

# The Deepest Regression Method

Stefan Van Aelst<sup>1</sup>, Peter J. Rousseeuw, Mia Hubert<sup>2</sup>, and Anja Struyf<sup>1</sup>

Revised version, August 11, 2000

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

<http://win-www.uia.ac.be/u/statist/>

## Abstract

Deepest regression (*DR*) is a method for linear regression introduced by Rousseeuw and Hubert [20]. The *DR* method is defined as the fit with largest regression depth relative to the data. In this paper we show that *DR* is a robust method, with breakdown value that converges almost surely to  $1/3$  in any dimension. We construct an approximate algorithm for fast computation of *DR* in more than two dimensions. From the distribution of the regression depth we derive tests for the true unknown parameters in the linear regression model. Moreover, we construct simultaneous confidence regions based on bootstrapped estimates. We also use the maximal regression depth to construct a test for linearity versus convexity/concavity. We extend regression depth and deepest regression to more general models. We apply *DR* to polynomial regression, and show that the deepest polynomial regression has breakdown value  $1/3$ . Finally, *DR* is applied to the Michaelis-Menten model of enzyme kinetics, where it resolves a long-standing ambiguity.

## 1 Introduction

Consider a dataset  $Z_n = \{\mathbf{z}_i = (x_{i1}, \dots, x_{i,p-1}, y_i); i = 1, \dots, n\} \subset \mathbb{R}^p$ . In linear regression we want to fit a hyperplane of the form  $y = \theta_1 x_1 + \dots + \theta_{p-1} x_{p-1} + \theta_p$  with  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^t \in \mathbb{R}^p$ . We denote the  $x$ -part of each data point  $\mathbf{z}_i$  by  $\mathbf{x}_i = (x_{i1}, \dots, x_{i,p-1})^t \in \mathbb{R}^{p-1}$ . The residuals of  $Z_n$  relative to the fit  $\boldsymbol{\theta}$  are denoted as  $r_i = r_i(\boldsymbol{\theta}) = y_i - \theta_1 x_{i1} - \dots - \theta_{p-1} x_{i,p-1} -$

---

<sup>1</sup>Research Assistant with the FWO, Belgium.

<sup>2</sup>Postdoctoral Fellow at the FWO, Belgium.

$\theta_p$ . To measure the quality of a fit, Rousseeuw and Hubert [20] introduced the notion of *regression depth*.

**Definition 1.** *The regression depth of a candidate fit  $\boldsymbol{\theta} \in \mathbb{R}^p$  relative to a dataset  $Z_n \subset \mathbb{R}^p$  is given by*

$$rdepth(\boldsymbol{\theta}, Z_n) = \min_{\mathbf{u}, v} \{ \#(r_i(\boldsymbol{\theta}) \geq 0 \text{ and } \mathbf{x}_i^t \mathbf{u} < v) + \#(r_i(\boldsymbol{\theta}) \leq 0 \text{ and } \mathbf{x}_i^t \mathbf{u} > v) \} \quad (1)$$

where the minimum is over all unit vectors  $\mathbf{u} = (u_1, \dots, u_{p-1})^t \in \mathbb{R}^{p-1}$  and all  $v \in \mathbb{R}$  with  $\mathbf{x}_i^t \mathbf{u} \neq v$  for all  $(\mathbf{x}_i^t, y_i) \in Z_n$ .

The regression depth of a fit  $\boldsymbol{\theta} \in \mathbb{R}^p$  relative to the dataset  $Z_n \subset \mathbb{R}^p$  is thus the smallest number of observations that need to be passed when tilting  $\boldsymbol{\theta}$  until it becomes vertical. Therefore, we always have  $0 \leq rdepth(\boldsymbol{\theta}, Z_n) \leq n$ .

In the special case of  $p = 1$  there are no  $\mathbf{x}$ -values, and  $Z_n$  is a univariate dataset. For any  $\theta \in \mathbb{R}$  we then have  $rdepth(\theta, Z_n) = \min(\#\{y_i \geq \theta\}, \#\{y_i \leq \theta\})$  which is the 'rank' of  $\theta$  when we rank from the outside inwards. For any  $p \geq 1$ , the regression depth of  $\boldsymbol{\theta}$  measures how balanced the dataset  $Z_n$  is about the linear fit determined by  $\boldsymbol{\theta}$ . It can easily be verified that regression depth is scale invariant, regression invariant, and affine invariant according to the definitions in Rousseeuw and Leroy ([21, page 116]).

Based on the notion of regression depth, Rousseeuw and Hubert [20] introduced the deepest regression estimator (*DR*) for robust linear regression. In Section 2 we give the definition of *DR* and its basic properties. We show that *DR* is a robust method with breakdown value that converges almost surely to  $1/3$  in any dimension  $p$  ( $p \geq 2$ ), when the good data come from a large semiparametric model. Section 3 proposes the fast approximate algorithm MEDSWEEP to compute *DR* in higher dimensions ( $p \geq 3$ ). Based on the distribution of the regression depth function, inference for the parameters is derived in Section 4. Tests and confidence regions for the true unknown parameters  $\theta_1, \dots, \theta_p$  are constructed. We also propose a test for linearity versus convexity of the dataset  $Z_n$  based on the maximal depth of  $Z_n$ . Applications of deepest regression to specific models are given in Section 5. First we give the definition of regression depth for more general models and show that it is monotone invariant. For general linear models we then compute the deepest regression according to this definition, and derive a monotone equivariance property. In the case of polynomial regression, we show that the deepest polynomial regression always has breakdown value at least  $1/3$ . We also apply the deepest regression to the Michaelis-Menten model, where it

provides a solution to the problem of ambiguous results obtained from the two commonly used parametrizations.

