
In this paper we introduce a notion of depth in the regression setting. It provides the ‘rank’ of any line (plane), rather than ranks of observations or residuals. In simple regression we can compute the depth of any line by a fast algorithm. For any bivariate data set Z_n of size n there exists a line with depth at least $n/3$. The largest depth in Z_n can be used as a measure of linearity versus convexity. In both simple and multiple regression we introduce the deepest regression method, which generalizes the univariate median and is equivariant for monotone transformations of the response. Throughout, the errors may be skewed and heteroskedastic. We also consider depth-based regression quantiles. They estimate the quantiles of y given x , as do the Koenker-Bassett regression quantiles, but with the advantage of being robust to leverage outliers. We explore the analogies between depth in regression and in location, where Tukey’s halfspace depth is a special case of our general definition. Also, Liu’s simplicial depth can be extended to the regression framework.

KEY WORDS: Asymmetric error distribution; Deepest regression; Depth envelopes; Depth quantiles; Geometry; Halfspace depth; Heteroskedasticity; Robust regression; Simplicial depth.

1. INTRODUCTION

In this paper we propose the notion of regression depth. We view depth as a property of a *fit* (typically determined by a vector θ of coefficients), rather than a property of an *observation*. In general we define the depth of a (candidate) fit θ to a given data set Z_n of size n as

Definition 1. The depth(θ, Z_n) is the smallest number of observations of Z_n that would need to be removed in order to make θ a nonfit.

Therefore, always $0 \leq \text{depth}(\theta, Z_n) \leq n$. (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).

Section 2 defines nonfits in simple regression, and constructs a fast algorithm for computing the regression depth of a line. We show that the largest regression depth relative to a data set Z_n is at least $n/3$, by constructing a line which is at least that deep. The maximal depth of Z_n is at its highest when the linear model holds well, and at its lowest when the data lie on a parabola or any other curve which is strictly convex or concave. The maximal regression depth thus reflects the degree of linearity in the data. We then consider the *deepest regression line* (that is, with largest regression depth), which generalizes the univariate median. The deepest regression line has a breakdown value of $1/3$. Around that line we construct depth envelopes, which are useful for graphical display.

Section 3 defines regression depth in the multiple regression setting with p coefficients, with the corresponding deepest regression fit and depth envelopes. The breakdown value of the deepest regression approaches $1/3$ in any dimension. Due to the monotone invariance of regres-

sion depth, the deepest regression is equivariant to monotone transformations of the response, unlike least squares, least absolute values (L^1) or least median of squares. Throughout the paper, the errors may be skewed and non-identically distributed (e.g. heteroskedastic).

In Section 4 we consider depth-based regression quantiles. They estimate the conditional quantile of y given x , as do the customary L^1 -based regression quantiles of Koenker and Bassett (1978), but with the additional advantage of being robust to leverage outliers.

Section 5 gives another interpretation of regression depth, using duality. For bivariate data the dual plot represents observations as lines, and candidate fits as points. This yields some interesting insights, and new results of geometry.

Section 6 focuses on the multivariate location setting. By applying Definition 1 we recover the halfspace depth of Tukey (1975) as a special case. This location depth has been used as a multivariate generalization of rank, see Green (1981) and Eddy (1985). The deepest location (Donoho and Gasko 1992) also generalizes the univariate median, and its depth is highest when the data are angularly symmetric. Interestingly, it turns out that (under additional conditions) both types of depth have precursors in the fifties, since simple regression depth is related to a test of Daniels (1954) in exactly the same way that bivariate location depth is related to a test of Hodges (1955). Section 7 explains how Liu’s simplicial depth (1990) for location can be extended to the regression context.

Both the deepest regression and L^1 regression are generalizations of the univariate median. Section 8 argues that the deepest regression is the more natural one, and also compares it with least median of squares and least trimmed squares (Rousseeuw 1984). Section 9 describes some directions for further research.

2. SIMPLE REGRESSION

Peter J. Rousseeuw is Professor and Mia Hubert is Assistant, Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen (UIA), Universiteitsplein 1, B-2610 Wilrijk, Belgium. We would like to thank Anja Struyf and Stefan Van Aelst for computational assistance. We are grateful to the Editor, Associate Editor, and three referees for comments which improved the presentation.

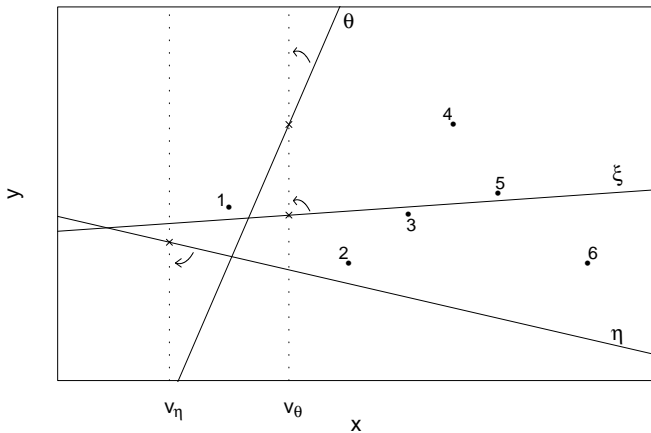


Figure 1. Bivariate data set with two nonfits θ and η , and a fit ξ with regression depth 2.

2.1 Definition of regression depth

In simple regression we want to fit a straight line $y = \theta_1 x + \theta_2$ to a data set $Z_n = \{(x_i, y_i); i = 1, \dots, n\} \subset \mathbb{R}^2$. All candidate fits will be denoted as $\theta = (\theta_1, \theta_2)$ so the first component is the slope estimate and the second is the intercept term. The residuals are then denoted as $r_i(\theta) = r_i = y_i - \theta_1 x_i - \theta_2$.

Definition 2. A candidate fit $\theta = (\theta_1, \theta_2)$ to Z_n will be called a **nonfit** iff there exists a real number $v_\theta = v$ which does not coincide with any x_i and such that

$$r_i(\theta) < 0 \text{ for all } x_i < v \text{ and } r_i(\theta) > 0 \text{ for all } x_i > v$$

or

$$r_i(\theta) > 0 \text{ for all } x_i < v \text{ and } r_i(\theta) < 0 \text{ for all } x_i > v.$$

Figure 1 shows a data set with 6 observations and two nonfits θ and η . Also the corresponding values v_θ and v_η are indicated. From this plot it is clear that the existence of v corresponds to the presence of a tilting point (marked by a cross) around which we can rotate the line until it is vertical, while not passing any observation. Note that a line lying above or below all the observations (such as the line η) is always a nonfit.

The notion of regression depth now follows immediately from the general concept of depth (Definition 1):

Definition 3. The regression depth (rdepth) of a fit θ relative to a data set Z_n is the smallest number of observations that need to be removed to make θ a nonfit. Equivalently, $rdepth(\theta, Z_n)$ is the smallest number of residuals that need to change sign.

For example, consider the line ξ in Figure 1. We can make it a nonfit by removing observations 4 and 5 (since one can then tilt ξ vertically without touching any remaining observations, e.g. using v_θ). Since ξ cannot be made a nonfit by removing fewer observations, $rdepth(\xi, Z_n) = 2$. Note that Definitions 2 and 3 allow for ties among the x_i

and that they do not require any distributional assumptions.

To compute $rdepth(\theta, Z_n)$ we first reorder the observations such that $x_1 \leq x_2 \leq \dots \leq x_n$ in $O(n \log n)$ time. Then we can compute the rdepth in $O(n)$ operations using

$$rdepth(\theta, Z_n) = \min_{1 \leq i \leq n} (\min\{L^+(x_i) + R^-(x_i), R^+(x_i) + L^-(x_i)\}) \quad (2.1)$$

where

$$L^+(v) = \#\{j; x_j \leq v \text{ and } r_j \geq 0\},$$

$$R^-(v) = \#\{j; x_j > v \text{ and } r_j \leq 0\},$$

and L^- and R^+ are defined accordingly. It therefore suffices to update $L^+(x_i)$, $L^-(x_i)$, $R^-(x_i)$ and $R^+(x_i)$ at each $i = 1, \dots, n$.

From Definition 3 it follows that regression depth is scale invariant, regression invariant and affine invariant, according to the definitions in Rousseeuw and Leroy (1987, page 116). In the population case $rdepth(\theta, H)$ is defined as the smallest probability mass that has to be removed, where H is the joint distribution of the (x, y) .

2.2 The maximal rdepth

Definition 3 implies that the rdepth of any fit is at most n . This upper bound is reached if all the (x_i, y_i) lie exactly on a straight line. In general the maximal rdepth will be lower. Theorem 1 establishes lower and upper bounds for the maximal rdepth.

Theorem 1. (a) At any data set $Z_n \subset \mathbb{R}^2$ it holds that

$$\max_{\theta} rdepth(\theta, Z_n) \geq \left\lceil \frac{n}{3} \right\rceil \quad (2.2)$$

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

(b) If the (x_i, y_i) are in general position, i.e., no three (x_i, y_i) lie on a line,

$$\max_{\theta} rdepth(\theta, Z_n) \leq \left\lceil \frac{n+2}{2} \right\rceil. \quad (2.3)$$

(c) For any (x, y) -distribution H on \mathbb{R}^2 with a density it holds that

$$\frac{1}{3} \leq \max_{\theta} rdepth(\theta, H) \leq \frac{1}{2}. \quad (2.4)$$

(d) If H has a density and satisfies

$$\text{med}[y|x] = \tilde{\theta}_1 x + \tilde{\theta}_2 \quad (2.5)$$

for some $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2) \in \mathbb{R}^2$ then

$$\max_{\theta} rdepth(\theta, H) = rdepth(\tilde{\theta}, H) = \frac{1}{2}. \quad (2.6)$$

Condition (2.5) is very weak, e.g. it does not require that the conditional distribution of $y|x$ be symmetric around $\tilde{\theta}_1 x + \tilde{\theta}_2$, or that it stays the same across different values of x . Note that our functional form is parametric whereas

the error distribution is nonparametric, so we have a semi-parametric model. This model is very large and allows for *skewness* and *heteroskedasticity*, which often occur in practice.

