

# Similarities between location depth and regression depth

Mia Hubert\*, Peter J. Rousseeuw, and Stefan Van Aelst†

November 25, 1999

Department of Mathematics and Computer Science, U.I.A.,  
Universiteitsplein 1, B-2610 Antwerpen, Belgium.

## Abstract

The location depth of Tukey (1975) is a multivariate generalization of rank, and leads to a multivariate median known as the Tukey median. Recently, Rousseeuw and Hubert (1999a) introduced a notion of depth in the regression setting. It provides the ‘rank’ of any line (plane), rather than ranks of observations or residuals. In general, depth is an integer assigned to a candidate fit relative to a data set. Both in location and in regression, the depth of a fit can also be defined with regard to a population distribution. In this paper we explore the analogies between location and regression depth. We compute the depth functions at elliptical distributions. We compare the lower and upper bounds for the maximal depth. We also consider the robustness, the asymptotics and the computation of the deepest location and the deepest regression. Finally we introduce the notion of centrality, which is a more quantitative version of depth that leads to affine equivariant estimators of location and regression with 50% breakdown value.

---

\*Postdoctoral Fellow at the FWO, Belgium.

†Research Assistant with the FWO, Belgium.

# 1 Introduction

The notion of *location depth* was proposed by Tukey (1975) as a graphical tool to picture bivariate data clouds, and as a multivariate generalization of rank. The deepest point is a kind of multivariate median (Donoho and Gasko 1992), the depth of which reflects the degree of angular symmetry in the data.

Rousseeuw and Hubert (1999a) extended the notion of depth to the linear regression setting. For this they proposed a unifying concept of depth, from which both location depth and regression depth can be derived. In Section 2 we recall this general definition of depth, and explain the relation with location and regression depth. In the population case, we look at the depth functions of elliptical distributions.

Section 3 concerns the maximal location and regression depth. We compare the lower and upper bounds in both settings. Section 4 is devoted to the fit with maximal depth. We show that the robustness properties and the asymptotics of the deepest location and the deepest regression are very similar, both in the finite-sample and in the population case. Section 5 compares the computation time of the location and the regression depth, and of the deepest fits.

Finally, in Section 6 we introduce the new notion of *centrality*. Both in location and in regression, the most central fit estimator is an affine equivariant estimator that has a 50% breakdown value.

## 2 Location and regression depth

### 2.1 A general definition of depth

Rousseeuw and Hubert (1999a) proposed a general definition of the depth of a (candidate) fit  $\theta$  relative to a given data set  $Z_n$ .

**Definition 1.** The  $\text{depth}(\theta, Z_n)$  is the smallest number of observations of  $Z_n$  that would need to be removed in order to make  $\theta$  a nonfit.

In the population case, we ask how much probability mass needs to be removed. A particular depth function is thus equivalent to the definition of a nonfit (since nonfits are exactly those fits with zero depth). The definition of a nonfit will depend on the statistical framework.

## 2.2 Location and regression depth at finite samples

When estimating location, the data set  $X_n$  consists of  $n$  observations in  $\mathbb{R}^p$ . Any  $\theta \in \mathbb{R}^p$  is then a candidate fit for the center of the data.

**Definition 2 (Location).** We call  $\theta \in \mathbb{R}^p$  a nonfit for a given data set  $X_n \subset \mathbb{R}^p$  iff  $\theta$  lies outside the convex hull of  $X_n$ .

For  $p = 1$ , the convex hull is just the interval spanned by the data. Definition 2 is equivalent to saying that there exists an affine hyperplane through  $\theta$  with all observations strictly on one side and none on the other side. Tukey's location depth (1975), which we will denote by  $ldepth(\theta, X_n)$ , is defined as the smallest number of data points contained in a closed halfspace of which the boundary passes through  $\theta$ . Therefore, Definitions 1 and 2 immediately yield Tukey's location depth.

The **deepest location** is now defined by

$$T_l^*(X_n) = \operatorname{argmax}_{\theta} ldepth(\theta, X_n) \quad (1)$$

and often called the *Tukey median*. If there are several solutions to (1), their average (i.e. their center of gravity) is taken. Note that for  $p = 1$  the Tukey median becomes the sample median. For general  $p$ , the deepest location can thus be seen as a multivariate median.

In (multiple) regression the data set is of the form  $Z_n = \{(x_{i1}, \dots, x_{i,p-1}, y_i)^t; i = 1, \dots, n\} \subset \mathbb{R}^p$ . We denote the  $x$ -part of each data point by  $x_i = (x_{i1}, \dots, x_{i,p-1})^t \in \mathbb{R}^{p-1}$ . We now want to fit the  $y_i$  by

$$\theta_1 x_{i1} + \dots + \theta_{p-1} x_{i,p-1} + \theta_p$$

that is, by an affine hyperplane in  $\mathbb{R}^p$ . Here again  $\theta = (\theta_1, \dots, \theta_p)^t \in \mathbb{R}^p$ .