## 2 Definition and properties of deepest regression

**Definition 2.** *In  $p$  dimensions the deepest regression estimator  $DR(Z_n)$  is defined as the fit  $\boldsymbol{\theta}$  with maximal  $rdepth(\boldsymbol{\theta}, Z_n)$ , that is*

$$DR(Z_n) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} rdepth(\boldsymbol{\theta}, Z_n). \quad (2)$$

Since the regression depth of a fit  $\boldsymbol{\theta}$  can only increase if we slightly tilt the fit until it passes through  $p$  observations (while not passing any other observations), it suffices to consider all fits through  $p$  data points in Definition 2. If several of these fits are ‘tied’ in the sense that they have the same (maximal) regression depth, we take their average. Note that the average does not necessarily have the maximal depth (see Mizera and Volauf [16] for a related discussion), but it has several other advantages. First, the  $DR$  is now uniquely defined, and its finite-sample efficiency is higher than the efficiency of one of the fits with maximal depth. Moreover, taking the average does not change the robustness as will be shown in Theorem 1.

Note that no distributional assumptions are made to define the deepest regression estimator of a dataset. The  $DR$  is a regression, scale, and affine equivariant estimator. For a univariate dataset, the deepest regression is its median. The  $DR$  thus generalizes the univariate median to linear regression.

In the population case, let  $(\mathbf{x}^t, y)$  be a random  $p$ -dimensional variable, with distribution  $H$  on  $\mathbb{R}^p$ . Then  $rdepth(\boldsymbol{\theta}, H)$  is defined as the smallest amount of probability mass that needs to be passed when tilting  $\boldsymbol{\theta}$  in any way until it is vertical. The deepest regression  $DR(H)$  is the fit  $\boldsymbol{\theta}$  with maximal depth. The natural setting of deepest regression is a large semiparametric model  $\mathcal{H}$  in which the functional form is parametric and the error distribution is nonparametric. Formally,  $\mathcal{H}$  consists of all distributions  $H$  on  $\mathbb{R}^p$  that satisfy the following conditions:

$H$  has a strictly positive density and there exists a  $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^p$  with

$$\operatorname{med}_H(y | \mathbf{x}) = (\mathbf{x}^t, 1)\tilde{\boldsymbol{\theta}}. \quad (\text{H})$$

Note that this model allows for skewed error distributions and heteroscedasticity. Van Aelst and Rousseeuw [26] have shown that the  $DR$  is a Fisher-consistent estimator of  $\tilde{\boldsymbol{\theta}}$  when the

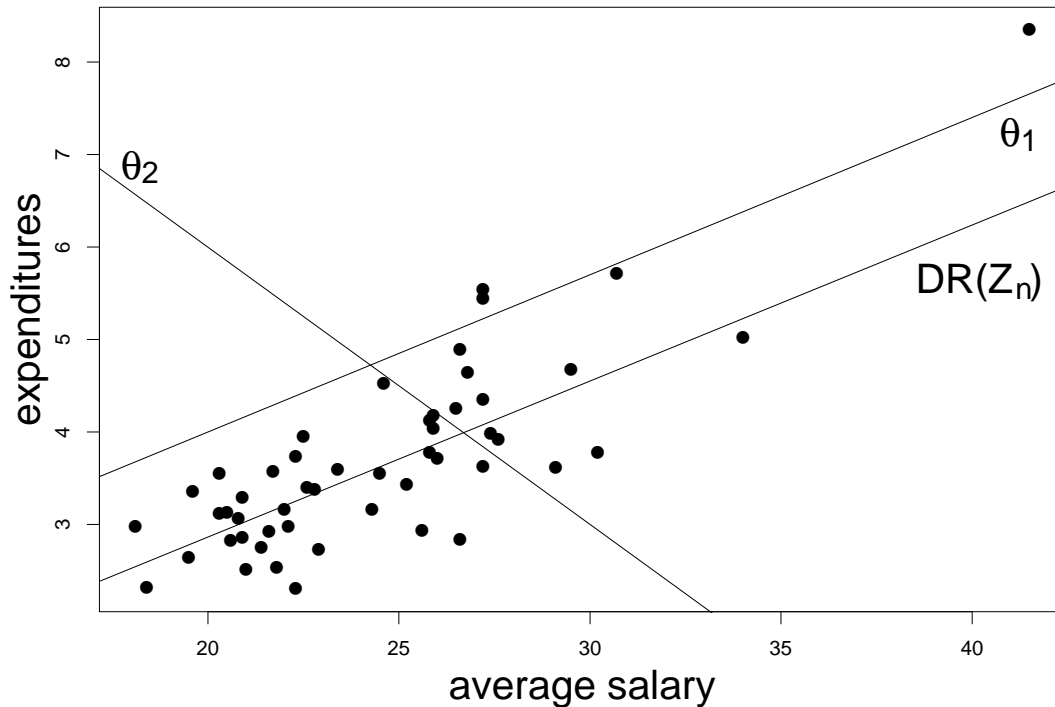


Figure 1: Educational spending data, with  $n = 51$  observations in  $p = 2$  dimensions. The lines  $\theta_1$  and  $\theta_2$  both have depth 2, and the deepest regression  $DR(Z_n)$  is the average of fits with depth 23.

data come from the natural semiparametric model  $\mathcal{H}$ . The asymptotic distribution of the deepest regression was obtained by He and Portnoy [11] in simple regression, and by Bai and He [2] in multiple regression.

Figure 1 shows the Educational Spending data, obtained from the DASL library at <http://lib.stat.cmu.edu/DASL>. This dataset lists the expenditures per pupil versus the average salary paid to teachers, for  $n = 51$  regions in the US. The fits  $\theta_1 = (0.17, 0.6)^t$  and  $\theta_2 = (-0.3, 12)^t$  both have regression depth 2, and the deepest regression  $DR(Z_n) = (0.17, -0.51)^t$  is the average of fits with depth 23. Figure 1 illustrates that lines with high regression depth fit the data better than lines with low depth. The regression depth thus measures the quality of a fit, which motivates our interest in the deepest regression  $DR(Z_n)$ .