Theorem 2 shows when the lower bounds in Theorem 1 are reached.

Theorem 2. (a) If all x_i are distinct and the (x_i, y_i) lie on a strictly convex (or strictly concave) curve, then

$$\max_{\theta} rdepth(\theta, Z_n) = \left\lceil \frac{n+2}{3} \right\rceil.$$

(b) If the probability mass of H is concentrated on a strictly convex (or concave) curve and has a density on that curve, then

$$\max_{\theta} rdepth(\theta, H) = \frac{1}{3}.$$

From Theorems 1 and 2 it follows that $\max_{\theta} rdepth(\theta, Z_n)$ can be seen as a measure of linearity for Z_n . Note that this quantity can be computed in $O(n^3)$ time from

$$\max_{\theta} rdepth(\theta, Z_n) = \max_{i < j} rdepth(\theta^{ij}, Z_n)$$

where θ^{ij} is the line passing through the observations i and j . We therefore only need to compute the $rdepth$ in each of the $O(n^2)$ lines θ^{ij} . If we first sort the x_i and then use the $O(n)$ algorithm of (2.1) to compute $rdepth(\theta^{ij}, Z_n)$, we spend $O(n \log n + n^3) = O(n^3)$ time.

2.3 The catline

We prove the lower bound in (2.2) by constructing a regression estimator that always has $rdepth$ at least $\lceil n/3 \rceil$. For any data set with $x_1 \leq \dots \leq x_n$ we divide the observations in three groups denoted by L, M , and R . If n is a multiple of 3, the left group L is formed by the first $m = n/3$ data points $\{(x_1, y_1), \dots, (x_m, y_m)\}$, the group M by the middle third, and R by the rightmost third. For $n = 3m + 1$ we take $\#M = m + 1$, whereas for $n = 3m + 2$ we take $\#L = \#R = m + 1$.

Definition 4. The **catline** is the line that simultaneously bisects the sets $L \cup M$ and $M \cup R$.

(We say that a line bisects a set of N points if neither of the two open halfplanes defined by that line contains more than $\lfloor N/2 \rfloor$ points). In the Appendix it is shown that the catline always exists, and that its $rdepth$ is at least $\lceil n/3 \rceil$. In the population case, the catline partitions the probability mass according to (2.7), where the horizontal line indicates the catline and the vertical lines separate the groups L, M and R :

$$\begin{array}{c|c|c} q & \frac{1}{3} - q & q \\ \hline \frac{1}{3} - q & q & \frac{1}{3} - q \end{array} \quad (2.7)$$

For each distribution H on \mathbb{R}^2 , there is a unique value of $0 \leq q \leq \frac{1}{3}$ satisfying (2.7).

Roughly speaking, the catline has the property that the number of positive residuals in L equals the number of negative residuals in M and the number of positive residuals in R . We call it the catline since it Cuts All Thirds (that is, L, M , and R).

Theorem 3. Suppose that all x_i are distinct. If the (x_i, y_i) lie on a straight line, or on a strictly convex (concave) curve, then

$$rdepth(\theta_{CAT}, Z_n) = \max_{\theta} rdepth(\theta, Z_n).$$

In (Hubert and Rousseeuw 1998) an efficient $O(n \log n)$ algorithm is constructed to compute the catline. There the properties of the catline are studied more thoroughly, e.g. its influence function is obtained. The breakdown value of the catline is shown to be $1/3$. Recall that the breakdown value $\varepsilon_n^*(\theta_{CAT}, Z_n)$ is the smallest fraction of Z_n that needs to be replaced to carry θ_{CAT} arbitrarily far away. (For background on the breakdown value, see Rousseeuw and Leroy 1987).

2.4 The deepest regression line

We define the **deepest regression** estimator $T_r^*(Z_n)$ as the θ with the largest $rdepth$:

$$\begin{aligned} T_r^*(Z_n) &= \operatorname{argmax}_{\theta} rdepth(\theta, Z_n) \\ &= \operatorname{argmax}_{\theta^{ij}} rdepth(\theta^{ij}, Z_n). \end{aligned} \quad (2.8)$$

If there are several θ^{ij} with that same $rdepth$, the average of those θ^{ij} is taken. Expression (2.8) gives a straightforward algorithm for computing T_r^* in $O(n^3)$ time (see Section 2.2), but further work is likely to yield a faster algorithm.

The regression depth of a fit θ says how well it is surrounded by data, i.e. how ‘balanced’ it is. When $n = 25$ and $rdepth(\theta, Z_n) = 10$, one cannot tip the line easily: no matter how we try to tilt it, there will always be at least 10 data points in our way. In this sense, the deepest regression fit is ‘most balanced,’ i.e. as much in equilibrium as possible.

Note that T_r^* is a generalization of the univariate median. For univariate data, $depth(\theta, Z_n) = \min(\#\{y_i \leq \theta\}, \#\{y_i \geq \theta\})$ so the median is most balanced. (For regression data with $x_i = 0$, the intercept of T_r^* equals the median of the y_i .) Another generalization of the univariate median is the least absolute values (or L^1) regression method, defined by minimizing $\sum_{i=1}^n |r_i|$. However, T_r^* and L^1 are quite different. For instance, the breakdown value of L^1 is zero, whereas that of the deepest regression line is at least 33%:

Theorem 4. For any data set $Z_n \subset \mathbb{R}^2$ with distinct x_i we have

$$\varepsilon_n^*(T_r^*, Z_n) \geq \frac{1}{n} \left(\left\lceil \frac{n}{3} \right\rceil - 1 \right) \approx \frac{1}{3}. \quad (2.9)$$

Example 1: Stars data. Figure 2 contains the Hertzsprung-Russell diagram of a star cluster in the direction of Cygnus (from Rousseeuw and Leroy 1987, page

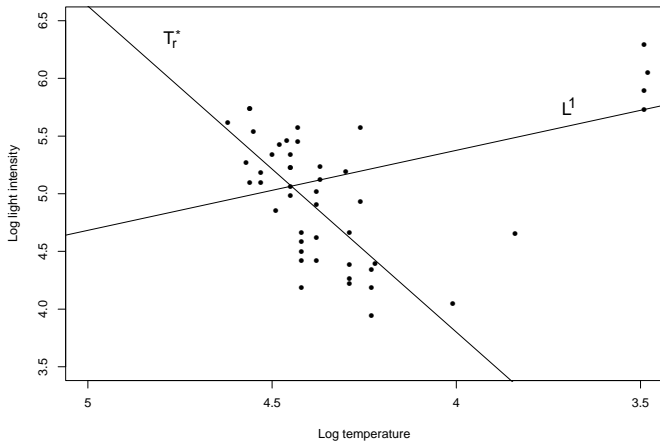


Figure 2. Hertzsprung-Russell diagram of a star cluster in the direction of Cygnus, with the deepest regression line T_r^* and the least absolute deviations fit L^1 which is attracted by the giant stars.

27). The logarithm of the star’s light intensity is plotted versus the logarithm of its surface temperature. We see that the L^1 line is attracted by the four leverage points in the upper right corner (the giant stars), whereas the deepest regression line fits the majority (the main sequence stars).

Note that the deepest regression line is but one step in the data analysis process. One of its uses is to detect outliers, which have large residuals from it. The outliers should then be investigated, and (depending on the context) they may be corrected, downweighted, or deleted, or the model may be changed.

Remark 1: Monotone equivariance. By definition, the regression depth of a fit only depends on the x_i and the **sign** of the residuals r_i . This allows for monotone transformations of the response y_i . Assume that the functional model is

$$y = g(\theta_1 x + \theta_2) \tag{2.10}$$

with g a strictly monotone link function. Typical examples of g include the logarithmic, the exponential, the square root, the square and the reciprocal transformation. We can then define the rdepth of the curved fit (2.10) as in Definition 3, and search for the deepest regression curve. But due to the invariance, there is an easier way: it suffices to put $\tilde{y}_i = g^{-1}(y_i)$ and to determine the deepest regression line $(\hat{\theta}_1, \hat{\theta}_2)$ of the transformed data (x_i, \tilde{y}_i) . Then we can backtransform the deepest regression line, yielding the deepest regression curve $y = g(\hat{\theta}_1 x + \hat{\theta}_2)$. Note that this monotone equivariance property (inherited from the univariate median) does not hold for least squares, L^1 , or existing positive-breakdown methods such as least trimmed squares (Rousseeuw 1984). The latter methods are not equivariant because monotone transformations change the relative sizes of residuals throughout the data set.

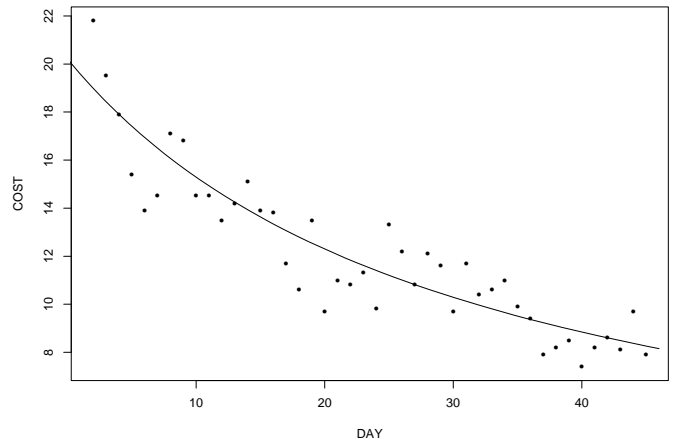


Figure 3. Plot of the cost data with the deepest regression curve, obtained by backtransforming the deepest regression line of the transformed data $\{(x_i, 1/y_i); i = 1, \dots, n\}$.

Example 2: Cost data. Figure 3 shows the cost per record transcribed (response COST) during 45 successive days (regressor DAY). This data set is available from the DASL library at <http://lib.stat.cmu.edu/DASL>. (Here we consider it as a regression data set rather than a time series.) A reciprocal transformation of COST (i.e. $\tilde{y}_i = g^{-1}(y_i) = 1/y_i$) yielded a scatterplot with roughly linear trend. In Figure 3 we have drawn the resulting deepest regression curve.