Following Rousseeuw and Hubert (1999a) we say

**Definition 3 (Regression).** A candidate fit  $\theta = (\theta_1, \dots, \theta_p)^t$  is called a nonfit to  $Z_n$  iff there exists an affine hyperplane  $V$  in  $x$ -space such that no  $x_i$  belongs to  $V$ , and such that  $r_i > 0$  for all  $x_i$  in one of its open halfspaces and  $r_i < 0$  for all  $x_i$  in the other open halfspace.

The regression depth of any hyperplane  $\theta$  now follows immediately from Definition 1. As in the location setup (1) we can now define the **deepest regression** estimator:

$$T_r^*(Z_n) = \operatorname{argmax}_{\theta} rdepth(\theta, Z_n).$$

If there are several  $\theta$  with that same rdepth, the average of those  $\theta$  is taken. Analogously to  $T_l^*$  we can see  $T_r^*$  as a kind of *regression median*. Note that  $T_r^*$  estimates the regression coefficients  $\theta_1, \dots, \theta_p$  without estimating any scale parameter, just like  $T_l^*$  which estimates location without estimating scatter.

Struyf and Rousseeuw (1999) have proved the following relation between the location and regression depth functions and the underlying empirical distribution.

**Theorem 1.** (a) The empirical distribution of any data set  $X_n \subset \mathbb{R}^p$  is uniquely determined by its halfspace depth function  $ldepth(\theta, X_n)$ .

(b) The empirical distribution of any data set  $Z_n \subset \mathbb{R}^p$  is uniquely determined by its regression depth function  $rdepth(\theta, Z_n)$ .

## 2.3 Location and regression depth at a population

**Definition 4 (Location).** At any distribution  $P$  on  $\mathbb{R}^p$  and at any  $\theta \in \mathbb{R}^p$

$$ldepth(\theta, P) = \inf_{\|u\|=1} P(H_{\theta,u})$$

where  $u$  is a unit vector in  $\mathbb{R}^p$  and  $H_{\theta,u} = \{x \in \mathbb{R}^p; u^t x \geq u^t \theta\}$  is a closed hyperplane.

In other words, the location depth of any  $\theta$  is defined as the smallest population mass of  $P$  contained in a closed halfspace with boundary through  $\theta$ .

Let us consider elliptical distributions  $P_{\mu,\Sigma}$  with density

$$f_{\mu,\Sigma}(x) = \frac{g((x - \mu)^t \Sigma^{-1} (x - \mu))}{\sqrt{\det(\Sigma)}} \quad (2)$$

with  $\mu \in \mathbb{R}^p$  and  $\Sigma$  a positive definite matrix of size  $p$ . We assume the function  $g$  to have a strictly negative derivative, so that  $P_{\mu,\Sigma}$  is unimodal. For spherical distributions  $P_{0,I_p}$  with marginal distribution  $P_{0,1}$  we find

$$ldepth(\theta, P_{0,I_p}) = 1 - P_{0,1}(\|\theta\|).$$

Since location depth is affine invariant, it follows that

$$ldepth(\theta, P_{\mu,\Sigma}) = ldepth(\Sigma^{-1/2}(\theta - \mu), P_{0,I_p}) = 1 - P_{0,1}(\|\Sigma^{-1/2}(\theta - \mu)\|)$$

where  $\Sigma^{-1/2}$  is the inverse of the symmetric square root of  $\Sigma$ .

The location depth function for the bivariate gaussian distribution  $P_{0,I_2} = N_2(0, I_2)$  is plotted in Figure 1. We see that the location depth function is more peaked than the bivariate distribution itself, and that the maximum is reached at  $\theta = (0, 0)^t$ . This phenomenon occurs at various other distributions. This is nicely illustrated in Rousseeuw and Ruts (1999) where the ldepth function of many population distributions is computed and plotted.