We define the finite-sample addition breakdown value  $\varepsilon_n^*$  of an estimator  $T_n$  as the smallest fraction of contamination that can be added to any dataset  $Z_n$  such that  $T_n$  becomes useless (see also Donoho and Gasko [7]). Let us consider an actual dataset  $Z_n$ . Denote by  $Z_{n+m}$  the dataset formed by adding  $m$  observations to  $Z_n$ . Then the breakdown value is

defined as

$$\varepsilon_n^*(T_n, Z_n) = \min\left\{\frac{m}{n+m}; \sup_{Z_{n+m}} \|T_{n+m}(Z_{n+m}) - T_n(Z_n)\| = \infty\right\}.$$

The addition breakdown value defined here is closely related to the replacement breakdown value (see Donoho and Huber [8]). Quantitative relationships allowing one to obtain the replacement breakdown value from the addition breakdown value are given by Zuo [29].

The breakdown value of the deepest regression is always positive, but it can be as low as  $1/(p+1)$  when the original data are themselves peculiar (Rousseeuw and Hubert [20]). Fortunately, it turns out that if the original data are drawn from the model, then the breakdown value converges almost surely to  $1/3$  in any dimension  $p$  ( $p \geq 2$ ).

**Theorem 1.** *Let  $Z_n = \{(\mathbf{x}_1^t, y_1), \dots, (\mathbf{x}_n^t, y_n)\}$  be a sample from a distribution  $H$  on  $\mathbb{R}^p$  ( $p \geq 2$ ) with  $H \in \mathcal{H}$ . Then*

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

(All proofs are given in the Appendix.) Theorem 1 says that the deepest regression does not break down when at least 67% of the data are generated from the semiparametric model  $\mathcal{H}$  while the remaining data (i.e., up to 33% of the points) may be anything. This result holds in any dimension. The  $DR$  is thus robust to leverage points as well as to vertical outliers. Moreover, Theorem 1 illustrates that the deepest regression is different from  $L^1$  regression, which is defined as  $L^1(Z_n) = \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^n |r_i(\boldsymbol{\theta})|$ . Note that  $L^1$  is another generalization of the univariate median to regression, but with zero breakdown value due to its vulnerability to leverage points.

In simple regression, Van Aelst and Rousseeuw [26] derived the influence function of the  $DR$  for elliptical distributions and computed the corresponding sensitivity functions. The influence functions of the  $DR$  slope and intercept are piecewise smooth and bounded, meaning that an outlier cannot affect  $DR$  too much, and the corresponding sensitivity functions show that this already holds for small sample sizes.

### 3 Computation

In  $p = 2$  dimensions the regression depth can be computed in  $O(n \log n)$  time with the algorithm described in (Rousseeuw and Hubert [20]). To compute the regression depth of a fit in  $p = 3$  or  $p = 4$  dimensions, Rousseeuw and Struyf [22] constructed exact algorithms

with time complexity  $O(n^{p-1} \log n)$ . For datasets with large  $n$  and/or  $p$  they also give an approximate algorithm that computes the regression depth of a fit in  $O(mp^3 + mpn + mn \log n)$  time. Here  $m$  is the number of  $(p-1)$ -subsets in  $x$ -space used in the algorithm. The algorithm is exact when all  $m = \binom{n}{p-1}$  such subsets are considered.

A naive exact algorithm for the deepest regression computes the regression depth of all  $O(n^p)$  fits through  $p$  observations and keeps the one(s) with maximal depth. This yields a total time complexity of  $O(n^{2p-1} \log n)$  which is very slow for large  $n$  and/or high  $p$ . Even if we use the approximate algorithm of Rousseeuw and Struyf [22] to compute the depth of each fit, the time complexity remains very high. For simple regression, researchers in computational geometry have obtained exact algorithms of complexity  $O(n \log^2 n)$  (van Kreveld et al. [28]) and even  $O(n \log n)$  (Langerman and Steiger [12]), i.e. little more than linear time. To speed up the computation in more than two dimensions, we will now construct the fast algorithm MEDSWEEP to approximate the deepest regression.

The MEDSWEEP algorithm is based on regression through the origin. For regression through the origin, Rousseeuw and Hubert [20] defined the regression depth (denoted as  $\text{rdepth}_0$ ) by requiring that  $v = 0$  in Definition 1. Therefore, the  $\text{rdepth}_0(\boldsymbol{\theta})$  of a fit  $\boldsymbol{\theta} \in \mathbb{R}^p$  relative to a dataset  $Z_n \subset \mathbb{R}^{p+1}$  is again the smallest number of observations that needs to be passed when tilting  $\boldsymbol{\theta}$  in any way until it becomes vertical. Rousseeuw and Hubert [20] have shown that in the special case of a regression line through the origin ( $p = 1$ ), the deepest regression ( $DR_o$ ) of the dataset  $Z_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$  is given by the slope

$$DR_o(Z_n) = \text{med}_{i=1}^n \frac{y_i}{x_i} \quad (4)$$

where observations with  $x_i = 0$  are not used. This estimator has minimax bias (Martin, Yohai and Zamar [14]) and can be computed in  $O(n)$  time.

We propose a sweeping method based on the estimator (4) to approximate the deepest regression in higher dimensions. Suppose we have a dataset  $Z_n = \{(x_{i1}, \dots, x_{i,p-1}, y_i); i = 1, \dots, n\}$ . We arrange the  $n$  observations as rows in an  $n \times p$  matrix  $\mathbf{X} = [X_1, \dots, X_{p-1}, Y]$  where the  $X_i$  and  $Y$  are  $n$ -dimensional column vectors.