To investigate the finite-sample performance of the deepest regression, we have generated $m = 10,000$ samples with various sample sizes n satisfying the linear model

$$y_i = \theta_1 x_i + \theta_2 + e_i \quad \text{for } i = 1, \dots, n \tag{2.11}$$

where $e_i \sim N(0, \sigma^2)$. Because of regression equivariance, we set $\theta = (\theta_1, \theta_2) = (0, 0)$. First, we have generated gaussian x_i . For each n , column 2 of Table 1 lists the relative efficiency of the deepest regression slope compared to the L^1 slope, showing that the deepest regression slope attains roughly 88% of the efficiency of its L^1 counterpart. But for uniformly distributed x_i the deepest regression slope is almost as efficient as the L^1 slope (whereas its breakdown value is much better). For the asymptotics of the deepest regression line, see Section 9.

2.5 Depth envelopes

Around the deepest regression line we can construct *depth*

Table 1. Relative efficiency (based on 10,000 simulations) of the deepest regression line compared to the L^1 line, when the x_i are gaussian or uniform.

n	gaussian x_i		uniform x_i	
	slope	intercept	slope	intercept
10	60.0%	78.8%	72.1%	73.3%
50	79.7%	86.4%	97.4%	84.4%
100	89.2%	85.2%	97.7%	85.6%
300	88.9%	82.5%	98.7%	81.0%
500	88.1%	83.3%	99.4%	81.4%

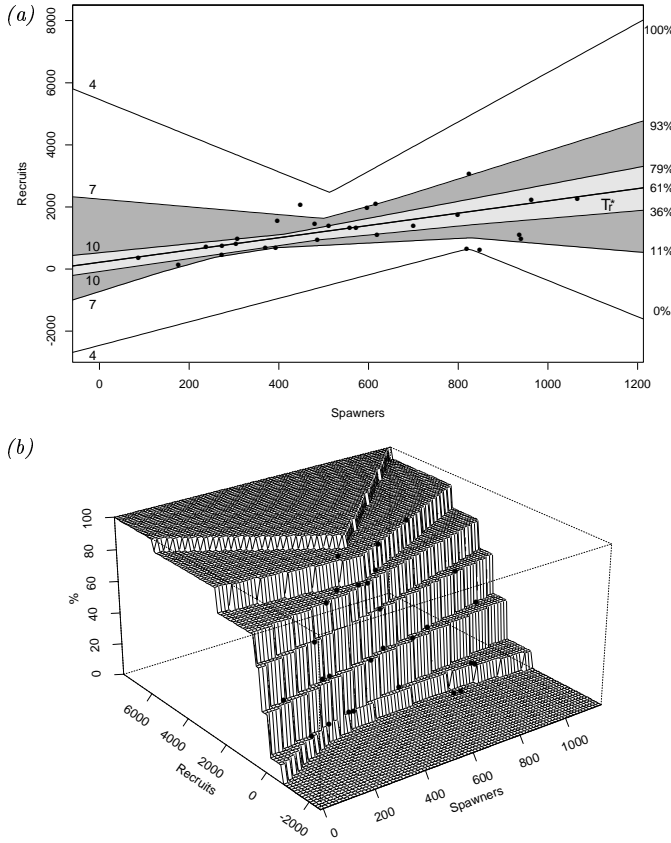


Figure 4. (a) The Skeena River data set ($n = 28$), its deepest regression line T_r^* (with depth 12) and its rdepth envelopes for $k = 4, 7$ and 10 . To the left of each envelope boundary its value of k is listed, and to its right the (cumulative) percentage of the data lying on or below it; (b) terrace plot of the depth envelope boundaries, with the percentages on the vertical axis.

envelopes E_k for $k \geq 2$, given by

$$E_k = \{(x, y); \min_{\theta}(\theta_1 x + \theta_2) \leq y \leq \max_{\theta}(\theta_1 x + \theta_2)\} \quad (2.12)$$

where $\theta = (\theta_1, \theta_2)$ needs to satisfy $rdepth(\theta, Z_n) \geq k$. It is proved in the Appendix that

$$E_k = \{(x, y); \min_{i,j}(\theta_1^{ij} x + \theta_2^{ij}) \leq y \leq \max_{i,j}(\theta_1^{ij} x + \theta_2^{ij})\} \quad (2.13)$$

where θ^{ij} ranges over all lines through two data points, for which $rdepth(\theta^{ij}, Z_n) \geq k$.

The upper and lower boundaries of the envelope E_k thus consist of line segments. The upper boundary is always convex, and the lower boundary is concave. Like the deepest regression, also the depth envelopes do not depend on assuming a particular type of error distribution.

Example 3: Skeena River data. Figure 4a plots the number of recruits versus the number of spawners from 1940 until 1967 for the Skeena River salmon stock (Carroll and Ruppert 1988). We see the deepest regression line T_r^* (with rdepth 12) and the depth envelopes for $k = 4, 7$, and 10 . (A referee suggested to system-

atically draw the depth envelopes for $k = 0.75k^*$, $k = 0.5k^*$, and $k = 0.25k^*$, with k^* the maximal rdepth.) Note that the set of envelopes always provides a (coarse) ordering of the observations. On the right hand side of Figure 4a we have indicated the percentage of the data lying on or below each envelope boundary. Figure 4b shows the same plot but with the percentages on a third (vertical) axis. It is a 3-dimensional variation on the usual empirical distribution function; for instance, cutting it with a vertical plane $x = x_0$ always yields a nondecreasing step function. For a population distribution H on \mathbb{R}^2 with a strictly positive density, the envelope boundaries become smooth curves, one for each rdepth α with $0 < \alpha < \alpha^* = \max_{\theta} rdepth(\theta, H)$. Also the surface in Figure 4b then becomes smooth.

3. MULTIPLE REGRESSION

3.1 Definition of regression depth

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

$$\theta_1 x_{i1} + \dots + \theta_{p-1} x_{i,p-1} + \theta_p = (\mathbf{x}_i, 1)\theta' \quad (3.1)$$

that is, by an affine hyperplane in \mathbb{R}^p , where $\theta = (\theta_1, \dots, \theta_p) \in \mathbb{R}^p$. Also here we do not make any distributional assumptions.

Definition 5. A fit $\theta = (\theta_1, \dots, \theta_p)$ is called a **nonfit** to Z_n iff there exists an affine hyperplane V in \mathbf{x} -space such that no \mathbf{x}_i belongs to V , and such that $r_i(\theta) > 0$ for all \mathbf{x}_i in one of its open halfspaces, and $r_i(\theta) < 0$ for all \mathbf{x}_i in the other open halfspace.

An example of a nonfit is shown in Figure 5. Here $p = 3$, so θ corresponds to a plane. The \mathbf{x} -space can be seen as the horizontal plane given by $y \equiv 0$, which contains the line V .

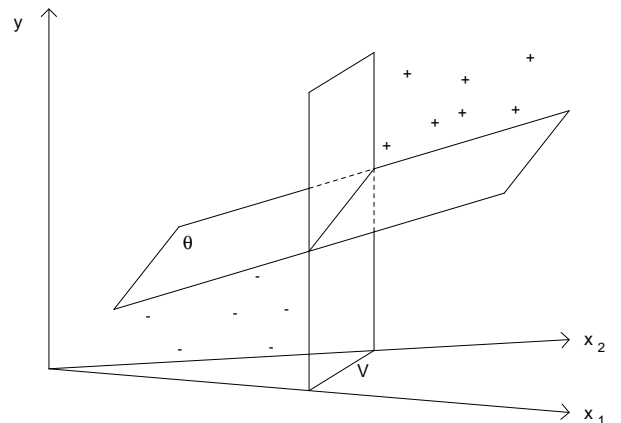


Figure 5. An example of a nonfit $\theta \in \mathbb{R}^3$. The affine hyperplane V in \mathbf{x} -space (\mathbb{R}^2) separates the observations with positive residuals from those with negative residuals.

(By definition, any $\boldsymbol{\eta}$ with all $r_i(\boldsymbol{\eta}) > 0$ or all $r_i(\boldsymbol{\eta}) < 0$ is also a nonfit, since it then suffices to choose V such that all \mathbf{x}_i lie on the same side of V .)

The regression depth of any $\boldsymbol{\theta} \in \mathbb{R}^p$ relative to $Z_n \subset \mathbb{R}^p$ is then given by Definition 3, as another instance of the general Definition 1.

Note that we can write $rdepth(\boldsymbol{\theta}, Z_n)$ in a form similar to (2.1). Any affine hyperplane V in \mathbf{x} -space splits the observations into two groups, which we will denote by $L(V)$ and $R(V)$. Set $L^+(V)$, resp. $L^-(V)$ the observations in $L(V)$ with positive, resp. negative residual. Similarly, $R(V)$ is partitioned into $R^+(V)$ and $R^-(V)$. Then

$$rdepth(\boldsymbol{\theta}, Z_n) = \min_V (\min\{L^+(V) + R^-(V), R^+(V) + L^-(V)\}) \quad (3.2)$$

where V ranges over all affine hyperplanes in \mathbf{x} -space. To compute the rdepth of a fit, we can reduce the search to a finite collection of hyperplanes V . This is outlined in (Rousseeuw and Struyf 1998) where exact algorithms are constructed that take $O(n^{p-1} \log n)$ time. Since this is only feasible for small p , that paper also proposes an approximate algorithm whose complexity is proportional to $n(p + \log n)$.

Remark 2. Another interpretation is obtained by noting that $\boldsymbol{\theta}$ is a nonfit to Z_n iff there exists an affine line L (take $L \perp V$) in \mathbf{x} -space such that, if we project all (\mathbf{x}_i, r_i) on the vertical plane through L , the line $r = 0$ is a nonfit for simple regression. Consequently, for $p \geq 3$ we find that

$$rdepth(\boldsymbol{\theta}, Z_n) = \min_L rdepth((0, 0), \pi_{L, \boldsymbol{\theta}}(Z_n))$$

where L ranges over all lines in \mathbf{x} -space, $\pi_{L, \boldsymbol{\theta}}(Z_n)$ is the projection of all $(\mathbf{x}_i, r_i(\boldsymbol{\theta}))$ on the vertical plane through L , and on the right hand side ‘rdepth’ stands for the simple regression depth of the fit $\boldsymbol{\eta} = (0, 0)$.

