Zonoid Data Depth: Theory and Computation

Rainer Dyckerhoff[1], Gleb Koshevoy[2] and Karl Mosler[1]

[1] Seminar für Wirtschafts- und Sozialstatistik, Universität zu Köln, 50923 Köln, Germany
[2] Central Institute of Mathematics and Economics, Russian Academy of Science, Krasikova 32, Moscow 117418, Russia

Abstract. A new notion of data depth in d-space is presented, called the zonoid data depth. It is affine equivariant and has useful continuity and monotonicity properties. An efficient algorithm is developed that calculates the depth of a given point with respect to a d-variate empirical distribution.

1 Data Depth

Data depth is a measure of centrality by which multivariate data can be ordered. Given a cloud of data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ in d-space, data depth measures how central an additional point $\mathbf{y}$ is situated with respect to the $\mathbf{x}_i$. This measure serves as the base of rank tests and robust procedures.

Every notion of data depth should be affine equivariant, which means that, if $\mathbf{y}$ and the $\mathbf{x}_i$ are subject to the same affine transformation, the two resulting depths are the same. Various such notions have been proposed by Mahalanobis (1936), Tukey (1975), Liu (1990), and others. See Liu and Singh (1993) and Rousseeuw and Leroy (1987, ch. 7).

Here we introduce a new definition, zonoid data depth, which has particularly nice properties. We present an efficient algorithm that calculates the data depth of a given point in $\mathbb{R}^d$ with respect to a given empirical distribution of d-variate data. It is monotone and continuous on $\mathbf{y}$, zero at infinity, and unity at the sample mean $\overline{\mathbf{x}}$. Moreover it is continuous on $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and monotone on dilations of $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Definition 1. Let $\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$. The *zonoid data depth* of $\mathbf{y}$ with respect to $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is

$$\operatorname{depth}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sup\{\alpha : \mathbf{y} \in D_\alpha(\mathbf{x}_1, \ldots, \mathbf{x}_n)\} \tag{1}$$

where

$$D_\alpha(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \left\{ \sum_{i=1}^n \lambda_i \mathbf{x}_i : \sum_{i=1}^n \lambda_i = 1, 0 \le \lambda_i, \alpha\lambda_i \le \frac{1}{n} \text{ for all } i \right\}. \tag{2}$$

$D_\alpha(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is the *α-trimmed region* of the empirical distribution generated by the $\mathbf{x}_i$ (Koshevoy and Mosler 1995), and we use the convention $\sup \emptyset = 0$. It is clear, that D_α is convex for every α. For $0 \leq \alpha \leq \frac{1}{n}$, D_α is the convex hull of the data. D_1 is a singleton containing their mean $\overline{\mathbf{x}}$. Moreover, D_α is monotone in the sense that $D_\alpha \subset D_\beta$ if $\alpha > \beta$.

Figure 1 exhibits several zonoid trimmed regions for a sample of 10 data points in two-space.

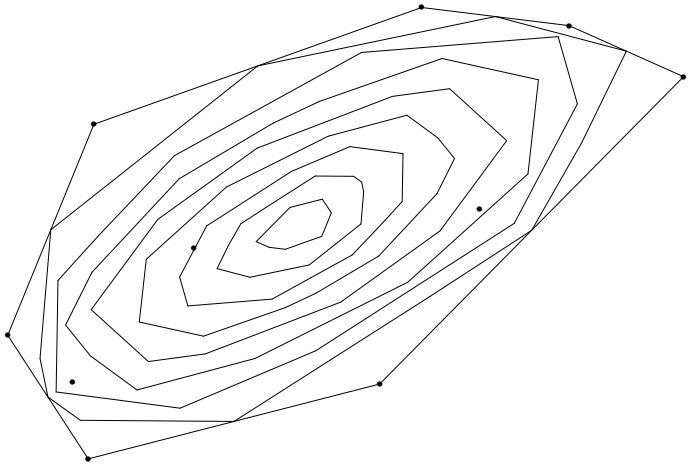

Figure 1. Zonoid trimming regions when $n = 10$ and $d = 2$. The trimming regions are drawn for $\alpha = 0.1, 0.2, \ldots 0.9$.