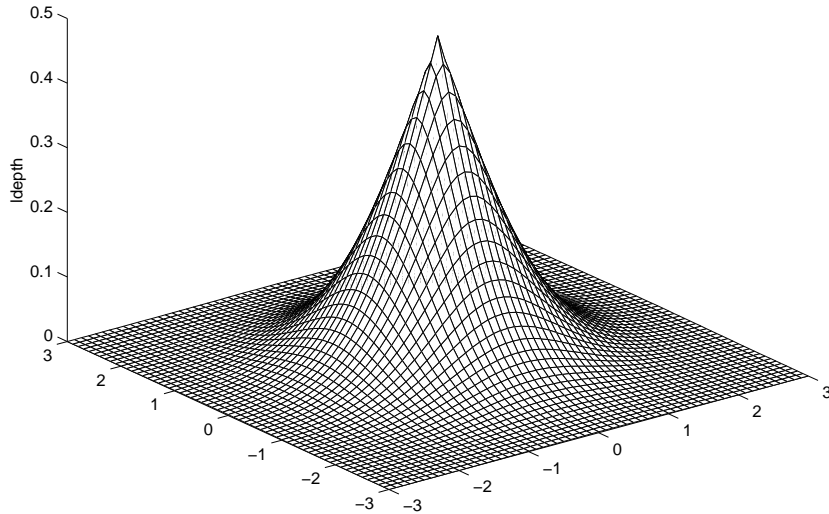


Figure 1: The location depth function at the bivariate gaussian distribution.

We can analogously define the rdepth of a fit with regard to a distribution  $H$ .

**Definition 5 (Regression).** Let  $H$  be the distribution of the random variable  $(x, y)$ . Then

$$rdepth(\theta, H) = \min_{u,v} \{H(y - (x^t, 1)\theta < 0 \text{ and } x^t u < v) + H(y - (x^t, 1)\theta > 0 \text{ and } x^t u > v)\}$$

where the minimum is over all unit vectors  $u \in \mathbb{R}^{p-1}$  and all  $v \in \mathbb{R}$  with  $H(x^t u = v) = 0$ .

Let us compute the regression depth for elliptical distributions  $H_{\mu,\Sigma}$  with density according to (2). Since regression depth is regression, scale and affine invariant (according to the definitions in Rousseeuw and Leroy 1987), it suffices to study the depth at spherical distributions  $H_{0,I_p}$ . From Van Aelst and Rousseeuw (1998) it follows that the minimal amount of probability mass that has to be removed to make  $\theta$  a nonfit is the probability mass passed when tilting the fit  $\theta$  around the intersection of  $\theta$  with  $T_r^*(H)$  until  $\theta$  becomes vertical. The direction in which to tilt  $\theta$  is such that  $\theta$  does not pass  $T_r^*(H)$ . Moreover, the

deepest regression is Fisher-consistent at spherical distributions (Van Aelst and Rousseeuw 1998). Therefore,  $T_r^*(H) = 0$ . We then obtain that the  $x$ -projection of the intersection of  $\theta$  with  $T_r^*(H)$  is given by  $x^t u_\theta = v_\theta$  with  $u_\theta = \theta_{sl}/\|\theta_{sl}\|$  and  $v_\theta = -\theta_p/\|\theta_{sl}\|$  where  $\theta_{sl} = (\theta_1, \dots, \theta_{p-1})^t$ . It follows that

$$\text{rdepth}(\theta, H_{0,I_p}) = H_{0,I_p}(y - (x^t, 1)\theta < 0 \text{ and } x^t u_\theta < v_\theta) + H_{0,I_p}(y - (x^t, 1)\theta > 0 \text{ and } x^t u_\theta > v_\theta). \quad (3)$$

Figure 2 shows the regression depth function for the bivariate gaussian distribution  $H_{0,I_2} = N_2(0, I)$ . Since for the bivariate gaussian distribution the probabilities  $H_{0,I_2}(y - (x, 1)\theta < 0 \text{ and } xu_\theta < v_\theta)$  and  $H_{0,I_2}(y - (x, 1)\theta > 0 \text{ and } xu_\theta > v_\theta)$  cannot be computed explicitly, we approximated them by means of numerical integration. The regression depth is a surface, with a sharp peak, which attains its maximum at  $(0, 0)^t$ . Note that this function is not symmetric in  $\theta_1$  (the slope) and  $\theta_2$  (the intercept), which clearly play a different role (whereas Figure 1 was symmetric).

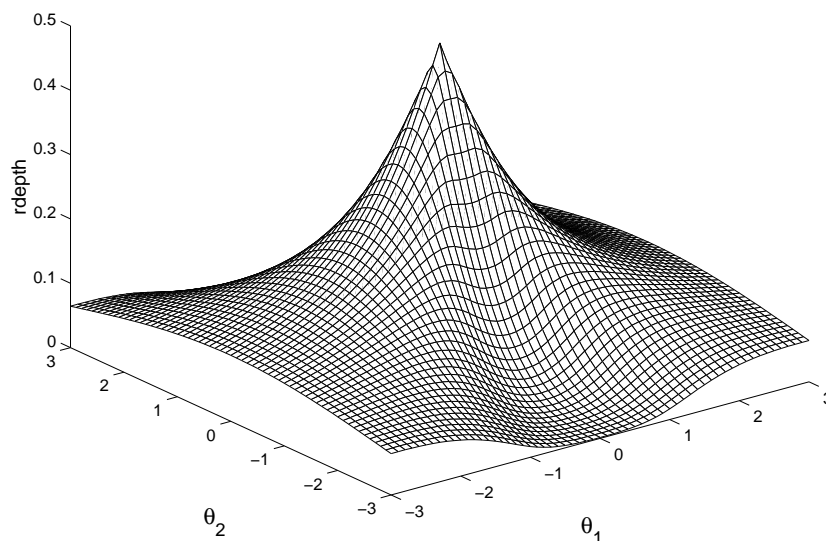


Figure 2: The regression depth function at the bivariate gaussian distribution.

