DIMACS Workshop on Data Depth, May 14–16, 2003

Data analysis by zonoid depth

Karl Mosler
Universität zu Köln
mosler@statistik.uni-koeln.de

1. Zonoid regions and zonoid depth:
   definitions,
   examples,
   general depth properties.
2. Special properties of the zonoid depth:
   convex, continuous on the point, dilation increasing, projection infimum, uniqueness of the distribution, Laws of Large Numbers.
3. Computation
4. Applications:
   - comparative dispersion,
   - comparative dependence,
   - point classification,
   - outlier identification.
5. Remarks and discussion.

1 Zonoid regions and zonoid depth

Let $\mathcal{P}_1$ denote the set of probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$ that have finite first moments.

Given a probability measure in $\mathcal{P}_1$ we define a family of central regions as follows.

Definition 1 (Zoonoid trimmed regions) .
For $P \in \mathcal{P}_1$, $0 < \alpha \leq 1$, *call*

$$D_\alpha(P) = \Bigg\{ \int_{\mathbb{R}^d} \boldsymbol{x}\, g(\boldsymbol{x})\, P(d\boldsymbol{x}) \quad : \quad g : \mathbb{R}^d \to \left[0, \frac{1}{\alpha}\right] \text{ measurable and } \int_{\mathbb{R}^d} g(\boldsymbol{x})\, P(d\boldsymbol{x}) = 1 \Bigg\}$$

the zonoid α-trimmed region *of* P. *For* $\alpha = 0$ *define*

$$D_0(P) = \mathbb{R}^d \,.$$

- D_α is convex, $\quad 0 \leq \alpha \leq 1$,
- $\bigcup_{\alpha > 0} D_\alpha =$ convex hull of the support of P,
- $D_1(P) = \{\boldsymbol{\epsilon}(P)\}$,
- The family of the D_α is nested,

$$D_{\alpha_2}(P) \subset D_{\alpha_1}(P) \quad \text{if } \ 0 \leq \alpha_1 < \alpha_2 \leq 1 \,.$$

Definition 2 (Zonoid depth of data) .
Let $\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$. *The number*

$$d_Z(\boldsymbol{y}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sup\{\alpha : \boldsymbol{y} \in D_\alpha(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)\}$$

is called the zonoid depth *of* $\boldsymbol{y}$ *with respect to* $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$.

(Note that $\sup \emptyset = 0$.*)*

Definition 3 (Zoonoid depth of a probability distribution)
For $P \in \mathcal{P}_1$ holds:

$$d_Z(\boldsymbol{y}|P) = \sup\{\alpha \,:\, \boldsymbol{y} \in D_\alpha(P)\}\,, \quad \boldsymbol{y} \in \mathbb{R}^d\,,$$

Then

$$D_\alpha(P) = \{\boldsymbol{y} \in \mathbb{R}^d : d_Z(\boldsymbol{y}|P) \geq \alpha\}\,, \quad 0 \leq \alpha \leq 1\,,$$

$$D_\alpha(P) = \left\{\boldsymbol{y} \in \mathbb{R}^d : gP \in \mathcal{A}(P, \boldsymbol{y}), ||g||_\infty \leq \frac{1}{\alpha}\right\}\,, \quad 0 < \alpha \leq 1\,.$$

where

$$\mathcal{A}(P, \boldsymbol{y}) = \{\nu \in \mathcal{P}_1 \,:\, \nu \text{ is } P\text{-continuous and } \boldsymbol{\epsilon}(\nu) = \boldsymbol{y}\,\}.$$

The **zoonoid depth** d satisfies the following general restrictions:

Affine invariance. For a regular $d \times d$-matrix $\boldsymbol{A}$ and $\boldsymbol{b} \in \mathbb{R}^d$ it holds

$$d(\boldsymbol{y}\boldsymbol{A}+\boldsymbol{b}|\boldsymbol{x}_1\boldsymbol{A}+\boldsymbol{b},\ldots,\boldsymbol{x}_n\boldsymbol{A}+\boldsymbol{b}) = d(\boldsymbol{y}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\,.$$

Vanishing at infinity. For any $(\boldsymbol{y}_k)_{k\in\mathbb{N}}$ with $\lim\limits_{k\to\infty}||\boldsymbol{y}_k|| = \infty$ it holds

$$\lim_{k\to\infty} d(\boldsymbol{y}_k|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = 0\,.$$

Monotone on rays. For each $\boldsymbol{y}$ of maximal depth and each $\boldsymbol{r} \in S^{d-1}$, the function

$$\lambda \mapsto d(\boldsymbol{y}+\lambda\boldsymbol{r}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\,, \quad \lambda \in \mathbb{R}\,,$$

is monotone decreasing.

Upper semicontinuity. For each $\alpha > 0$ the set

$$D_\alpha(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) := \left\{\boldsymbol{y}\in\mathbb{R}^d \,|\, d(\boldsymbol{y}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \geq \alpha\right\}$$

is closed.