Zonoid data depth differs from the existing notions: Tukey's depth (Tukey 1975), simplicial depth (Liu 1990), majority depth (see Liu and Singh 1993). Our notion has many properties in general which these notions have under some restrictions only; see e.g. Liu and Singh (1993) for properties of Tukey's, simplicial and majority depths. Koshevoy and Mosler (1995) demonstrate that the zonoid data depth equals twice Tukey's data depth of a properly transformed distribution.

In Section 2 a theorem is given that collects the main continuity and monotonicity properties of zonoid data depth. Section 3 presents the algorithm.

2 Properties of Zonoid Data Depth

The depth of $\mathbf{y}$ equals zero if $\mathbf{y}$ lies outside the convex hull $\text{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$; it equals one if $\mathbf{y}$ is the arithmetic mean. From infinity to the mean the data depth increases monotonically and is continuous on $\mathbf{y} \in \text{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. If

$\mathbf{y}$ and the $\mathbf{x}_i$ are transformed by the same affine transform then the depth remains the same. Further, inside the convex hull, the data depth is continuous on the $\mathbf{x}_i$; it increases if the distribution of the $\mathbf{x}_i$ becomes more variable in terms of a dilation. More precisely, the main properties of the zonoid data depth are summarized in the following theorem.

Theorem 2.

(i) *(Zero at infinity)* $\sup_{\|\mathbf{y}\| \geq M} \operatorname{depth}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_n) \to 0$ *as* $M \to \infty$.
(ii) *(Continuous on* $\mathbf{y}$*) At every* $\mathbf{y}^0 \in \operatorname{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$*, the function* $\mathbf{y} \mapsto \operatorname{depth}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_n)$ *is continuous.*
(iii) *(Continuous on the* $\mathbf{x}_i$*) At every* $(\mathbf{x}_1^0, \ldots, \mathbf{x}_n^0) \in \operatorname{int}\operatorname{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$*, the function* $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \mapsto \operatorname{depth}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_n)$ *is continuous.*
(iv) *(Unity only at expectation) If* $\mathbf{y} \neq \overline{\mathbf{x}}$ *then* $\operatorname{depth}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_n) < 1 = \operatorname{depth}(\overline{\mathbf{x}}|\mathbf{x}_1, \ldots, \mathbf{x}_n)$.
(v) *(Monotone on* $\mathbf{x}$*) For every* $\mathbf{y} \in \mathbb{R}^d$*,* $\operatorname{depth}(c\mathbf{y} + \overline{\mathbf{x}}|\mathbf{x}_1, \ldots, \mathbf{x}_n)$ *is monotone decreasing on* $c \geq 0$.
(vi) *(Affine equivariant) For any given matrix* A *and vector* b*,* $\operatorname{depth}(A\mathbf{y} + b|A\mathbf{x}_1 + b, \ldots, A\mathbf{x}_n + b) = \operatorname{depth}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_n)$.
(vii) *(Monotone on dilation)* $\operatorname{depth}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_n) \leq \operatorname{depth}(\mathbf{y}|\mathbf{z}_1, \ldots, \mathbf{z}_n)$ *if* $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ *is a dilation of* $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Koshevoy and Mosler (1995) have defined the zonoid data depth in the following, more general context: Let $\mathbf{y} \in \mathbb{R}^d$ and μ be a d-variate probability distribution that has a finite expectation vector $E(\mu)$. The zonoid data depth of $\mathbf{y}$ with respect to μ is defined by

$$\operatorname{depth}_\mu(\mathbf{y}) = \sup\{\alpha \,:\, \mathbf{y} \in D_\alpha(\mu)\}. \tag{3}$$

Here $D_\alpha(\mu)$ denotes the zonoid α-trimmed region of μ (Koshevoy and Mosler 1995),

$$D_\alpha(\mu) = \Big\{ \int_{\mathbb{R}^d} x g(x)\, d\mu(x) \,:\, g : \mathbb{R}^d \to [0, \frac{1}{\alpha}] \text{ measurable and } \int_{\mathbb{R}^d} g(x)\, d\mu(x) = 1 \Big\}.$$

If μ is an empirical distribution generated by $\mathbf{x}_1, \ldots, \mathbf{x}_n$, it can be easily seen that the Definition (3) becomes (1). The theorem thus follows from Koshevoy and Mosler (1995, Th. 8.1).