### 3 The maximal location and regression depth

Following Donoho and Gasko (1992) we have

**Theorem 2 (Location).** (a) At any data set  $X_n \subset \mathbb{R}^p$  it holds that

$$\left\lceil \frac{n}{p+1} \right\rceil \leq \max_{\theta} ldepth(\theta, X_n) \leq n \quad (4)$$

where the ceiling  $\lceil \lambda \rceil$  is the smallest integer  $\geq \lambda$ .

(b) If  $X_n$  is in general position (i.e., no more than  $p$  observations lie in any  $(p-1)$ -dimensional affine subspace), then

$$\max_{\theta} ldepth(\theta, X_n) \leq \left\lceil \frac{n}{2} \right\rceil. \quad (5)$$

(c) For any distribution  $P$  on  $\mathbb{R}^p$  with a density it holds that

$$\frac{1}{p+1} \leq \max_{\theta} ldepth(\theta, P) \leq \frac{1}{2}.$$

(d) If  $P$  is angularly symmetric about some  $\tilde{\theta} \in \mathbb{R}^p$ , then

$$\max_{\theta} ldepth(\theta, P) = ldepth(\tilde{\theta}, P) = \frac{1}{2}.$$

Note that a distribution  $P$  on  $\mathbb{R}^p$  is said to be angularly symmetric about  $\theta$  iff  $P(\theta + A) = P(\theta - A)$  for any cone  $A$  emanating from the origin.

Actually, the maximal depth at a given data set  $X_n$  depends on its shape. The upper bound (5) is reached at highly symmetric data sets, whereas the lower bound in (4) is attained at very asymmetric data sets. This property can e.g. be used to explore the skewness of a multivariate distribution (Liu et al. 1999).

In Rousseeuw and Hubert (1999a), Hubert and Rousseeuw (1998), and Mizera (1998) it is shown that very similar properties hold for the maximal regression depth, as summarized in the following theorem.

**Theorem 3 (Regression).** (a) For any data set  $Z_n \subset \mathbb{R}^p$  it holds that

$$\left\lceil \frac{n}{p+1} \right\rceil \leq \max_{\theta} rdepth(\theta, Z_n) \leq n.$$

(b) If  $Z_n$  is in general position,

$$\max_{\theta} rdepth(\theta, Z_n) \leq \left\lceil \frac{n+p}{2} \right\rceil.$$

(c) For any distribution  $H$  on  $\mathbb{R}^p$  with a density, we have

$$\frac{1}{p+1} \leq \max_{\theta} rdepth(\theta, H) \leq \frac{1}{2}.$$

(d) If  $H$  has a strictly positive density function on  $\mathbb{R}^p$  such that  $\text{med}(y|x) = (x^t, 1)\tilde{\theta}$  for some  $\tilde{\theta} \in \mathbb{R}^p$ , then

$$\max_{\theta} rdepth(\theta, H) = rdepth(\tilde{\theta}, H) = \frac{1}{2}.$$

The lower bound  $1/(p+1)$  in (c) is reached when all probability mass of  $H$  is concentrated on the 'moment curve'  $\{(u, u^2, \dots, u^p); u > 0\}$ . For simple regression ( $p = 2$ ) the maximal regression depth can therefore be used as a test of linearity versus convexity/concavity (Rousseeuw, Van Aelst and Hubert, 1999).

## 4 The deepest location and the deepest regression

The deepest location  $T_l^*$  and the deepest regression  $T_r^*$  both appear to be highly robust estimators having similar properties. Let us first compare their breakdown values. Roughly speaking, the *breakdown value* of an estimator with regard to a data set  $X_n$  or a distribution  $P$  is the smallest fraction of  $X_n$  or  $P$  that needs to be replaced to carry the estimator arbitrarily far away. (For background on the breakdown value, see Rousseeuw and Leroy 1987).