Equivalenty, the **zonoid regions** are

- **affine equivariant,**
- **bounded,**
- **starshaped with respect to deepest points,**
- **closed.**

2 Special properties of the zonoid depth

- **Convexity:** $D_\alpha(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ is convex, $\quad 0 \le \alpha \le 1$,
- **Unique maximum:**

$$d_Z(\boldsymbol{y}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = 1 \quad \Leftrightarrow \quad \boldsymbol{y} = \overline{\boldsymbol{x}}\,,$$

- There holds

$$d_Z(\boldsymbol{y}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \begin{cases} = 0 & \text{if } \boldsymbol{y} \notin conv(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n), \\ \ge \frac{1}{n} & \text{if } \boldsymbol{y} \in conv(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)\,. \end{cases}$$

- **Continuous on the point:**
The function

$$\boldsymbol{y} \mapsto d_Z(\boldsymbol{y}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$$

is continuous on $\boldsymbol{y} \in conv\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$.

- **Continuous on the data:**
The function

$$(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \mapsto d_Z(\boldsymbol{y}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$$

is continuous at any $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$
for which $\boldsymbol{y}$ is in the relative interior of $conv\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$.

- **Increasing on dilation:** If $(\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$ is a dilation of $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ then

$$d_Z(\boldsymbol{y}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \le d_Z(\boldsymbol{y}|\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)\,.$$

- **Projection infimum:** For $P \in \mathcal{P}_1$ and $\boldsymbol{x}_1, \dots \boldsymbol{x}_n \in \mathbb{R}^d$, it holds

$$d_Z(\boldsymbol{y}|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \inf_{\boldsymbol{p} \in S^{d-1}} d_Z(\langle \boldsymbol{p}, \boldsymbol{y}\rangle | \langle \boldsymbol{p}, \boldsymbol{x}_1\rangle, \dots, \langle \boldsymbol{p}, \boldsymbol{x}_n\rangle)\,.$$

Laws of Large Numbers

Let P_n denote the empirical distribution from iid sampling of P. There holds:

- **Uniform LLN for zoonoid regions**

$$\lim_{n\to\infty} \sup_{\alpha_1 \le \alpha \le 1} |D_\alpha(P_n) - D_\alpha(P)| = 0\,, \quad a.s., \quad 0 < \alpha_1 \le 1\,.$$

- **LLN for zonoid depth**

$$\lim_{n\to\infty} |d(y|P_n) - d(y|P)| = 0\,, \quad y \in \mathrm{conv}(P)\,, \quad a.s.$$

- **Uniform LLN for zonoid depth**
  If the zonoid depth w.r.t. P is continuous on $\boldsymbol{y} \in \mathbb{R}$ the convergence is uniform (Dyckerhoff, 2003):

$$\lim_{n\to\infty} \sup_{y\in\mathbb{R}} |d(y|P_n) - d(y|P)| = 0\,, \quad a.s.$$

  In particular, if P has an L-density, the zonoid depth w.r.t. P is continuous on $\boldsymbol{y} \in \mathbb{R}$.

3 Computation

- **Computation of the depth of a single point**
- **Computation of a depth contour**
- **Robustness**

Based on the projection property, the depth can be numerically approximated by choosing a number of directions and calculating the pertaining univariate depths and their minimum over all directions.

4 Applications

4.1 Comparative dispersion

Definition 4 (Central regions order (=Lift zonoid order))
.
Let P, $Q \in \mathcal{P}_1$. Then

$$P \preceq_{lz} Q \quad \Leftrightarrow \quad D_\alpha(P) \subset D_\alpha(Q) \quad \textit{for all} \ \ \alpha \in [0,1]\,.$$

For $\boldsymbol{X} \in \mathcal{P}_1$ holds

$$\mathrm{E}(\boldsymbol{X}) \preceq_{lz} \boldsymbol{X}\,,$$
$$\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}) \preceq_{lz} \alpha(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})) \quad \text{if} \ \ \alpha > 1\,.$$

The order is implied by the convex (=dilation) order and implies equal expectations,

$$\boldsymbol{X} \preceq_{conv} \boldsymbol{Y} \quad \Rightarrow \quad \boldsymbol{X} \preceq_{lz} \boldsymbol{Y} \quad \Rightarrow \quad \mathrm{E}(\boldsymbol{X}) = \mathrm{E}(\boldsymbol{Y})\,.$$

The preorder does not indicate differences in location but in scale.
It compares distributions with respect to some sort of dispersion, *viz.* the size and shape of zonoid central regions.

In a family of elliptical distributions the lift zonoid order corresponds to an ordering of parameter matrices:
Let $\boldsymbol{X} \sim \mathcal{E}(\boldsymbol{a}, \boldsymbol{R}, \psi)$ and $\boldsymbol{Y} \sim \mathcal{E}(\boldsymbol{b}, \boldsymbol{S}, \psi)$. Then $\boldsymbol{X} \preceq_{lz} \boldsymbol{Y}$ iff

$$\boldsymbol{a} = \boldsymbol{b} \ \text{ and } \ \boldsymbol{S} - \boldsymbol{R} \ \text{ is positive semidefinite}\,.$$

Particularly, two normal variables are lift-zonoid ordered iff the means are equal and the covariance matrices are Loewner ordered.