3 Computation

We consider the data matrix

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n),$$

whose columns are the vectors $\mathbf{x}_i$, $i = 1, \ldots, n$, and denote $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$, $\mathbf{1} = (1, \ldots, 1)'$, $\mathbf{0} = (0, \ldots, 0)'$. The prime indicates the transpose.

The data depth (1) of a point $\mathbf{y}$ in $\mathbb{R}^d$ can be computed as follows.

$$\left.\begin{array}{ll} \text{Minimize} & \gamma \\ \text{subject to} & \mathbf{X}\boldsymbol{\lambda} = \mathbf{y} \\ & \boldsymbol{\lambda}'\mathbf{1} = 1 \\ & \gamma\mathbf{1} - \boldsymbol{\lambda} \geq \mathbf{0}, \quad \boldsymbol{\lambda} \geq \mathbf{0} \end{array}\right\} \qquad \text{(LP)}$$

(LP) is a linear program in the real variables $\lambda_1, \ldots, \lambda_n$ and γ. If γ^* is the optimal value of the objective then

$$\text{depth}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{1}{n\gamma^*}\,.$$

If (LP) has no feasible solution, then it is clear that $\mathbf{y} \notin \text{conv}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

Although the above LP can be easily solved by the standard simplex method when n is small, a more subtle approach is needed for large-scale problems. Our algorithm exploits the special structure of the set of constraints by a Dantzig-Wolfe decomposition. (LP) can be written

$$\left.\begin{array}{ll} \text{Minimize} & \gamma \\ \text{subject to} & \mathbf{X}\boldsymbol{\lambda} = \mathbf{y} \\ & (\lambda_1, \ldots, \lambda_n, \gamma)' \in S \end{array}\right\} \qquad \text{(LP}')$$

where

$$S = \{(\lambda_1, \ldots, \lambda_n, \gamma)' \in \mathbb{R}^{n+1} \,:\, \sum_{i=1}^{n} \lambda_i = 1, 0 \leq \lambda_i \leq \gamma \leq 1, \text{ for all } i\}\,.$$

Because S is a bounded polyhedral set, any point in S is a convex combination of the extreme points. Fortunately, the extreme points of S are explicitly known.

Proposition 3. *The set V of extreme points of S is given by*

$$V = \left\{ \frac{1}{|I|}(\boldsymbol{\delta}_I, 1)' \,:\, \emptyset \neq I \subset \{1, \ldots, n\} \right\}.$$

Here

$$\boldsymbol{\delta}_I = (\delta_I(1), \delta_I(2), \ldots, \delta_I(n)), \qquad \delta_I(k) = \begin{cases} 1, \textit{ if } k \in I, \\ 0, \textit{ if } k \notin I. \end{cases}$$

By Proposition 3, whose proof is left to the reader, (LP′) can be decomposed as follows. The *master problem*, with variables β_I, $\emptyset \neq I \subset \{1, \ldots, n\}$, is

$$\left.\begin{aligned} \text{Minimize} \quad & \sum_I \frac{1}{|I|}\beta_I \\ \text{subject to} \quad & \sum_I \frac{1}{|I|}(\mathbf{X}\boldsymbol{\delta}_I')\beta_I = \mathbf{y} \\ & \sum_I \beta_I = 1 \\ & \beta_I \geq 0 \ \text{ for all } I \end{aligned}\right\} \qquad \text{(MP)}$$

In every simplex step of (MP) a new pivot column is selected by solving the *subproblem*

$$\left.\max_I \frac{1}{|I|}(\mathbf{w}\mathbf{X}\boldsymbol{\delta}_I' - 1) + \alpha \qquad \right\} \qquad \text{(SP)}$$

where $(\mathbf{w}, \alpha)$ is the vector of simplex multipliers of the master problem.