By definition, any fit passing through k points will have a regression depth of at least k . Moreover, in such situations there also exists an upper bound on rdepth:

Theorem 5. Exact fit property. If the number of observations lying on $\boldsymbol{\theta}$ is k (where $0 \leq k \leq n$), then

$$k \leq rdepth(\boldsymbol{\theta}, Z_n) \leq \left\lceil \frac{n+k}{2} \right\rceil. \quad (3.3)$$

For $k = n$, this confirms that $rdepth(\boldsymbol{\theta}, Z_n) = n$.

Illustration 1. Since the least absolute values (L^1) estimator always passes through at least p observations, its rdepth is at least p . The same holds for the version of the least trimmed squares estimator (Rousseeuw 1984) obtained by fitting (some or all) p -subsets exactly.

Illustration 2. The least squares (LS) estimator is never a nonfit. In fact, if Z_n is such that $X_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ has full rank and $n \geq 2p$, then $rdepth(\boldsymbol{\theta}_{LS}, Z_n) \geq 1$. This bound is sharp. (See the Appendix.)

As in simple regression, we can consider the rdepth of a fit with regard to a *population*. Let H be the joint distribution of the (\mathbf{x}, y) . Then a nonfit $\boldsymbol{\theta}$ is defined as in Definition 5 with $H(\mathbf{x} \in V) = 0$, and $rdepth(\boldsymbol{\theta}, H)$ is the smallest probability mass that has to be removed to make $\boldsymbol{\theta}$ a nonfit. Theorem 6 shows that $rdepth(\boldsymbol{\theta}, Z_n)$ is a consistent estimator for $rdepth(\boldsymbol{\theta}, H)$ if Z_n is sampled from H .

Theorem 6. If Z_n is sampled from a distribution H with a density, then

$$\frac{rdepth(\boldsymbol{\theta}, Z_n)}{n} \xrightarrow[n \rightarrow \infty]{a.s.} rdepth(\boldsymbol{\theta}, H). \quad (3.4)$$

3.2 Maximal rdepth and deepest regression

The following results provide upper bounds on the maximal regression depth, generalizing those of Theorem 1 to multiple regression. A subset of \mathbb{R}^p is said to be in general position if no more than p observations lie in any $(p - 1)$ -dimensional affine subspace.

Theorem 7. (a) If the (\mathbf{x}_i, y_i) are in general position,

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_n) \leq \left\lceil \frac{n+p}{2} \right\rceil. \quad (3.5)$$

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

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) \leq \frac{1}{2}. \quad (3.6)$$

(c) If H has a density and

$$\text{med}[y|\mathbf{x}] = \tilde{\theta}_1 x_1 + \dots + \tilde{\theta}_{p-1} x_{p-1} + \tilde{\theta}_p \quad (3.7)$$

for some $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \dots, \tilde{\theta}_p) \in \mathbb{R}^p$ then

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) = rdepth(\tilde{\boldsymbol{\theta}}, H) = \frac{1}{2}. \quad (3.8)$$

Moreover, we conjecture that the lower bound $n/3$ generalizes to $n/(p + 1)$.

Conjecture 1. (a) For any data set $Z_n \subset \mathbb{R}^p$ it holds that

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_n) \geq \left\lceil \frac{n}{p+1} \right\rceil. \quad (3.9)$$

(b) For any distribution H on \mathbb{R}^p with a density it holds that

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) \geq \frac{1}{p+1}. \quad (3.10)$$

Remark 3. As in Theorem 2, we can show that there exist configurations at which the lower bounds in (3.9) and (3.10) are reached. Suppose the probability mass of H is concentrated on the ‘moment curve’ $\{(u, u^2, \dots, u^p); u > 0\}$ and has a density on that curve. Then $\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) = 1/(p + 1)$.

The deepest regression T_r^* is defined as the $\boldsymbol{\theta}$ that maximizes $rdepth(\boldsymbol{\theta}, Z_n)$. It is the ‘most balanced’ fit, which

does not want to tilt about any V . We can find T_r^* by computing the rdepth of all fits through p observations. In combination with the exact $O(n^{p-1} \log n)$ algorithm for $rdepth(\boldsymbol{\theta}, Z_n)$ this would yield an overall computation time of $O(n^{2p-1} \log n)$. Therefore, we are working on faster (approximate) algorithms for T_r^* .

In multiple regression the deepest regression is still equivariant for monotone transformations of y , as can be seen from (2.5) and (3.2), unlike L^1 and least trimmed squares. We now derive the breakdown value of T_r^* .

Corollary of Conjecture 1. If Conjecture 1 holds, and the \mathbf{x}_i are in general position,

$$\varepsilon_n^*(T_r^*, Z_n) \geq \frac{1}{n} \left(\left\lceil \frac{n}{p+1} \right\rceil - p + 1 \right) \approx \frac{1}{p+1}. \quad (3.11)$$

This tells us that the breakdown value of the deepest regression is always positive, but it can be $1/(p+1)$ when the original ('uncontaminated') data are themselves peculiar, e.g. when they lie on the moment curve. 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 8. Let $Z_n = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ be a sample from a distribution H on \mathbb{R}^p ($p \geq 3$) with a strictly positive density, which satisfies (3.7). Then

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

Note that this result does not depend on Conjecture 1. The condition that H has a density excludes cases where the support of \mathbf{x} has Lebesgue measure zero, such as a hyperplane (i.e. multicollinearity). Theorem 8 says that the deepest regression does not break down when at least 67% of the data are generated from the model, while the remaining data (i.e., up to 33% of the points) may be anything.

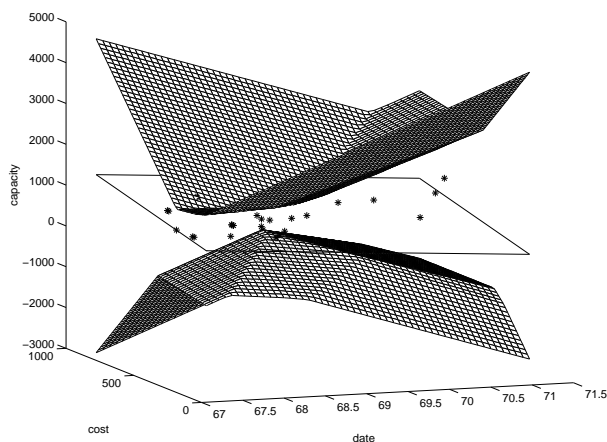


Figure 6. Deepest regression plane and the envelope with depth 6 for the Nuclear Power data set.

The depth envelopes around T_r^* are again given by

$$E_k = \{(\mathbf{x}, y); \min_J (\theta_1^J x_1 + \dots + \theta_{p-1}^J x_{p-1} + \theta_p^J) \leq y \leq \max_J (\theta_1^J x_1 + \dots + \theta_{p-1}^J x_{p-1} + \theta_p^J)\}$$

over all J with $rdepth(\boldsymbol{\theta}^J, Z_n) \geq k$. Here J is a p -subset of Z_n and $\boldsymbol{\theta}^J$ is the fit that passes through the observations in J . The boundaries of E_k are now piecewise planar surfaces.

Example 4: Nuclear Power data. In Figure 6 we see the deepest regression plane and the depth envelope with $k = 6$ for the Nuclear Power data set (from the DASL library at <http://lib.stat.cmu.edu/DASL>). The regressors are the date of construction and the cost of 32 light water nuclear power plants, and the response is their net capacity. In Figures 4a and 6 we see that each envelope E_k is rather narrow in the region of the available \mathbf{x}_i whereas they become much wider outside that region, where any fit is but an extrapolation.

3.3 Regression through the origin

In the setting of regression through the origin, we want to fit the y_i by

$$\theta_1 x_{i1} + \dots + \theta_p x_{ip} = \mathbf{x}\boldsymbol{\theta}' \quad (3.12)$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ and $Z_n = \{(x_{i1}, \dots, x_{ip}, y_i); i = 1, \dots, n\} \subset \mathbb{R}^{p+1}$. We assume that any observations with $(x_{i1}, \dots, x_{ip}) = (0, \dots, 0)$ have been deleted. For the definition of a nonfit $\boldsymbol{\theta}$, we modify Definition 5 by requiring that V passes through the origin (of \mathbf{x} -space). Then the regression depth $rdepth_0(\boldsymbol{\theta}, Z_n)$ is defined accordingly, as in Definition 3 and (3.2).

The $rdepth_0$ remains the same if we carry out the following construction. First we make sure that $x_{ip} \neq 0$ for all $i = 1, \dots, n$ (if necessary, we carry out a nonsingular linear transformation on $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ to achieve this). Next, we put $\tilde{Z}_n = \{(\tilde{x}_{i1}, \dots, \tilde{x}_{ip}, \tilde{y}_i); i = 1, \dots, n\}$ with $\tilde{x}_{i1} := x_{i1}/x_{ip}$, $\tilde{x}_{i2} := x_{i2}/x_{ip}$, \dots , $\tilde{x}_{ip} := x_{ip}/x_{ip} = 1$, and $\tilde{y}_i := y_i/x_{ip}$. Then

$$rdepth_0(\boldsymbol{\theta}, Z_n) = rdepth_0(\boldsymbol{\theta}, \tilde{Z}_n) = rdepth(\boldsymbol{\theta}, Z_n^1) \quad (3.13)$$

where the right hand side is the 'plain' rdepth (for regression with intercept) applied to $Z_n^1 = \{(\tilde{x}_{i1}, \dots, \tilde{x}_{i,p-1}, \tilde{y}_i); i = 1, \dots, n\} \subset \mathbb{R}^p$. Hence $rdepth_0$ has the same properties as rdepth.