**Theorem 4.** (a) (Donoho and Gasko, 1992) At any data set  $X_n \subset \mathbb{R}^p$ , it holds that

$$\varepsilon_n^*(T_l^*, X_n) \geq \frac{1}{n} \left\lceil \frac{n}{p+1} \right\rceil \approx \frac{1}{p+1}.$$

(b) (Rousseeuw and Hubert, 1999a) If the  $x_i$  are in general position,

$$\varepsilon_n^*(T_r^*, Z_n) \geq \frac{1}{n} \left( \left\lceil \frac{n}{p+1} \right\rceil - p + 1 \right) \approx \frac{1}{p+1}.$$

The breakdown value of the deepest location and regression can thus be as low as  $1/(p+1)$  at some peculiar situations. However, if the original data are drawn from the model, then the breakdown value converges almost surely to  $1/3$  in any dimension  $p$ .

**Theorem 5.** (a) (Donoho and Gasko, 1992) Let  $X_n$  be a sample of size  $n$  from an absolutely continuous angularly symmetric distribution on  $\mathbb{R}^p$  ( $p \geq 3$ ). Then

$$\varepsilon_n^*(T_l^*, X_n) \xrightarrow[n \rightarrow \infty]{a.s.} \frac{1}{3}.$$

(b) (Van Aelst et al., 1999) Let  $Z_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$  be a sample from a distribution  $H$  on  $\mathbb{R}^p$  ( $p \geq 3$ ) with a density, which satisfies  $\text{med}(y|x) = (x^t, 1)\tilde{\theta}$  for some  $\tilde{\theta}$ . Then

$$\varepsilon_n^*(T_r^*, Z_n) \xrightarrow[n \rightarrow \infty]{a.s.} \frac{1}{3}.$$

Moreover, we see that also in the population case the breakdown value of the deepest location and the deepest regression is  $1/3$ .

**Theorem 6.** (a) (Zhang, 1998) If  $P$  is angularly symmetric about some  $\tilde{\theta}$ ,

$$\varepsilon^*(T_l^*, P) = \frac{1}{3}.$$

(b) (Van Aelst and Rousseeuw, 1998) If  $H$  has a strictly positive density function on  $\mathbb{R}^p$  that satisfies  $\text{med}(y|x) = (x^t, 1)\tilde{\theta}$  for some  $\tilde{\theta}$ , then

$$\varepsilon^*(T_r^*, H) = \frac{1}{3}.$$

Another measure of robustness of an estimator is the *influence function*, which measures the effect of adding a small amount of contamination at a single point (Hampel et al. 1986). The influence function of the deepest regression was recently obtained (Van Aelst and Rousseeuw, 1998). It is a bounded and piecewise smooth function. For location however, the influence function of the Tukey median is not yet known.

Recently, He and Portnoy (1998) and Bai and He (1998) studied the asymptotic distribution of the deepest location and regression in  $p$  dimensions. They proved that the limiting distribution of both estimators depend on a Gaussian process. Since the moments of this distribution are not yet known, the asymptotic efficiency of both methods is still an open question, but some simulations have been carried out.

## 5 Computational aspects

The time complexity for the computation of the location depth or the regression depth of a fit are tabulated in Table 1.

In two dimensions, both the location depth and the regression depth of a fit can be computed in  $O(n \log n)$  time (Rousseeuw and Ruts 1996, Rousseeuw and Hubert 1999a). These results were used by Rousseeuw and Struyf (1998) to construct  $O(n^{p-1} \log n)$  time algorithms for  $ldepth(\theta, X_n)$  and  $rdepth(\theta, Z_n)$  in any dimension  $p$ . Since this is too slow for large  $n$  and/or high  $p$ , the latter authors have also proposed approximate algorithms. For location they obtain a complexity of  $O(mp^3 + mpn)$  time, and for regression a complexity of  $O(mp^3 + mpn + mn \log n)$  time. Here the parameter  $m$  can be chosen by the user, and determines the accuracy of the approximation.

Table 1: The time complexity of currently available exact and approximate algorithms for the location depth or the regression depth of a fit.

dimension	ldepth( $\theta, X_n$ )	rdepth( $\theta, Z_n$ )
$p = 2$ (exact)	$O(n \log n)$ Rousseeuw and Ruts (1996)	$O(n \log n)$ Rousseeuw and Hubert (1999a)
$p > 2$ (exact)	$O(n^{p-1} \log n)$ Rousseeuw and Struyf (1998)	$O(n^{p-1} \log n)$ Rousseeuw and Struyf (1998)
$p > 2$ (approx.)	$O(mp^3 + mpn)$ Rousseeuw and Struyf (1998)	$O(mp^3 + mpn + mn \log n)$ Rousseeuw and Struyf (1998)

To compute the deepest location in two dimensions, an algorithm with time complexity  $O(n^2 \log^2 n)$  was constructed by Rousseeuw and Ruts (1998). It is expected that a faster algorithm should be possible. Van Kreveld et al. (1999) constructed an algorithm for the deepest regression in two dimensions in  $O(n \log^2 n)$  time, which is close to linear. They considered the problem in the dual space, i.e. the fit space. Indeed, there is a nice relation between regression depth and arrangements of hyperplanes (Rousseeuw and Hubert 1999b). Therefore, several techniques from computational geometry have been used to construct this fast algorithm.