4.2 Comparative dependency

Definition 5 (Central regions dependence order) .
Let $\mathcal{D}$ be a family of central regions, e.g. zonoid regions.
Consider two d-variate random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$
which have the same univariate marginals.
Define $\boldsymbol{X} \preceq_{\mathcal{D}} \boldsymbol{Y}$ if

$$\mathrm{vol}_d D_\alpha(F_{\boldsymbol{X}}) \geq \mathrm{vol}_d D_\alpha(F_{\boldsymbol{Y}}) \quad \textit{holds for every } \alpha\,.$$

$\preceq_{\mathcal{D}}$ *is called the* zonoid regions dependence order.

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are elliptically distributed,
$\boldsymbol{X} \sim \mathcal{E}(P, \boldsymbol{R}, \psi)$, $\boldsymbol{Y} \sim \mathcal{E}(P, \boldsymbol{S}, \psi)$, then

$$\boldsymbol{X} \preceq_{\mathcal{D}} \boldsymbol{Y} \quad \Leftrightarrow \quad |\det \boldsymbol{R}| \;\geq\; |\det \boldsymbol{S}|\,.$$

4.3 Point classification

(Joint work with Richard Hoberg)

Classification problem:
A multivariate data set C is partioned into given classes

$$C_1, \ldots, C_k\,, \qquad C_j = \{x_{j1}, \ldots, x_{jn_j}\}\,.$$

An additional data point y has to be assigned to the class to which, in some sense, it 'fits best'.

Depth classification rule:
Any depth function d provides a rule

$$classd(y) = argmin_j d(y|C_j)$$

which assigns y to that class C_j in which y is deepest.

Centroid classification rule:
Use the *Euclidean depth*

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

Mahalanobis classification rule:
Use the *Mahalanobis depth*

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

with Σ_j denoting the covariance matrix of group C_j.

With the Euclidean (resp. Mahalanobis) depth the distance from class C_j is measured in a circularly (resp. elliptically) symmetric way.
With other notions of data depth eventual asymmetries of the data can be reflected.

Problem: Most known depths vanish outside the convex hull of the set C_j.

By this, a point y lying outside the convex hulls of all classes cannot be classified by the depth.

Zonoid-Mahalanobis depth:

$$d_{ZMah}(y|C_j) = \max\left\{d_Z(y|C_j), \frac{1}{\max_j n_j} d_{Mah}(y|C_j)\right\} .$$

- This maximum is a depth and everywhere positive.
- The depth equals the zonoid depth inside the convex hull of C_j
  and is a multiple of the Mahalanobis depth outside.

Comparison of alternative rules:
The *zonoid-Mahalanobis rule* has been applied to several small benchmark data sets from the literature

and compared with known classification rules such as nearest neighbour, hypervolume, histogram and other rules.

The classification rules have been evaluated with respect to their *apparent error rates* as well as to their *leave-one-out error rates.*

In the result, the zonoid-Mahalanobis classification rule comes out 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 based on density estimators, the new rule avoids the – often problematic – choice of bandwidth.

4.4 Outlier identification

(Joint work with Katharina Cramer)

The problem of identifying an outlier is a problem of classification into two classes..

Classical procedure:
Assume independent data, estimate the mean vector $\boldsymbol{u}$ and the covariance matrix Σ of the regular process by $\hat{\boldsymbol{u}}$ and $\hat{\Sigma}$ and, by this, a family of ellipsoids around $\hat{\boldsymbol{u}}$ such that

- **(Heuristic approach)**
  k points lie outside the ellipsoid and are identified as outliers (where k is given) (e.g. Hadi, 1992) or
- **(Model based approach)**
  the ellipsoid has probability $1 - p$ under the regular distribution (where p is given) (Becker and Gather, 1999).

Criticism:

- Elliptical symmetry is not necessarily given,
- tail behavior may be non-normal.

Depth based procedure:
The complement of a depth based central region with properly chosen parameter α serves as an **outlying region** in the sense of Becker and Gather (1999). Two different approaches to depth trimming:

(a) Trimming with some given percentage p of the data:
$p\%$ of the data are *per definitionem* regarded as outliers.

(b) Trimming with some given α: α corresponds to the p-quantile of some chosen reference model (e.g. normal variables, $p = 0.99$).

"Practical advice":
(1) If some prior information on the slope and the tail behavior of the distribution (not necessarily symmetric) of the regular process is available,
choose a probability model and select p (or α).

(2) If the regular process is not specified at all,
use a k-outlier approach, that is, in a sequential way identify k outliers, $k = 1, 2, 3, \cdots,$
and judge the results by some model evaluation criterion.

Additional problems to be solved: masking and swamping.

5 Remarks and discussion