**Step 1:** In the first step we construct the sweeping variables  $X_1^S, \dots, X_{p-1}^S$ . We start with  $X_1^S = X_1$ . To obtain  $X_j^S$  ( $j > 1$ ) we successively sweep  $X_1^S, \dots, X_{j-1}^S$  out of the original

variable  $X_j$ . In general, to sweep  $X_k^S$  out of  $X_l$  ( $k < l$ ) we compute

$$\hat{\beta}_{lk} = \text{med}_{j \in J} \frac{x_{jl} - \text{med}_{i=1}^n x_{il}}{x_{jk}^S - \text{med}_{i=1}^n x_{ik}^S} \quad (5)$$

where  $J$  is the collection of indices for which the denominator is different from zero, and then we replace  $X_l$  by  $X_l - \hat{\beta}_{lk} X_k$ . If  $k < l - 1$  we can now sweep the next variable  $X_{k+1}^S$  out of this new  $X_l$ . If  $k = l - 1$  then  $X_l^S = X_l$ . Thus we obtain the sweeping variables

$$\begin{aligned} X_1^S &= X_1 \\ X_2^S &= X_2 - \hat{\beta}_{21} X_1^S \\ &\vdots \\ X_{p-1}^S &= X_{p-1} - \hat{\beta}_{p-1,1} X_1^S - \cdots - \hat{\beta}_{p-1,p-2} X_{p-2}^S. \end{aligned}$$

**Step 2:** In the second step we successively sweep  $X_1^S, \dots, X_{p-1}^S$  out of  $Y$ . Put  $Y^0 = Y$ . For  $k = 1, \dots, p - 1$  we now compute

$$\hat{\beta}_k = \text{med}_{j \in J} \frac{y_j^{k-1} - \text{med}_{i=1}^n y_i^{k-1}}{x_{jk}^S - \text{med}_{i=1}^n x_{ik}^S} \quad (6)$$

with  $J$  as before, and we replace the original  $Y^{k-1}$  by  $Y^k = Y^{k-1} - \hat{\beta}_k X_k^S$ . Thus we obtain

$$Y^S = Y - \hat{\beta}_1 X_1^S - \cdots - \hat{\beta}_{p-1} X_{p-1}^S. \quad (7)$$

The process (6)-(7) is iterated until convergence is reached. In each iteration step all the coefficients  $\hat{\beta}_1, \dots, \hat{\beta}_{p-1}$  are updated. Usually only a few iterations are needed, and in any case the number of iterations has been limited to 100. After the iteration process, we take the median of  $Y^S$  to be the intercept  $\hat{\beta}_p$ .

**Step 3:** By backtransforming  $\hat{\beta}_1, \dots, \hat{\beta}_p$  we obtain the regression coefficients  $(\hat{\theta}_1^S, \dots, \hat{\theta}_p^S)^t$  corresponding to the original variables  $X_1, \dots, X_{p-1}, Y$ . The obtained fit  $\hat{\theta}^S$  is then slightly adjusted until it passes through  $p$  observations, because we know that this can only improve the depth of the fit. We start by making the smallest absolute residual zero. Then for each of the directions  $X_1, \dots, X_{p-1}$  we tilt the fit in that direction until it passes an observation while not changing the sign of any other residual. This yields the fit  $\hat{\theta}$ .

**Step 4:** In the last step we approximate the depth of the final fit  $\hat{\boldsymbol{\theta}}$ . Let  $\mathbf{u}_1^S, \dots, \mathbf{u}_{p-1}^S$  be the directions corresponding to the variables  $X_1^S, \dots, X_{p-1}^S$ , then we compute the minimum over  $\mathbf{u} \in \{\mathbf{e}_1, -\mathbf{e}_1, \dots, \mathbf{e}_{p-1}, -\mathbf{e}_{p-1}, \mathbf{u}_1^S, -\mathbf{u}_1^S, \dots, \mathbf{u}_{p-1}^S, -\mathbf{u}_{p-1}^S\}$  instead of over all unit vectors  $\mathbf{u} \in \mathbb{R}^{p-1}$  in the right hand side of expression (1).

Since computing the median takes  $O(n)$  time, the first step of the algorithm needs  $O(p^2n)$  time and the second step takes  $O(hpn)$  time where  $h$  is the number of iterations. The adjustments in step 3 also take  $O(p^2n)$  time, and computing the approximate depth in the last step can be done in  $O(pn \log n)$  time. The time complexity of the MEDSWEEP algorithm thus becomes  $O(p^2n + hpn + pn \log n)$  which is very low.

To measure the performance of our algorithm we carried out the following simulation. For different values of  $p$  and  $n$  we generated  $m = 10,000$  samples  $Z^{(j)} = \{(x_{i1}, \dots, x_{i,p-1}, y_i); i = 1, \dots, n\}$  from the standard gaussian distribution. For each of these samples we computed the deepest regression  $(\hat{\theta}_1^{(j)}, \dots, \hat{\theta}_p^{(j)})^t$  with the MEDSWEEP algorithm and measured the total time needed for these 10,000 estimates. For each  $n$  and  $p$  we also computed the bias of the intercept, which is the average of the 10,000 intercepts, and the bias of the vector of the slopes, which we measure by

$$b(\hat{\theta}_1, \dots, \hat{\theta}_{p-1}) = \sqrt{\frac{1}{p-1} ((\text{ave}_j \hat{\theta}_1^{(j)})^2 + \dots + (\text{ave}_j \hat{\theta}_{p-1}^{(j)})^2)}. \quad (8)$$

We also give the mean squared error of the vector of the slopes, given by

$$\text{MSE}(\hat{\theta}_1, \dots, \hat{\theta}_{p-1}) = \frac{1}{m} \sum_{j=1}^m \frac{1}{p-1} \sum_{i=1}^{p-1} (\hat{\theta}_i^{(j)} - \theta_i)^2 \quad (9)$$

where the true values  $\theta_i; i = 1, \dots, p-1$  equal zero, and the mean squared error of the intercept, given by  $\frac{1}{m} \sum_{j=1}^m (\hat{\theta}_p^{(j)})^2$ . Table 1 lists the bias and mean squared error of the vector of the slopes, while the bias and mean squared error of the intercept are given in Table 2. Note that the bias and mean squared error of the slope vector and the intercept are low, and that they decrease with increasing  $n$ . From Tables 1 and 2 we also see that the mean squared error does not seem to increase with  $p$ .