Let us consider the special case of a regression line through the origin ($p = 1$), i.e. with $Z_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ where we want to fit y_i by θx_i . Since $x_i = x_{i1} = x_{ip}$ we find $Z_n^1 = \{\frac{y_i}{x_i}, \dots, \frac{y_n}{x_n}\} \subset \mathbb{R}$, hence the deepest regression is given by

$$\hat{\theta} = \text{med}_{i=1}^n \frac{y_i}{x_i}. \quad (3.14)$$

This estimator has minimax bias (Martin, Yohai and Zamar 1989). Note that $\hat{\theta}$ differs from the L^1 slope, which may even be the *highest* y_j/x_j if its $|x_j|$ is large enough.

4. DEPTH QUANTILES

For any number $0 < \tau < 1$, the τ -th regression quantile β_τ of Koenker and Bassett (1978) is defined as the hyperplane minimizing

$$2 \sum_{i=1}^n \{ \tau |r_i| 1(r_i \geq 0) + (1 - \tau) |r_i| 1(r_i < 0) \}. \quad (4.1)$$

Here all positive residuals receive weight 2τ , and all negative residuals have weight $2(1 - \tau)$. For $\tau = 0.5$ the objective (4.1) reduces to $\sum_{i=1}^n |r_i|$ hence $\beta_{0.5}$ is the least absolute values (L^1) regression.

Based on our formula (2.1), Roger Koenker (personal communication, 1997) proposed to extend the quantile idea to regression depth. In general, we define the τ -th depth quantile θ_τ as the hyperplane maximizing

$$2 \min_V (\min \{ \tau L^+(V) + (1 - \tau) R^-(V), \tau R^+(V) + (1 - \tau) L^-(V) \}) \quad (4.2)$$

which extends (3.2) by incorporating the weights 2τ and $2(1 - \tau)$. Clearly, $\theta_{0.5}$ reduces to the deepest regression. For any τ the maximum of (4.2) is reached at a hyperplane through p points, which yields a naive algorithm to compute θ_τ .

The L^1 -based quantiles β_τ and the depth-based quantiles θ_τ have several properties in common. Like the β_τ , also the θ_τ are \sqrt{n} -consistent estimators of the conditional τ -quantile of y given x , provided the latter function is linear (He and Portnoy 1998). This also explains why the depth quantiles are useful tools to detect and visualize heteroskedasticity.

Example 5: Faculty Salaries data. In Figure 7 we have plotted the salary of assistant professors versus the average salary of professors, at the top 50 universities of the Association of American Universities. This data set also comes from the DASL library. The depth quantiles θ_τ (for $\tau = 0.1, 0.2, \dots, 0.9$) clearly show the heteroskedasticity in the data.

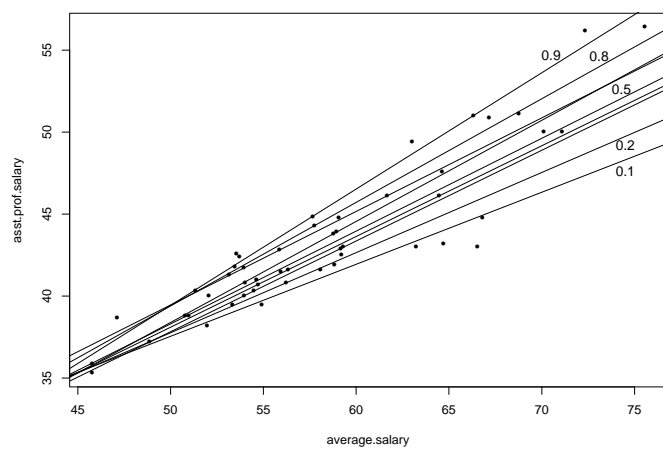


Figure 7. The depth quantiles of the Faculty Salaries data indicate heteroskedasticity.

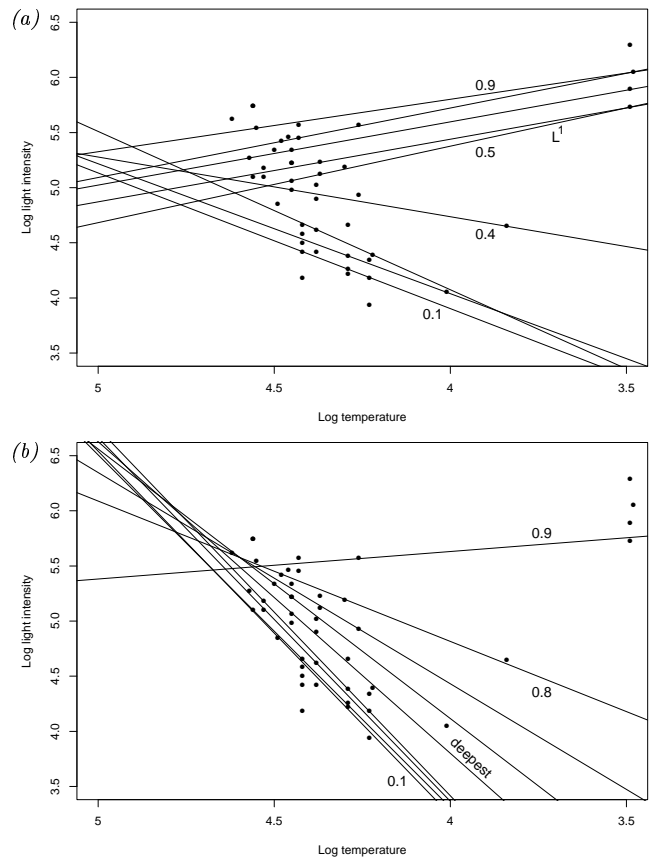


Figure 8. (a) Data of Figure 2, with the L^1 -based quantiles for $\tau = 0.1, 0.2, \dots, 0.9$ which are attracted by the leverage points. (b) The depth quantiles, of which only one is affected.

ity in the data.

For outlier-free data as in Example 5, the depth quantiles θ_τ are often close to the L^1 -based quantiles β_τ . However, the θ_τ have the additional advantage of being robust to leverage points. This creates new perspectives for data analysis, for instance in fields like economics where the utility of regression quantiles is widely recognized and where leverage points often occur.

Example 6: Stars data. Figure 8a shows the L^1 -based regression quantiles β_τ for the stars data of Figure 2, including the L^1 fit $\beta_{0.5}$. We see that $\beta_{0.5}, \beta_{0.6}, \beta_{0.7}, \beta_{0.8}$ and $\beta_{0.9}$ go through leverage points, but also $\beta_{0.1}, \beta_{0.2}, \beta_{0.3}$ and $\beta_{0.4}$ are attracted by them and do not fit the majority (the main sequence stars). By contrast, the depth quantiles θ_τ in Figure 8b are unaffected, except for $\theta_{0.9}$ (because over 10% of the data lie in the upper right direction).

5. THE DUAL PLOT

In a sense, the dual is simply the space of all potential fit vectors θ , so it can be seen as the parameter space. However, the dual space also represents the original data.

In simple regression, the axes of the dual plot are labelled θ_1 and θ_2 (instead of x and y). Dualization then transforms a line L given by $y = \theta_1 x + \theta_2$ to the point

$D(L) = (\theta_1, \theta_2)$ and transforms a point $\mathbf{z} = (x, y)$ to the set $D(\mathbf{z})$ of all fits (θ_1, θ_2) that pass through \mathbf{z} . Therefore, $D(\mathbf{z}_i)$ is the line given by $\theta_2 = -x_i\theta_1 + y_i$. The dual plot preserves the incidence and the ordering of lines and points, in the sense that a point \mathbf{z} lying below/on/above a line L in the primal plot corresponds to a line $D(\mathbf{z})$ below/through/above the point $D(L)$ in the dual plot. This property is important because $rdepth$ is based on signs of residuals. Also note that observations with the same x -value correspond to parallel lines in the dual plot.

Figure 9 is the dual plot of Figure 1 and thus contains 6 lines L_1, \dots, L_6 . No lines L_i are parallel since the x_i in Figure 1 are all distinct. The nonfits θ and η are now points in the dual plot. In fact, a point θ in the dual plot corresponds to a nonfit iff there exists a direction \mathbf{u} (with $\|\mathbf{u}\| = 1$) such that the halfline $[\theta, \theta + \mathbf{u} >$ does not intersect any L_i . Figure 9 shows such a halfline emanating from θ , indicated by an arrow. Here we adopt the convention that two parallel lines intersect at infinity, hence \mathbf{u} is restricted to the set of directions not parallel to any line L_i (this corresponds to the condition in Definition 2 that v_θ may not coincide with any x_i). Also note that whenever θ lies on a line L_i then any halfline $[\theta, \theta + \mathbf{u} >$ intersects L_i hence θ is not a nonfit (this corresponds to the case when $y_i = \theta_1 x_i + \theta_2$ hence $r_i = 0$). The set of all nonfits θ we call the *exterior* of the dual plot.

By definition, the $rdepth$ of a fit θ is (in dual space) the smallest number of lines L_i that need to be removed to set θ free (i.e., so that it lies in the exterior of the remaining arrangement of lines). Equivalently, it is the smallest number of lines intersected by any halfline $[\theta, \theta + \mathbf{u} >$. In Figure 9 we see that any halfline $[\xi, \xi + \mathbf{u} >$ intersects at least 2 lines, hence $rdepth(\xi, Z_n) = 2$.

This defines the depth of a point relative to an arrangement of lines, which is a new concept in the geometry literature. Also the dual version of our Theorem 1(a) is

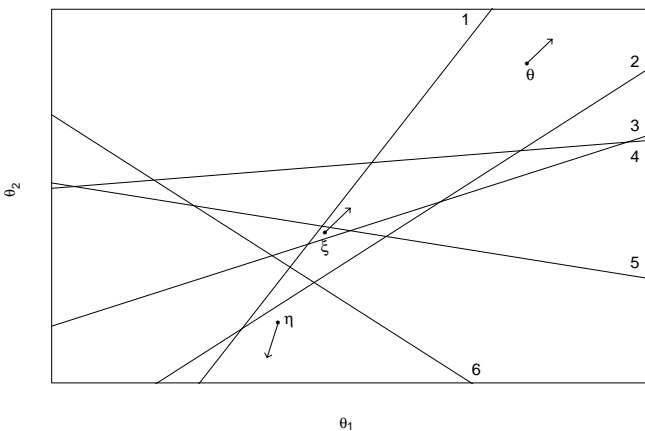


Figure 9. Dual plot of Figure 1, where θ_1 is the slope and θ_2 the intercept.

new:

Theorem 1'. For any set of n lines in the plane there exists a point θ which can only be set free by removing at least $\lceil n/3 \rceil$ lines.