A naive exact algorithm for the deepest location in  $p > 2$  dimensions would require to compute the location depth at all intersections of  $p$  hyperplanes through  $p$  observations. For this we need  $O(n^{3p-1} \log n)$  time. In regression we have to compute the regression depth of all  $O(n^p)$  fits through  $p$  observations and keep the one(s) with maximal depth. This yields a  $O(n^{2p-1} \log n)$  time algorithm for the deepest regression. To speed up the computation, approximate algorithms have been constructed by Struyf and Rousseeuw (2000) for location and by Van Aelst et al. (1999) for regression. The complexities of these algorithms can be found in Table 2. Here again the default values of the parameters  $k$  and  $m$  can be increased to improve the accuracy of the algorithm, or they can be decreased to speed up the computation. In regression, the value  $h$  denotes the number of iterations until convergence, and is bounded

Table 2: Time complexity of currently available algorithms for the deepest location and for the deepest regression.

dimension	$T_l^*$	$T_r^*$
$p = 2$ (exact)	$O(n^2 \log^2 n)$ Rousseeuw and Ruts (1998)	$O(n \log^2 n)$ van Kreveld et al. (1999)
$p > 2$ (exact)	$O(n^{3p-1} \log n)$	$O(n^{2p-1} \log n)$
$p > 2$ (approx.)	$O(kmn \log n + knp)$ Struyf and Rousseeuw (2000)	$O(p^2 n + hpn + pn \log n)$ Van Aelst et al. (1999)

by 300.

## 6 Centrality

In Theorem 4 we saw that the breakdown value of the deepest location and regression estimators decreases as  $p$  increases, at least at some special configurations. It can be argued that this is due to the fact that  $ldepth$  and  $rdepth$  depend only on a kind of combinatorial structure, in the sense that they are invariant to ‘order-preserving’ transformations (see Rousseeuw and Hubert 1999a, Section 8). To obtain a higher breakdown value for any configuration we therefore need to go beyond the qualitative information contained in the multivariate ranking. In location this can be achieved by the minimum volume ellipsoid estimator and the minimum covariance determinant estimator (Rousseeuw 1985, Rousseeuw and Van Driessen 1999a), which use the quantitative notion of volume and attain a breakdown value of 50%. In regression, we can for instance use the least trimmed squares estimator (Rousseeuw 1984, Rousseeuw and Van Driessen 1999b). Note that the latter method uses quantitative information, namely the absolute values of the residuals, whereas  $rdepth$  depends only on their signs.

Here we will construct a more quantitative version of depth, which we will call

**centrality.** In location, we define the centrality of some  $\theta \in \mathbb{R}^p$  relative to  $X_n \subset \mathbb{R}^p$  as

$$lcent(\theta, X_n) = \inf_{\|u\|=1} M_l / (M_l + |\text{med}_i u^t(x_i - \theta)|) \quad (6)$$

where  $M_l$  (in which ‘l’ stands for location) is given by

$$M_l = \text{med}_i |u^t x_i - \text{med}_j u^t x_j| \quad (7)$$

and does not depend on  $\theta$ . Both (6) and (7) use quantitative information. Clearly,  $lcent$  is a dimensionless quantity between 0 and 1, like  $ldepth$ . The more  $\theta$  is centrally located, the larger  $lcent(\theta, X_n)$  becomes. This suggests the **most central fit** estimator, given by

$$T_l^c(X_n) = \underset{\theta}{\text{argmax}} lcent(\theta, X_n). \quad (8)$$

Note that this is a robust estimator because we have used a robust measure of centrality. In fact, for  $\theta = x_i$  there is a relation between  $lcent(x_i, X_n)$  and the ‘outlyingness’ of  $x_i$  as defined by Stahel (1981) and Donoho (1982). They used the outlyingness of  $x_i$  to compute a weight  $w(x_i)$ , yielding the Stahel-Donoho estimator which is a weighted mean of the  $x_i$ . The new estimator  $T_l^c$  in (8) is more radical in that it looks for the ‘innermost’ candidate fit. One could also measure centrality in a nonrobust way, e.g. by replacing the univariate medians in (6) and (7) by averages, but then the deepest fit  $T_l^c(X_n)$  becomes  $\bar{X}_n$  which has a zero breakdown value.