Table 3 lists the average time the MEDSWEEP algorithm needed for the computation of the  $DR$ , on a Sun SparcStation 20/514. We see that the algorithm is very fast.

To illustrate the MEDSWEEP algorithm we generated 50 points in 5 dimensions, according to the model  $y = x_1 - x_2 + x_3 - x_4 + 1 + e$  with  $x_1, x_2, x_3, x_4$  and  $e$  coming from

Table 1: Bias (8) and mean squared error (9) of the *DR* slope vector, obtained by generating  $m = 10,000$  standard gaussian samples for each  $n$  and  $p$ . The *DR* fits were obtained with the MEDSWEEP algorithm. The results for the bias have to be multiplied by  $10^{-4}$  and the results for the MSE by  $10^{-3}$ .

| n    |      | p     |       |       |       |
|------|------|-------|-------|-------|-------|
|      |      | 3     | 4     | 5     | 10    |
| 50   | bias | 18.27 | 28.53 | 22.47 | 21.42 |
|      | MSE  | 52.32 | 52.23 | 52.54 | 58.65 |
| 100  | bias | 8.41  | 11.16 | 8.98  | 12.68 |
|      | MSE  | 25.32 | 25.99 | 26.15 | 27.17 |
| 300  | bias | 5.98  | 6.29  | 7.39  | 10.89 |
|      | MSE  | 8.27  | 8.26  | 8.41  | 8.42  |
| 500  | bias | 1.68  | 2.91  | 9.10  | 5.49  |
|      | MSE  | 4.89  | 5.02  | 4.93  | 4.97  |
| 1000 | bias | 3.58  | 6.77  | 3.54  | 3.98  |
|      | MSE  | 2.51  | 2.46  | 2.46  | 2.50  |

the standard gaussian distribution. The *DR* fit obtained with MEDSWEEP is  $y = 0.98x_1 - 1.01x_2 + 0.97x_3 - 1.00x_4 + 1.14$  with approximate depth 21. The algorithm needed 14 iterations till convergence. In a second example, we generated 50 points according to the model  $y = 100x_1 + x_2 - 2x_3 + 3x_4 - 4 + e$  with standard gaussian  $x_1, x_2, x_3, x_4$  and  $e$ . After 25 iterations the MEDSWEEP algorithm yielded the fit  $y = 99.99x_1 + 0.99x_2 - 2.03x_3 + 3.00x_4 - 3.86$  with approximate depth 21. Note that in both cases the coefficients obtained by the algorithm approximate the true parameters in the model very well. The MEDSWEEP algorithm is available from our website <http://win-www.uia.ac.be/u/statis/> where its use is explained.

Table 2: Bias and mean squared error of the *DR* intercepts, obtained by generating  $m = 10,000$  standard gaussian samples for each  $n$  and  $p$ . The *DR* fits were obtained with the MEDSWEEP algorithm. The results for the bias have to be multiplied by  $10^{-4}$  and the results for the MSE by  $10^{-3}$ .

| n    |      | p      |       |       |        |
|------|------|--------|-------|-------|--------|
|      |      | 3      | 4     | 5     | 10     |
| 50   | bias | -48.70 | 14.38 | 13.23 | -19.64 |
|      | MSE  | 32.62  | 35.23 | 36.72 | 44.35  |
| 100  | bias | -3.32  | 12.21 | 9.92  | -1.70  |
|      | MSE  | 16.01  | 16.92 | 16.77 | 18.14  |
| 300  | bias | -7.99  | -2.37 | 3.49  | 4.54   |
|      | MSE  | 5.31   | 5.22  | 5.23  | 5.47   |
| 500  | bias | -5.53  | -4.33 | -4.26 | 2.39   |
|      | MSE  | 3.15   | 3.21  | 3.21  | 3.18   |
| 1000 | bias | -3.40  | -5.10 | 0.75  | -0.01  |
|      | MSE  | 1.56   | 1.55  | 1.58  | 1.62   |

## 4 Inference

### 4.1 Tests for parameters

In simple regression, the semiparametric model assumptions of condition (H) state that  $\text{med}(y|x) = \tilde{\theta}_1 x + \tilde{\theta}_2$  and that the errors  $e_i = y_i - \tilde{\theta}_1 x_i - \tilde{\theta}_2$  are independent with  $P(e_i > 0) = 1/2 = P(e_i < 0)$ , hence  $P(e_i = 0) = 0$ . Then it is possible to compute  $F_n(k) := P(\text{rdepth}(\tilde{\boldsymbol{\theta}}, Z'_n) \leq k)$  where  $Z'_n$  has the same  $\{x_1, \dots, x_n\}$  as the actual dataset  $Z_n$ . By invariance properties,

$$F_n(k) = P(\text{rdepth}(\mathbf{0}, \{(x_i, e_i); i = 1, \dots, n\}) \leq k) \quad (10)$$

where the  $e_i$  are i.i.d. from (say) the standard gaussian. Thus we can compute  $F_n(k)$  by simulating (10). When there are *no ties* among the  $x_i$  we can even compute  $F_n(k)$  explicitly making use of formula (4.4) in Daniels [6], yielding

$$F_n(k) = 2(n - 2k) \sum_{j=0}^{j'} B(n, \frac{1}{2})(n - k + j(n - 2k)) \quad (11)$$