For any collection I of lines we define the region $S(I)$ consisting of all points θ with $rdepth(\theta, I) \geq 1$. If $|I| = 1$ this is the line itself, and for $|I| = 2$ it is the union of both lines. In the typical case $|I| = 3$ it is the union of the three lines with the triangle they determine. The following result is also new:

Theorem 9. Any set of n lines in the plane can be partitioned into $k = \lceil n/3 \rceil$ subsets I_1, \dots, I_k such that $S(I_1), \dots, S(I_k)$ have a nonempty intersection.

In multiple regression the dual space is again the set of all possible fits θ , hence it is now p -dimensional. An observation $\mathbf{z}_i = (x_{i1}, \dots, x_{i,p-1}, y_i)$ is mapped to the set $D(\mathbf{z}_i)$ of all θ that pass through \mathbf{z}_i , so $D(\mathbf{z}_i)$ is the hyperplane H_i given by

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

(For $p = 3$, the dual space is \mathbb{R}^3 and each observation corresponds to a plane.) For any θ , we define $rdepth(\theta, Z_n)$ as the smallest number of hyperplanes H_i that need to be removed to set it free. Equivalently, we can search for the direction \mathbf{u} for which the halfline $[\theta, \theta + \mathbf{u} >$ intersects the fewest hyperplanes H_i . Conjecture 1(a) is thus equivalent to

Conjecture 1'. For any set of n hyperplanes in \mathbb{R}^p there exists a point θ which can only be set free by removing at least $\lceil n/(p+1) \rceil$ hyperplanes.

We also conjecture that Theorem 9 remains true in higher dimensions:

Conjecture 2. Any set of n hyperplanes in \mathbb{R}^p can be partitioned into $k = \lceil n/(p+1) \rceil$ subsets I_1, \dots, I_k such that $S(I_1), \dots, S(I_k)$ have a nonempty intersection.

This would imply Conjecture 1' because for any point θ in that intersection $rdepth(\theta, Z_n) \geq k$. If we were to replace hyperplanes by points, and $S(I_j)$ by the convex hull of I_j , then Theorem 9 and Conjecture 2 correspond to results of Birch (1959) and Tverberg (1966).

6. RELATION WITH LOCATION DEPTH

The general definition of depth (Definition 1) can be applied to the multivariate location setting as well. The data set X_n then consists of n observations $\mathbf{x}_i \in \mathbb{R}^p$, and a candidate fit θ is itself a p -variate point which should describe the position of the data cloud. We now call θ a nonfit for X_n iff θ lies outside the convex hull of X_n (for $p = 1$, the convex hull is just the interval spanned by the data). Equivalently, there exists an affine hyperplane through

θ with all observations strictly on one side and none on the other side. Tukey's (1975) halfspace depth is defined as the smallest number of data points contained in any closed halfspace of which the boundary passes through θ . Therefore, Definition 1 yields the halfspace depth, which we will call *location depth* and denote as $ldepth(\theta, X_n)$.

For bivariate data, $ldepth(\theta, X_n)$ can be computed in $O(n \log n)$ time with the algorithm of (Rousseeuw and Ruts 1996). In higher dimensions ($p \geq 3$), Rousseeuw and Struyf (1998) have constructed exact and approximate algorithms for $ldepth(\theta, X_n)$.

Remark 4. Under additional conditions, both $ldepth$ and $rdepth$ have interesting connections to earlier work. Under the (restrictive) assumptions that $p = 2$ and that the observations are iid according to a bivariate distribution which is centrosymmetric about some θ , Tukey's location depth $ldepth(\theta_0, X_n)$ corresponds to Hodges' (1955) test statistic for the null hypothesis $\theta = \theta_0$. For simple regression, in the special case that the data were generated by the linear model (2.11) with independent errors e_i drawn from a continuous distribution with zero median, the regression depth of some presupposed (θ_1^0, θ_2^0) evaluated at Z_n corresponds to a test of Daniels (1954) for the null hypothesis $(\theta_1, \theta_2) = (\theta_1^0, \theta_2^0)$. Daniels also computed the exact sampling distribution of the test statistic under the null hypothesis. In fact, the tests of Daniels and Hodges have the same null distribution (Hill 1960).

The following result concerns the maximal attainable depth for a given data set, and should be compared with Theorem 7 and Conjecture 1.

Theorem 10. (Rado 1946, Donoho and Gasko 1992.) At any data set $X_n \subset \mathbb{R}^p$ it holds that

$$\max_{\theta} ldepth(\theta, X_n) \geq \left\lceil \frac{n}{p+1} \right\rceil. \tag{6.1}$$

If X_n is in general position we also have

$$\max_{\theta} ldepth(\theta, X_n) \leq \left\lfloor \frac{n}{2} \right\rfloor. \tag{6.2}$$

Actually, the maximal depth at a given data set X_n depends on its shape. The upper bound (6.2) is reached at highly symmetric data sets, whereas the lower bound in (6.1) is attained at very asymmetric data sets. (In regression the maximal depth depends on how linear the data are: see Theorems 1 and 2.)

Similar results hold for a population distribution P on \mathbb{R}^p . Let us define $ldepth(\theta, P)$ as the smallest probability mass that needs to be removed to make θ a nonfit. Equivalently, it is the smallest mass in any closed halfspace with boundary through θ . In the population case, $ldepth$ lies between 0 and 1. Let us recall the notion of angular symmetry:

Definition 6. A distribution P on \mathbb{R}^p is angularly symmetric about θ iff $P(\theta + A) = P(\theta - A)$ for any cone A emanating from the origin.

Note that angular symmetry of P is weaker than the usual notion of centrosymmetry, in which A can be any measurable set. Angular symmetry of the empirical distribution means that for any observation \mathbf{x}_i in an open halfline $]\theta, \theta + \mathbf{u} >$ there must be an \mathbf{x}_j in $]\theta, \theta - \mathbf{u} >$, whereas centrosymmetry adds the requirement that $\|\mathbf{x}_i - \theta\| = \|\theta - \mathbf{x}_j\|$. In the following result, angular symmetry plays the same role as condition (2.5) in regression.

Theorem 11. (a) For any distribution P on \mathbb{R}^p with a density it holds that

$$\frac{1}{p+1} \leq \max_{\theta} ldepth(\theta, P) \leq \frac{1}{2}. \tag{6.3}$$

(b) If moreover P is angularly symmetric about some $\tilde{\theta}$, we find

$$\max_{\theta} ldepth(\theta, P) = ldepth(\tilde{\theta}, P) = \frac{1}{2}. \tag{6.4}$$

As in regression, the **deepest location** estimator is defined by

$$T_l^*(X_n) = \operatorname{argmax}_{\theta} ldepth(\theta, X_n) \tag{6.5}$$

and is sometimes called the *Tukey median*. For $p = 1$ it becomes the univariate sample median. For $p = 2$ the bivariate Tukey median can currently be computed in $O(n^2 \log^2 n)$ time with the algorithm constructed in (Rousseeuw and Ruts 1998).

Theorem 12. At any data set $X_n \subset \mathbb{R}^p$ it holds that

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

This lower bound on the breakdown value of T_l^* was already obtained by Donoho and Gasko (1992) under the additional assumption that X_n be in general position.

7. EXTENDING LIU'S SIMPLICIAL DEPTH TO REGRESSION

In Section 6 we saw that regression depth is similar to halfspace location depth. Here we will construct a regression counterpart to the simplicial location depth $ldepth^{(S)}(\theta, X_n)$ of Liu (1990), which is the proportion of data simplices containing θ . By analogy, we define $rdepth^{(S)}(\theta, Z_n)$ as the proportion of dual simplices containing θ :

Definition 7. The $rdepth^{(S)}$ of $\theta \in \mathbb{R}^p$ relative to $Z_n \subset \mathbb{R}^p$ is defined as

$$rdepth^{(S)}(\theta, Z_n) = \binom{n}{p+1}^{-1} \sum_{i_1 < \dots < i_{p+1}} I(\theta \in S(H_{i_1}, \dots, H_{i_{p+1}})). \tag{7.1}$$

Here, H_{i_1} is the hyperplane (in dual space) corresponding to the observation \mathbf{z}_{i_1} and S is the simplex determined

by $p + 1$ hyperplanes. Note that $\text{ldepth}^{(S)}$ and $\text{rdepth}^{(S)}$ do not fall under Definition 1, because they count $(p + 1)$ -tuples of observations whereas ldepth and rdepth count individual observations (at most n). The maximal $\text{rdepth}^{(S)}$ is given by:

Theorem 13. Under the conditions of Theorem 7(c) we have

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

The maximal values of rdepth and $\text{rdepth}^{(S)}$ are thus reached in the same situation. This again corresponds to the location setting, where the maximal values of ldepth and $\text{ldepth}^{(S)}$ are both reached under angular symmetry.

In simple regression with $x_1 < \dots < x_n$ we can rewrite (7.1) by the ‘primal’ formula

$$\text{rdepth}^{(S)}(\theta, Z_n) = \binom{n}{3}^{-1} \sum_{i < j < k} A(r_i(\theta), r_j(\theta), r_k(\theta)) \quad (7.2)$$

where $A(r_i, r_j, r_k)$ is 1 if the residuals r_i, r_j and r_k have alternating signs, and 0 otherwise. We then find a result similar to Theorems 2 and 3:

Theorem 14. If all the probability mass of H on \mathbb{R}^2 is concentrated on a strictly convex (or concave) curve, then the maximal $\text{rdepth}^{(S)}$ is $(\frac{1}{3})^3 = \frac{1}{27}$ and it is attained by the catline.