If the maximum objective of the subproblem is greater than zero and maximized at $I = I^*$, then the new pivot column for the master problem is calculated as

$$B^{-1}\left(\frac{1}{|I^*|}\boldsymbol{\delta}_{I^*}\mathbf{X}', 1\right)',$$

where B^{-1} is the basis inverse of the master problem. The pivot row for the simplex step is then determined by the usual minimal ratio test, and the tableau is updated. This process is continued until the maximum objective of the subproblem equals zero. Then the current solution of the master problem is optimal and the algorithm is stopped.

Summary of the algorithm

1. **Initialization.** Find a basic feasible solution of the system $\mathbf{X}\boldsymbol{\lambda} = \mathbf{y}$, $\boldsymbol{\lambda}'\mathbf{1} = 1$, $\boldsymbol{\lambda} \geq \mathbf{0}$, using the two-phase method. Then $\beta_{\{i\}} = \lambda_i$, $i = 1, \ldots, n$, $\beta_I = 0$, $|I| \geq 2$, is a basic feasible solution of the master problem. Initialize the revised simplex tableau for the master problem.
2. **Solution of the subproblem.**
   (a) Compute $\mathbf{w}\mathbf{X}$ and arrange the components in decreasing order. Let (i) be the index of the i-th largest component of $\mathbf{w}\mathbf{X}$.
   (b) Find k^* which maximizes

$$\frac{1}{k}\left(\sum_{i=1}^{k}(\mathbf{w}\mathbf{X})_{(i)} - 1\right), \quad k \in \{1, \ldots, n\}.$$

   (c) The maximum objective of the subproblem is given by

$$z^* = \frac{1}{k^*}\left(\sum_{i=1}^{k^*}(\mathbf{w}\mathbf{X})_{(i)} - 1\right) + \alpha$$

   and the maximum is achieved at $I^* = \{(1), (2), \ldots, (k^*)\}$.

(d) If $z^* = 0$ stop; the basic feasible solution of the last master step is optimal. Otherwise continue with the next step.

3. **Update of the master tableau.** Let

$$\mathbf{c} = B^{-1}\left(\frac{1}{|I^*|}\boldsymbol{\delta}_{I^*}\mathbf{X}', 1\right)'$$

where B^{-1} is the basis inverse of the master problem. Join the new pivot column $\binom{z^*}{\mathbf{c}}$ with the master tableau. Determine the pivot row for the simplex step by the usual minimal ratio test and update the master tableau. Continue with Step 2.

Further, the algorithm generates an increasing sequence of lower bounds on the data depth and a (not necessarily decreasing) sequence of upper bounds.

Table 1 summarizes some computation times. A sample of size n was drawn from a standard normal distribution. $\mathbf{y}$ was calculated as the arithmetic mean of the first ten points in the sample. The table shows the computation times in seconds on a 100 MHz Pentium™.

Table 1. Computation times [in seconds] of the algorithm

n \ d	2	3	4	5	10
1000	0.21	0.43	0.76	0.87	4.11
2000	0.54	1.09	1.75	2.36	8.73
4000	1.48	2.03	4.22	5.32	24.88
8000	3.35	6.81	10.76	16.20	71.07
16000	9.72	14.72	23.39	36.52	150.38

References

Koshevoy, G. and Mosler, K. (1995). Zonoid trimming for multivariate distributions. Mimeo.

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

Liu, R. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association* **88**, 252–260.

Mahalanobis, P.C. (1936). On the generalized distance in statistics. *Proceedings of the National Academy India* **12**, 49–55.

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

Tukey, J.W. (1975). Mathematics and picturing data. *Proceedings of the 1974 International Congress of Mathematicians, Vancouver* **2**, 523–531.