Table 3: Computation time (in seconds) of the MEDSWEEP algorithm for a sample of size  $n$  with  $p$  dimensions. Each time is an average over 10,000 samples.

| n    | p     |       |       |      |
|------|-------|-------|-------|------|
|      | 3     | 4     | 5     | 10   |
| 50   | 0.023 | 0.040 | 0.057 | 0.20 |
| 100  | 0.071 | 0.15  | 0.25  | 0.51 |
| 300  | 0.21  | 0.42  | 0.67  | 1.38 |
| 500  | 0.39  | 0.73  | 1.12  | 2.24 |
| 1000 | 0.66  | 1.43  | 2.18  | 4.42 |

for  $k \leq [(n-1)/2]$ , and  $F_n(k) = 1$  otherwise. Here  $j' = [k/(n-2k)]$  and each term is a probability of the binomial distribution  $B(n, 1/2)$ , which stems from the number of  $e_i$  in  $\{e_1, \dots, e_n\}$  with a particular sign. For increasing  $n$  we can approximate  $B(n, 1/2)$  by a gaussian distribution due to the central limit theorem, so (11) can easily be extended to large  $n$ .

The distribution of the regression depth allows us to test one or several regression coefficients. To test the combined null hypothesis  $(\tilde{\theta}_1, \tilde{\theta}_2)^t = (0, 0)^t$  we compute  $k = rdepth((0, 0)^t, Z_n)$  and the corresponding  $p$ -value equals  $F_n(k)$  and can be computed from (11). Consider the dataset in Figure 2 about  $n = 41$  species of animals (Van den Bergh [27]). The plot shows the logarithm of the weight of a newborn versus the logarithm of the weight of its mother. The deepest regression line  $DR = (0.86, -2.12)^t$  has depth 19. For this dataset  $rdepth((0, 0)^t, Z_n) = 1$ , yielding the  $p$ -value  $F_{41}(1) = 0.00000$  which is highly significant. To test the significance of the slope ( $H_0 : \tilde{\theta}_1 = 0$ ) we compute  $\max rdepth((0, \theta_2)^t, Z_n)$  over  $\theta_2$ . This is easy, because we only have to compute the  $rdepth$  of all horizontal lines passing through an observation. For the animal dataset, the maximal  $rdepth((0, \theta_2)^t, Z_n)$  equals 5. Therefore the corresponding  $p$ -value is  $P(rdepth(\tilde{\theta}, Z'_n) \leq 5) = F_{41}(5) = 0.00002$ , so  $H_0$  is rejected. This  $p$ -value 0.00002 should be interpreted in the same way as the  $p$ -value associated with  $R^2$  or the F-test in LS regression. Analogously, to test  $\tilde{\theta}_2 = 0$  we compute  $\max rdepth((\theta_1, 0)^t, Z_n) = 6$  by considering all lines through the origin and an observation, yielding the  $p$ -value  $F_{41}(6) = 0.0001$  which is also highly significant.

More generally, we can test the hypothesis  $H_0 : \tilde{\theta}_i = \theta_0$  for  $i = 1, 2$  by computing the maximal regression depth of all lines with  $\theta_i = \theta_0$  that pass through an observation.

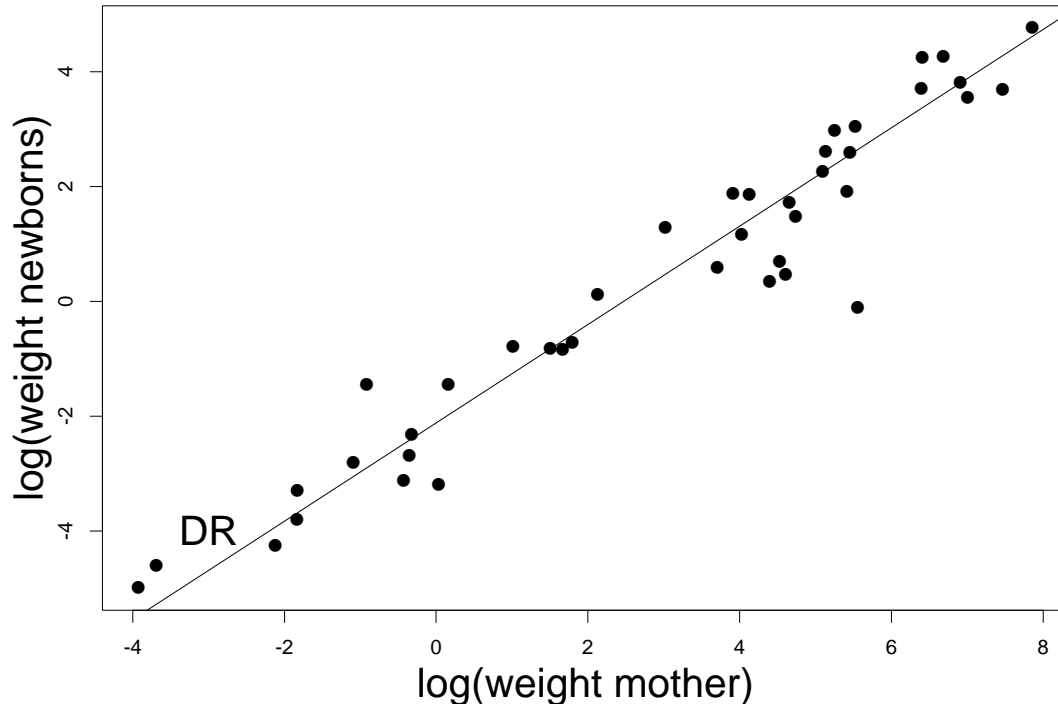


Figure 2: Logarithm of the weight of a newborn versus the logarithm of the weight of its mother for  $n = 41$  species of animals, with the *DR* line which has depth 19.