8. DISCUSSION

Let us compare the deepest regression T_r^* with the L^1 method. They extend two different definitions of the univariate median. The easiest of these is to order the observations y_i and to take the one(s) in the middle, an approach which dates back to the prehistory of statistics and may be older than the arithmetic mean. (Actually, this definition asks for the deepest location using univariate depth.) The second definition minimizes $\sum_{i=1}^n |y_i - \theta|$ and is much more recent, being proposed by Laplace in the eighteenth century. The second definition is less intuitive than the first, e.g. for seeing that the median is equivariant for monotone transformations of the y_i .

Extending Laplace’s definition to regression yields the L^1 method, which has been well-developed and studied. On the other hand, the oldest definition of the median was not generalized to linear regression before the present paper. The difficulty was to generalize ranks to higher dimensions, which is achieved by regression depth. We think that the deepest regression T_r^* is a more natural generalization of the median than L^1 , because T_r^* is based only on the ordering structure, which yields monotone equivariance and makes it suitable for the general semiparametric model including skewed error distributions and heteroskedasticity. Also, T_r^* and its associated regression quantiles are robust to leverage points, unlike L^1 and its quantiles. When the errors are gaussian and homoskedastic, T_r^* is only slightly less efficient than L^1 . The initial

algorithms for T_r^* take much more time than the current algorithms for L^1 , but of course the latter have been developed and refined for many years.

The combinatorial ordering structure of a regression data set can be described in the dual (Section 5). For an arrangement of hyperplanes in \mathbb{R}^p the ordering structure is characterized by Grünbaum’s incidence graph, as explained in Edelsbrunner (1987, Chapter 7). Let us now consider continuous transformations of the data which preserve that incidence graph. Then the deepest regression T_r^* is *equivariant for all such order-preserving transformations* (this property generalizes the monotone equivariance of the univariate median). A consequence of this equivariance is that there are configurations at which the breakdown value of T_r^* is $1/(p + 1)$. This happens for all estimators that depend only on the ordering structure, because there are configurations consisting of $p + 1$ subsets of observations such that (a) there is an order-preserving transformation which maps each subset to the next, and (b) one of these subsets can be sent arbitrarily far away by an order-preserving transformation which keeps the other data fixed.

From the viewpoint of the breakdown value, methods depending only on the ordering structure cannot compete with positive-breakdown methods such as least median of squares (LMS) and least trimmed squares (LTS) regression (Rousseeuw 1984). When the ‘good’ data are generated by the model, T_r^* attains a breakdown value of $1/3$, whereas LMS, LTS and their offspring (such as S-estimators) always attain 50%. This is because the latter methods also use the *size* of the residuals, which is more than the ordering structure. But on the other hand, LMS/LTS and their relatives are not equivariant for monotone transformations of the y_i or for order-preserving transformations. Unlike the deepest regression T_r^* , they are not consistent for the full semiparametric model with skewness and heteroskedasticity. (Note that the LMS does not generalize the univariate median: for univariate data it becomes the midpoint of the shortest half, which is similar to estimating a mode.) Therefore, neither T_r^* or LMS can replace the other. In applications where robustness and outlier detection are the most important (e.g. in computer vision) one needs to use LMS or its relatives. But in applications where there are not so many outliers (this can be checked with LMS/LTS/S beforehand) and the emphasis is on monotonicity and the possibility of skewness and heteroskedasticity, the deepest regression T_r^* is the natural choice.

In the setting of multivariate location the situation is similar. The L^1 estimator is the location θ minimizing $\sum_{i=1}^n \|\mathbf{x}_i - \theta\|$, and has 50% breakdown but is not equivariant for linear transformations. On the other hand, the deepest location T_l^* is equivariant for linear transformations, and is suitable for the semiparametric model with angularly symmetric distributions. For a configuration of n points in \mathbb{R}^p the ordering structure is characterized by the n^p matrix of Goodman and Pollack (1983). Since the deepest location T_l^* only depends on this ordering structure, it is equivariant for all order-preserving transformations. Consequently, its breakdown value is $1/(p + 1)$ at certain

configurations, but for data generated from the model it becomes $1/3$. Also the deepest location T_l^* cannot replace typical positive-breakdown methods such as the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators of (Rousseeuw 1984), but it is perfectly suited for the large semiparametric model.

9. FURTHER DEVELOPMENTS

Several issues raised in this paper offer the opportunity to collaborate with specialists of, among other things, rank statistics and computational geometry (e.g., for Conjecture 1').

After we circulated the original version of this paper in November 1996, the asymptotics of the deepest regression line were derived by He and Portnoy (1998). Under the model (2.11) where the errors e_i are independent and have a distribution with zero median, they prove that the deepest regression T_r^* is a consistent estimator of the conditional median line $y = \theta_1 x + \theta_2$. Moreover, the limiting distribution of $\sqrt{n}(T_r^* - \theta)$ is nearly gaussian. The resulting relative asymptotic efficiencies (obtained numerically) are close to our simulation results for $n = 500$. The limiting distributions of T_r^* and T_l^* should be studied further, especially their second moment and the comparison with a gaussian distribution.

Location depth has already seen quite a few applications, e.g. to a new type of quality index and to limiting p -values (Liu and Singh 1993, 1997), and to multivariate control charts (Liu 1995). We are currently looking at some applications of regression depth, e.g. to enzyme kinetics and economics. It would also be interesting to know whether the results of Dümbgen (1992) can be extended to $rdepth^{(S)}$.

The introduction of a general notion of depth in Definition 1, of which both regression depth and location depth are special cases, also invites the extension of depth to other statistical settings such as the estimation of scale and scatter matrices, nonlinear regression, and so on.

APPENDIX

Proof of Theorem 1

The inequality (2.3) follows directly from Theorem 5 with $k \leq 2$. In order to establish the lower bound in (2.2) we first note that the catline of Section 2.3 always exists, since the ham-sandwich theorem (see, e.g., Edelsbrunner 1987, page 69) ensures the existence of a line that simultaneously bisects $P_1 = L \cup M$ and $P_2 = M \cup R$. We now have to prove that its $rdepth$ is at least $\lceil n/3 \rceil$. Consider a tilt point v in L , hence to the left of $P_2 = M \cup R$. Denote by P_2^+ (resp. P_2^-, P_2^0) the number of strictly positive (resp. negative, zero) residuals in P_2 . From the definition of the catline, it follows that $P_2^+ \leq \lceil n/3 \rceil$, $P_2^- \leq \lceil n/3 \rceil$, $P_2^+ + P_2^- + P_2^0 = 2\lceil n/3 \rceil$ if $n = 3m$, and $P_2^+ + P_2^- + P_2^0 = 2\lceil n/3 \rceil + 1$ if $n \neq 3m$. Let us now tilt the line upward (downward) about v until it becomes vertical. Doing so it passes all points of P_2 with positive (negative) residuals, hence at least $\min(P_2^+ + P_2^-, P_2^- + P_2^0) \geq \lceil n/3 \rceil$ points. For a tilt point v in M or R the reasoning is analogous, hence

$rdepth(\theta_{CAT}, Z_n) \geq \lceil n/3 \rceil$. In the population case the proof is similar, using the scheme (2.7).

Proof of Theorem 2

(a) A line with maximal depth passes through two observations (or it could be slightly tilted or shifted until it does fit 2 points), thereby dividing the observations in three groups with alternating residual signs. The $rdepth$ of this line is then $2 +$ the size of the smallest group, which is bounded by $2 + \lceil (n-2)/3 \rceil = \lceil (n+2)/3 \rceil$. Part (b) is analogous.

Proof of Theorem 3

Assume the observations are ordered by their x -coordinates and set $m = \lceil n/3 \rceil$, then the line through the observations \mathbf{z}_{m+1} and \mathbf{z}_{2m+1} if $n \neq 3m+2$ [resp. through \mathbf{z}_{m+1} and \mathbf{z}_{2m+2} if $n = 3m+2$] has maximal depth and is also the catline.

Proof of Theorem 4

As in the proof of the Corollary of Conjecture 1, with $p = 2$.

Proof of Statement (2.13)

Take any $(x, y) \in E_k$. Then $y = \lambda(\theta_1^{(1)}x + \theta_2^{(1)}) + (1-\lambda)(\theta_1^{(2)}x + \theta_2^{(2)})$ for some $0 \leq \lambda \leq 1$ and $rdepth(\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}), Z_n) \geq k$ for $i = 1, 2$. Thus $\theta^{(i)} \in \text{conv}\{\theta; rdepth(\theta, Z_n) \geq k\} = \text{conv}\{\theta^{ij}; rdepth(\theta^{ij}, Z_n) \geq k\}$. (Here, 'conv' stands for the convex hull). Finally it is easy to verify that $y \in \text{conv}\{\theta_1^{ij}x + \theta_2^{ij}; rdepth(\theta^{ij}, Z_n) \geq k\}$ which equals the interval between the minimum and maximum of these $n(n-1)/2$ values.

Proof of Theorem 5

Consider the data set Z_{n-k} obtained by removing these k observations from Z_n . In the dual space (see Section 5) this means that no hyperplane contains θ . Therefore, for any direction \mathbf{u} with $\|\mathbf{u}\| = 1$ it holds that $\#\{\text{intersections of } [\theta, \theta + \mathbf{u} > \text{ with hyperplanes}\} + \#\{\text{intersections of } [\theta, \theta - \mathbf{u} > \text{ with hyperplanes}\} = n - k$, hence

$$0 \leq rdepth(\theta, Z_{n-k}) \leq \left\lceil \frac{n-k}{2} \right\rceil.$$

Reinserting the k observations (hyperplanes) yields

$$0 + k \leq rdepth(\theta, Z_n) \leq \left\lceil \frac{n-k}{2} \right\rceil + k = \left\lceil \frac{n+k}{2} \right\rceil.$$

Proof of Illustration 2

Assume that all residuals differ from zero, otherwise the result is trivial. Define $I = \{i; r_i > 0\}$ and $J = \{i; r_i < 0\}$, then one of them (say, I) contains at least p observations. Put $c = \sum_{i \in I} r_i > 0$. Since the LS estimates satisfy the normal equations, we have $\sum_{i=1}^n r_i = 0$ and thus $c = -\sum_{j \in J} r_j$. Since $\sum_{i=1}^n r_i \mathbf{x}_i = 0$ we have $\mathbf{x}^* := \sum_{i \in I} (r_i/c) \mathbf{x}_i = \sum_{j \in J} (-r_j/c) \mathbf{x}_j = (\sum_{i=1}^n |r_i| \mathbf{x}_i) / (\sum_{i=1}^n |r_i|)$. Set $\lambda_i = r_i/c$ and $\gamma_j = -r_j/c$. Then \mathbf{x}^* belongs to the interior of the convex hull of $A = \{\mathbf{x}_i, i \in I\}$ and to the convex hull of $B = \{\mathbf{x}_j, j \in J\}$. Hence, there can be no hyperplane in \mathbf{x} -space that separates A from B . Therefore θ_{LS} is not a nonfit. For an example where $rdepth(\theta_{LS}, Z_n) = 1$ take $\mathbf{z}_1 = (0, 0, 1)$ and $\mathbf{z}_i = (x_i, y_i, 0)$ for $2 \leq i \leq n$, where the

(x_i, y_i) are evenly spaced on the unit circle in \mathbb{R}^2 .

