126.

HARDY, A. (1994). An examination of procedures for determining the number of clusters in a data set 169–177.

HARDY, A. (1996). On the number of clusters. *Computational Statistics and Data Analysis* **23**, 83–96.

HOBERG, R. (2003). *Clusteranalyse, Klassifikation und Datentiefe.* Josef Eul, Lohmar, Köln.

KOSHEVOY, G. (2002). The Tukey depth characterizes the atomic measure. *Journal of Multivariate Analysis* **83**, 360–364.

Koshevoy, G. (200x). Projections of lift zonoids, the Oja depth and the Tukey depth. *Annals of Statistics* . To appear.

Koshevoy, G. and Mosler, K. (1997). Zonoid trimming for multivariate distributions. *Annals of Statistics* **25**, 1998–2017.

Lachenbruch, P. A. (1968). Estimation of Error Rates in Discriminant Analysis. *Technometrics* **10**, 1–11.

Lachenbruch, P. A. (1975). *Discriminant Analysis.* Hafner, New York.

Liu, R. (1990). On a notion of data depth based on random simplices. *Annals of Statistics* **18**, 405–414.

Mosler, K. (2002). *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach.* Springer, New York.

Mosler, K. (2003). Central regions and dependency. *Methodology and Computing in Applied Probability* **5**, 5–21.

Rousseeuw, P. J. and Ruts, I. (1996). Algorithm AS 307. Bivariate location depth. *Applied Statistics* **45**, 516–526.

Rousseeuw, P. J. and Struyf, A. (1998). Computing location depth and regression depth in higher dimensions. *Statistics and Computing* **8**, 193–203.

Ruspini, E. H. (1970). Numerical methods for fuzzy clustering. *Information Sciences* **2**, 319–350.

Struyf, A. and Rousseeuw, P. J. (1999). Halfspace depth and regression depth characterize the empirical distribution. *Journal of Multivariate Analysis* **69**, 135–153.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proceedings National Academy of Sciences* 1423–1426.

Zuo, Y. and Serfling, R. (2000a). General notions of statistical depth function. *Annals of Statistics* **28**, 461–482.

Zuo, Y. and Serfling, R. (2000b). Nonparametric notions of multivariate "scatter measure" and "more scattered" based on statistical depth functions. *Journal of Multivariate Analysis* **75**, 62–78.

Karl Mosler (corresponding author), Richard Hoberg
Universität zu Köln
Albertus-Magnus-Platz
50923 Köln, Germany
mosler@statistik.uni-koeln.de