In regression we define the centrality of a candidate fit  $\theta$  relative to  $Z_n \subset \mathbb{R}^p$  as

$$rcent(\theta, Z_n) = \inf_{\substack{\|u\|=1 \\ v \in \mathbb{R}}} M_r / \left( M_r + \left| \text{med}_i \frac{r_i(\theta)}{u^t x_i - v} \right| \right) \quad (9)$$

where this time

$$M_r = \text{med}_i |y_i - \text{med}_j y_j| / \text{med}_i |u^t x_i - v|.$$

Again  $rcent(\theta, Z_n)$  lies between 0 and 1, and it measures how well  $\theta$  fits the data. The **most central fit** regression estimator becomes

$$T_r^c(Z_n) = \underset{\theta}{\text{argmax}} rcent(\theta, Z_n).$$

Note that  $lcent$ ,  $T_l^c$ ,  $rcent$ , and  $T_r^c$  can also be computed for population distributions. The following theorem shows that the maximal centrality (like the maximal depth) measures how well the data can be fitted.

**Theorem 7.** (a) If the distribution  $P$  on  $\mathbb{R}^p$  is angularly symmetric about some  $\tilde{\theta}$  then

$$\max_{\theta} lcent(\theta, P) = lcent(\tilde{\theta}, P) = 1.$$

(b) If  $H$  satisfies the conditions of Theorem 3(d) then

$$\max_{\theta} rcent(\theta, H) = rcent(\tilde{\theta}, H) = 1.$$

**Proof.** (a) Since  $P$  is angularly symmetric, and has a density,  $P(H_{\theta,u}) = P(\text{int } H_{\theta,u}) = P(\text{int } H_{\theta,-u}) = P(H_{\theta,-u})$ . Therefore  $\text{med}_P(u^t(x - \tilde{\theta})) = 0$ , hence (6) yields  $lcent(\tilde{\theta}, P) = 1$ . (b) Assume without loss of generality that  $\tilde{\theta} = 0$ . Since  $\text{med}_H(y|x) = 0$  and  $H$  has a strictly positive density, we have for each  $x$  that  $H(y < 0|x) = H(y \leq 0|x) = H(y \geq 0|x) = H(y > 0|x)$ . Take now any unit vector  $u$  and any number  $v$ , then  $u^t x - v > 0$  or  $u^t x - v < 0$  (since  $H(u^t x = v) = 0$ ). If  $u^t x - v > 0$  then  $H(\frac{y}{u^t x - v} \leq 0|x) = H(\frac{y}{u^t x - v} < 0|x) = H(y < 0|x) = H(y > 0|x) = H(\frac{y}{u^t x - v} \geq 0|x)$ , so  $\text{med}_H(\frac{y}{u^t x - v}|x) = 0$ . The same holds for  $u$  and  $v$  that satisfy  $u^t x - v < 0$ . Therefore, (9) yields  $rcent(0, H) = 1$ .

In general, it holds that:

**Theorem 8.** (a) The most central location estimator  $T_l^c$  is affine equivariant and has a 50% breakdown value.

(b) The most central regression estimator  $T_r^c$  is regression, scale, and affine equivariant, and has a 50% breakdown value.

Note that for univariate data  $T_l^c$  becomes the sample median. Therefore  $T_l^c$  is a *multivariate generalization of the median* which inherits the 50% breakdown value and is affine equivariant. This follows from the fact that  $T_l^c$  corresponds with one of the projection estimators of Tyler (1994, formula (3.8)). To our knowledge, it is the first such multivariate median.

Analogously, for univariate data (i.e., all  $x_i = 1$ ) also  $T_r^c$  becomes the sample median of the  $y_i$ . Therefore  $T_r^c$  generalizes the median to multiple regression, again inheriting the 50% breakdown value while satisfying all the equivariance properties. Note that  $T_r^c$  belongs to the class of projection estimators of Maronna and Yohai (1993).

**Remark.** The key quantity in (6) is  $|\text{med}_i u^t(x_i - \theta)|$  which depends on  $\theta$ . A qualitative analog is the smallest number of  $u^t(x_i - \theta)$  on either side of zero, yielding

$$\inf_{\|u\|=1} \min \left\{ \sum_{i=1}^n I(u^t(x_i - \theta) > 0), \sum_{i=1}^n I(u^t(x_i - \theta) < 0) \right\} = ldepth(\theta, X_n)$$

which recovers the location depth. Starting from (9), we analogously find

$$\inf_{\substack{\|u\|=1 \\ v \in \mathbb{R}}} \min \left\{ \sum_{i=1}^n I\left(\frac{r_i(\theta)}{u^t x_i - v} > 0\right), \sum_{i=1}^n I\left(\frac{r_i(\theta)}{u^t x_i - v} < 0\right) \right\} = rdepth(\theta, Z_n)$$

which recovers the notion of regression depth.