Proof of Theorem 6

Let V be an affine hyperplane in \mathbf{x} -space, and \bar{V} the vertical hyperplane in \mathbb{R}^p through V . Further let $u = \pm 1$ determine a direction in which to tilt θ until it becomes vertical. For each V and u we then define $A_{\theta, V, u}$ as the double wedge formed by tilting θ around $\bar{V} \cap \theta$ in the direction of u until the fit becomes vertical. Further denote by H_n the empirical distribution of the observed data Z_n . Define

$$\delta_H(H_n, H) = \sup_{\theta, V, u} |H_n(A_{\theta, V, u}) - H(A_{\theta, V, u})|.$$

If Z_n is sampled from H it holds that

$$\delta_H(H_n, H) \xrightarrow[n \rightarrow \infty]{a.s.} 0.$$

This follows from the generalization of the Glivenko-Cantelli theorem as formulated and proved in (Pollard 1984, Theorem 14 and Lemma 18), and the fact that $A_{\theta, V, u}$ can be constructed by taking a finite number of unions and intersections of half-spaces. Define now $\Pi(\theta) = \inf_{V, u} H(A_{\theta, V, u})$ and its empirical version $\Pi_n(\theta) = \inf_{V, u} H_n(A_{\theta, V, u})$. It is clear that $\Pi(\theta) = rdepth(\theta, H)$ and $\Pi_n(\theta) = rdepth(\theta, Z_n)/n$. Moreover, we have that $|\Pi_n(\theta) - \Pi(\theta)| \leq \delta_H(H_n, H) \rightarrow 0$ which implies (3.4).

Proof of Theorem 7

The inequality in (a) follows from Theorem 5 with $k \leq p$. The population cases (b) and (c) are analogous.

Proof of Remark 3

For simplicity of notation, we will write down the proof for $p = 3$. However, the construction remains valid in higher dimensions. Take any 3 points on the curve, denoted by p_1, p_2 and p_3 , and denote by θ the plane through these points. We will now show that $rdepth(\theta, H) \leq 1/4$. The plane θ splits the probability mass on the curve in 4 disjoint regions A, B, C and D such that the curve lies alternately below and above θ . Consider first the line $p_1 p_2$ through the points p_1 and p_2 . If we tilt θ around this line, we have to pass region D in one direction, and $A \cup B \cup C$ if we turn in the other direction. Thus $rdepth(\theta, H) \leq \min(H(D), H(A \cup B \cup C))$. Tilting around the line $p_2 p_3$ yields $rdepth(\theta, H) \leq \min(H(A), H(B \cup C \cup D))$. Then consider the line through p_1 parallel to $p_2 p_3$, yielding $rdepth(\theta, H) \leq \min(H(C), H(A \cup B \cup D))$. Finally, taking the line through p_3 parallel to $p_1 p_2$ yields $rdepth(\theta, H) \leq \min(H(B), H(A \cup C \cup D))$. Summarizing, $rdepth(\theta, H) \leq \min(H(A), H(B), H(C), H(D)) \leq 1/4$ since A, B, C and D partition the mass of H . Finally, we can consider the plane θ that splits the probability mass on the curve in 4 disjoint regions A, B, C and D with an equal probability (of $1/4$). It is clear that this plane satisfies $rdepth(\theta, H) = 1/4$.

Proof of Corollary of Conjecture 1

Replacing $m < \lceil n/(p+1) \rceil - p + 1$ data points of Z_n yields a new sample Z'_n . Conjecture 1 says that $rdepth(T_r^*(Z'_n), Z'_n) \geq \lceil n/(p+1) \rceil \geq m + p$. Since Z_n can be obtained by replacing m points of Z'_n it follows that $rdepth(T_r^*(Z'_n), Z_n) \geq p$. Hence

$T_r^*(Z'_n) \in \{\theta; rdepth(\theta, Z_n) \geq p\}$, which is bounded since the \mathbf{x}_i are in general position.

Proof of Theorem 8

See Van Aelst et al. (1999).

Proof of Theorem 9

Take $\theta = \theta_{CAT}$. If $n = 3k$ we can partition Z_n into k triplets (z_i, z_j, z_h) with $z_i \in L, z_j \in M, z_h \in R$ and either $r_i \geq 0, r_j \leq 0$, and $r_h \geq 0$ or $r_i \leq 0, r_j \geq 0$, and $r_h \leq 0$. If $n = 3k - 1$ or $n = 3k - 2$ both $|L \cup M|$ and $|M \cup R|$ are odd, hence at least one residual is zero. We can then isolate the zero residual(s) in some I_h with $|I_h| \leq 2$ and partition the remaining points. In each of these situations, θ belongs to $\cap_j S(I_j)$.

Proof of Theorem 11

See Donoho and Gasko (1992, page 1818). The generalization from centrosymmetry to angular symmetry is straightforward.

Proof of Theorem 12

This is similar to the proof of the Corollary of Conjecture 1. Now we replace $m < \lceil n/(p+1) \rceil$ data points to obtain a contaminated sample X'_n . From $ldepth(T_l^*(X'_n), X'_n) \geq \lceil n/(p+1) \rceil \geq m + 1$ it follows that $ldepth(T_l^*(X'_n), X_n) \geq 1$. Therefore $T_l^*(X'_n)$ lies in $\text{conv}(X_n)$ which is bounded.

REFERENCES

- Birch, B.J. (1959), "On 3N Points in a Plane," *Proceedings of the Cambridge Philosophical Society*, 55, 289-293.
- Carroll, R.J., and Ruppert, D. (1988), *Transformation and Weighting in Regression*, New York: Chapman and Hall.
- Daniels, H.E. (1954), "A Distribution-Free Test for Regression Parameters," *Annals of Mathematical Statistics*, 25, 499-513.
- 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.
- Dümbgen, L. (1992), "Limit Theorems for Simplicial Depth," *Statistics and Probability Letters*, 14, 119-128.
- Eddy, W.F. (1985), "Ordering of Multivariate Data," in *Computer Science and Statistics: Proceedings of the 16th Symposium on the Interface*, ed. L. Billard, Amsterdam: North-Holland, 25-30.
- Edelsbrunner, H. (1987), *Algorithms in Combinatorial Geometry*, Berlin, Springer-Verlag.
- Goodman, J.E., and Pollack, R. (1983), "Multidimensional sorting," *SIAM Journal on Computing*, 12, 484-507.
- Green, P.J. (1981), "Peeling Bivariate Data," in *Interpreting Multivariate Data*, ed. V. Barnett, New York: John Wiley, 3-19.
- He, X., and Portnoy, S. (1998), "Asymptotics of the Deepest Line," in *Statistical Inference and Related Topics: A Festschrift in Honor of A.K.Md.E. Saleh*, New York: Nova Science Publications Inc. (to appear).
- Hill, B.M. (1960), "A Relationship Between Hodges' Bivariate Sign Test and a Nonparametric Test of Daniels," *Annals of Mathematical Statistics*, 31, 1190-1192.
- Hodges, J.L. Jr. (1955), "A Bivariate Sign Test," *Annals of Mathematical Statistics*, 26, 523-527.
- Hubert, M., and Rousseeuw, P.J. (1998), "The Catline for Deep Regression," *Journal of Multivariate Analysis*, 66, 270-296.
- Koenker, R., and Bassett, G.J. (1978), "Regression Quantiles," *Econometrica*, 46, 33-50.
- Liu, R.Y. (1990), "On a Notion of Data Depth Based on Random

- Simplices," *The Annals of Statistics*, 18, 405-414.
- Liu, R.Y. (1995), "Control Charts for Multivariate Processes," *Journal of the American Statistical Association*, 90, 1380-1387.
- Liu, R.Y., and Singh, K. (1993), "A Quality Index based on Data Depth and Multivariate Rank Tests," *Journal of the American Statistical Association*, 88, 252-260.
- Liu, R.Y., and Singh, K. (1997), "Notions of Limiting P-values Based on Data Depth and Bootstrap," *Journal of the American Statistical Association*, 92, 266-277.
- Martin, R.D., Yohai, V.J., and Zamar, R.H. (1989), "Min-max Bias Robust Regression," *The Annals of Statistics*, 17, 1608-1630.
- Rado, R. (1946), "A Theorem on General Measure," *Journal of the London Mathematical Society*, 21, 291-300.
- Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871-880.
- 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 Struyf, A. (1998), "Computing Location Depth and Regression Depth in Higher Dimensions," *Statistics and Computing*, 8, 193-203.
- Tukey, J.W. (1975), "Mathematics and the Picturing of Data," *Proceedings of the International Congress of Mathematicians*, Vancouver, 2, 523-531.
- Tverberg, H. (1966), "A Generalization of Radon's Theorem," *Journal of the London Mathematical Society*, 41, 123-128.
- Van Aelst, S., Hubert, M., Struyf, A., and Rousseeuw, P.J. (1999), "Applications of Deepest Regression," in preparation.