For example, to test the hypothesis  $H_0 : \tilde{\theta}_1 = 1$  for the animal data (i.e., the hypothesis that the weight of a newborn is proportional to the weight of its mother) we compute  $\max rdepth((1, \theta_2)^t, Z_n) = 10$  and the corresponding  $p$ -value  $F_{41}(10) = 0.021$ , which is significant at the 5% level but not at the 1% level.

These tests generalize easily to higher dimensions and situations with ties among the  $x_i$ , but then we can no longer use the exact formula (11) which is restricted to the *bivariate* case without ties in  $\{x_1, \dots, x_n\}$ . Therefore, in these cases we compute  $F_n(k)$  by simulating (10).

Let us consider the stock return dataset (Benderly and Zwick [3]) shown in Figure 3 with  $n = 28$  observations in  $p = 3$  dimensions. The regressors are output growth and inflation (both as percentages), and the response is the real return on stocks. The deepest regression obtained with the MEDSWEEP algorithm equals  $DR = (3.41, -2.22, 6.66)^t$  with approximate depth 11. To test the null hypothesis  $(\tilde{\theta}_1, \tilde{\theta}_2)^t = (0, 0)^t$  that both slopes are zero (this would be done with the  $R^2$  in LS regression) we compute the maximal  $rdepth((0, 0, \theta_3)^t, Z_n)$  over all  $\theta_3$  (i.e. over all  $y_i$  in the dataset). By computing the exact  $rdepth$  of these 28 hor-

horizontal planes (by the fast algorithm of Rousseeuw and Struyf [22]) we obtain the value 6. Simulation yields the corresponding  $p$ -value  $F_{28}(6) = 0.22$ , which is not significant. To test the significance of the intercept ( $H_0 : \tilde{\theta}_3 = 0$ ) we compute  $rdepth((\theta_1, \theta_2, 0)^t, Z_n)$  over all  $(\theta_1, \theta_2)^t$ . That is, we compute the depth of all planes through two observations and the origin. For the example this yields  $\max rdepth((\theta_1, \theta_2, 0)^t, Z_n) = 11$  with corresponding  $p$ -value  $F_{28}(11) \approx 1$ , which is not at all significant.

## 4.2 Confidence regions

In order to construct a confidence region for the unknown true parameter vector  $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \dots, \tilde{\theta}_p)^t$  we use a bootstrap method. Starting from the dataset  $Z_n = \{(\mathbf{x}_i^t, y_i); i = 1, \dots, n\} \in \mathbb{R}^p$ , we construct a bootstrap sample by randomly drawing  $n$  observations, with replacement. For each bootstrap sample  $Z_n^{(j)}$ ,  $j = 1, \dots, m$  we compute its deepest regression fit  $\hat{\boldsymbol{\theta}}^{(j)}$ . Note that there will usually be a few outlying estimates  $\hat{\boldsymbol{\theta}}^{(j)}$  in the set of bootstrap fits  $\{\hat{\boldsymbol{\theta}}^{(j)}; j = 1, \dots, m\}$ , which is natural since some bootstrap samples contain disproportionately many outliers. Therefore we don't construct a confidence ellipsoid based on the classical mean and covariance matrix of the  $\{\hat{\boldsymbol{\theta}}^{(j)}; j = 1, \dots, m\}$ , but we use the robust minimum covariance determinant estimator (MCD) proposed by (Rousseeuw [18,19]).

The MCD looks for the  $h \geq n/2$  observations of which the empirical covariance matrix has the smallest possible determinant. Then the center  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_p)^t$  of the dataset is defined as the average of these  $h$  points, and the covariance matrix  $\hat{\Sigma}$  of the dataset is a certain multiple of their covariance matrix. To obtain a confidence ellipsoid of level  $\alpha$  we compute the MCD of the set of bootstrap estimates with  $h = \lceil (1 - \alpha)m \rceil$ . The  $(1 - \alpha)\%$  confidence ellipsoid  $E_{1-\alpha}$  is then given by

$$E_{1-\alpha} = \{\boldsymbol{\theta} \in \mathbb{R}^p; (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t \hat{\Sigma}^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq c^2\}. \quad (12)$$

Here  $c := RD(\hat{\boldsymbol{\theta}}^{(j)})_{\lceil (1-\alpha)m \rceil; m}$  is the  $\lceil (1 - \alpha)m \rceil$  order statistic of the robust distances of the bootstrap estimates  $\{\hat{\boldsymbol{\theta}}^{(j)}; j = 1, \dots, m\}$ , where the robust distance (Rousseeuw and Leroy [21]) of  $\hat{\boldsymbol{\theta}}^{(j)}$  is given by

$$RD(\hat{\boldsymbol{\theta}}^{(j)}) = \sqrt{(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}})^t \hat{\Sigma}^{-1} (\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}})}. \quad (13)$$

From this confidence ellipsoid  $E_{1-\alpha}$  in fit space we can also derive the corresponding regression confidence region for  $\hat{y} = \tilde{\theta}_1 x_1 + \dots + \tilde{\theta}_{p-1} x_{p-1} + \tilde{\theta}_p$  defined as

$$R_{1-\alpha} = \{(\mathbf{x}^t, y) \in \mathbb{R}^p; \min((\mathbf{x}^t, 1)\boldsymbol{\theta}) \leq y \leq \max((\mathbf{x}^t, 1)\boldsymbol{\theta}) \text{ where } \boldsymbol{\theta} \in E_{1-\alpha}\}.$$