## References

- Bai, Z., and He, X. (1998), “Asymptotic distributions of the maximal depth estimators for regression and multivariate location,” *The Annals of Statistics*, to appear.
- Donoho, D.L. (1982), “Breakdown Properties of Multivariate Location Estimators,” Ph.D. Qualifying paper, Harvard University.
- Donoho, D.L., and Gasko, M. (1992), “Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness,” *The Annals of Statistics*, 20, 1803-1827.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.
- He, X., and Portnoy, S. (1998), “Asymptotics of the deepest line,” in *Applied Statistical Sciences III: Nonparametric statistics and related topics*, edited by Ahmed, S.E., Ahsanullah, M., and Sinha, B.K., 71-81, Nova Science Publications Inc., New York.
- Hubert, M., and Rousseeuw, P.J. (1998), “The catline for deep regression,” *Journal of Multivariate Analysis*, 66, 270-296.
- Liu, R.Y., Parelius, J.M., and Singh, K. (1999), “Multivariate analysis by data depth: descriptive statistics, graphics and inference,” *The Annals of Statistics*, to appear.
- Maronna, R.A. and Yohai, V.J. (1993), “Bias-robust Estimates of Regression Based on Projections,” *The Annals of Statistics*, 21, 965-990.

- Mizera, I. (1998), "On depth and deep points: a calculus," submitted.
- Rousseeuw, P.J. (1984), "Least median of squares regression," *Journal of the American Statistical Association*, 79, 871-880.
- Rousseeuw, P.J. (1985), "Multivariate estimation with high breakdown point," *Mathematical Statistics and Applications, Vol. B*, W. Grossmann, G. Pflug, I. Vincze and W. Wertz, eds., Dordrecht: Reidel, 283-297.
- Rousseeuw, P.J., and Hubert, M. (1999a), "Regression depth," *Journal of the American Statistical Association*, 94, 388-402.
- Rousseeuw, P.J., and Hubert, M. (1999b), "Depth in an arrangement of hyperplanes," *Discrete and Computational Geometry*, 22, 167-176.
- Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley.
- Rousseeuw, P.J., and Ruts, I. (1996), "AS 307: Bivariate Location Depth," *Applied Statistics*, 45, 516-526.
- Rousseeuw, P.J., and Ruts, I. (1998), "Constructing the bivariate Tukey median," *Statistica Sinica*, 8, 827-839.
- Rousseeuw, P.J., and Ruts, I. (1999), "The depth function of a population distribution," *Metrika*, 49, 213-244.
- Rousseeuw, P.J., and Struyf, A. (1998), "Computing location depth and regression depth in higher dimensions," *Statistics and Computing*, 8, 193-203.
- Rousseeuw, P.J., Van Aelst, S., and Hubert, M. (1999), "Rejoinder to the discussion of regression depth", *Journal of the American Statistical Association*, 94, 419-433.
- Rousseeuw, P.J., and Van Driessen, K. (1999a), "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, 41, 212-223.
- Rousseeuw, P.J., and Van Driessen, K. (1999b), "Computing LTS regression for large data sets," submitted.

- Stahel, W.A. (1981), “Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators,” Ph.D. thesis (in German), ETH, Zürich.
- Struyf, A., and Rousseeuw, P.J. (2000), “High-dimensional computation of the deepest location,” *Computational Statistics and Data Analysis*, to appear.
- Struyf, A., and Rousseeuw, P.J. (1999), “Halfspace depth and regression depth characterize the empirical distribution,” *Journal of Multivariate Analysis*, 69, 135-153.
- Tukey, J.W. (1975), “Mathematics and the Picturing of Data,” *Proceedings of the International Congress of Mathematicians*, Vancouver, 2, 523-531.
- Tyler, D.E. (1994), “Finite sample breakdown points of projection based multivariate location and scatter statistics,” *The Annals of Statistics*, 22, 1024-1044.
- Van Aelst, S., and Rousseeuw, P.J. (1998), “Robustness properties of deepest regression,” *Journal of Multivariate Analysis*, to appear.
- Van Aelst, S., Rousseeuw, P.J., Hubert, M., and Struyf, A. (1999), “Deepest regression in practice,” Technical report, University of Antwerp.
- van Kreveld, M., Mitchell, J., Rousseeuw, P.J., Sharir, M., Snoeyink, J., and Speckmann, B. (1999), “Efficient algorithms for maximum regression depth,” Proc. 15th ACM Symposium on Computational Geometry, 31-40.
- Zhang, J. (1998), “Some extensions of Tukey’s depth function,” submitted.