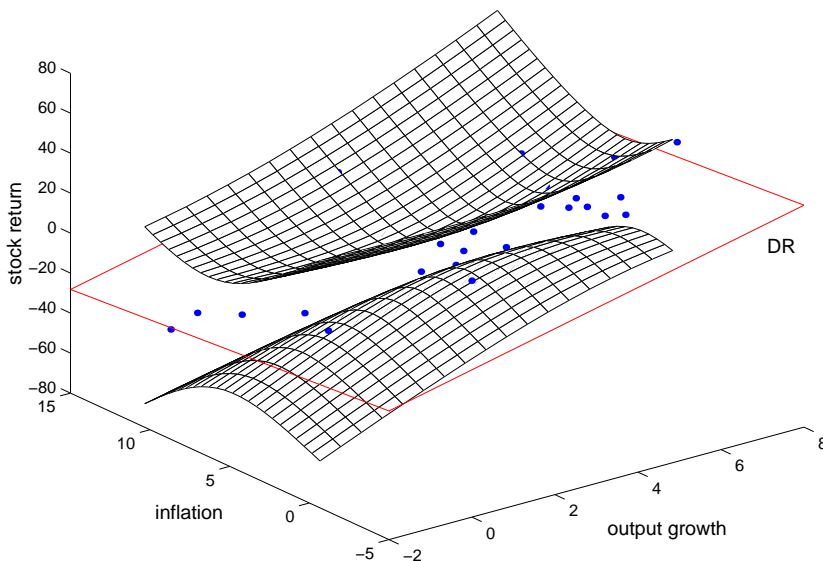


Figure 3: Stock return dataset with its deepest regression plane, and the upper and lower surface of the 95% confidence region  $R_{0.95}$  based on  $m = 1,000$  bootstrap samples.

**Theorem 2.** This region  $R_{1-\alpha}$  equals the set

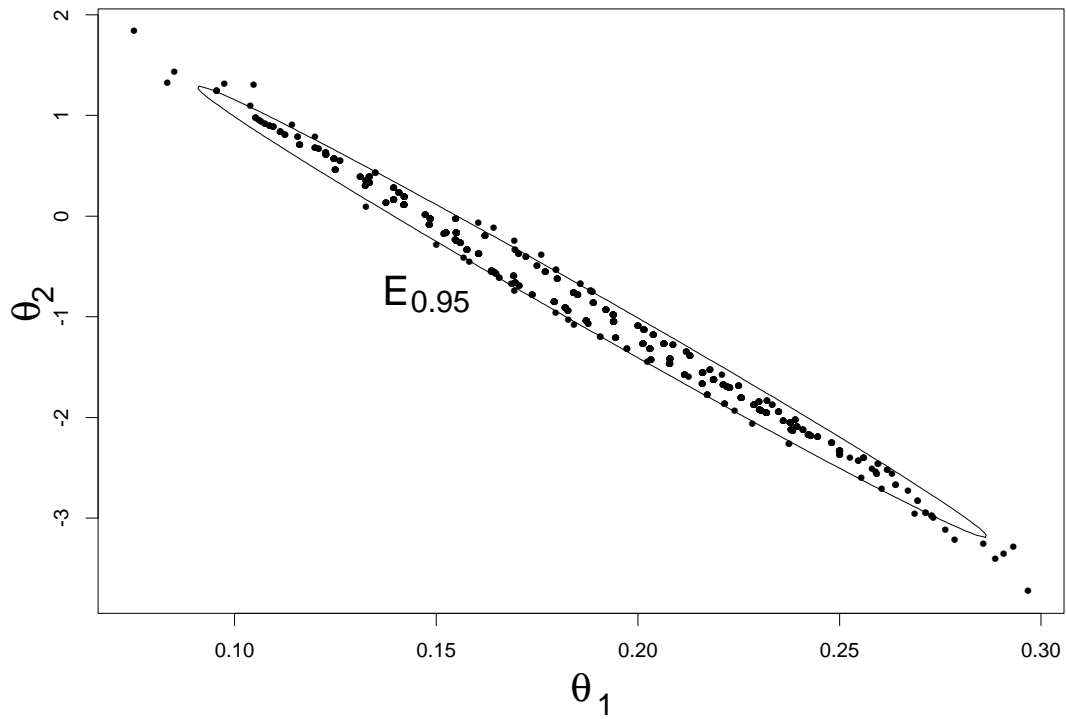
$$\{(\mathbf{x}^t, y) \in \mathbb{R}^p; (\mathbf{x}^t, 1)\hat{\boldsymbol{\theta}} - c\sqrt{(\mathbf{x}^t, 1)\hat{\Sigma}(\mathbf{x}^t, 1)^t} \leq y \leq (\mathbf{x}^t, 1)\hat{\boldsymbol{\theta}} + c\sqrt{(\mathbf{x}^t, 1)\hat{\Sigma}(\mathbf{x}^t, 1)^t}\} \quad (14)$$

with the same constant  $c$  as in (12).

Let us consider the Educational Spending data of Figure 1. Figure 4a shows the deepest regression estimates of 1000 bootstrap samples, drawn with replacement from the original data. Using the fast MCD algorithm of Rousseeuw and Van Driessen [23] we find the center in Figure 4a to be  $(0.19, -0.95)^t$  which corresponds well to the  $DR$  fit  $y = 0.19x - 1.05$  of the original data. As a confidence region for  $(\tilde{\theta}_1, \tilde{\theta}_2)^t$  we take the 95% tolerance ellipse  $E_{0.95}$  based on the MCD center and scatter matrix, which yields the corresponding confidence region  $R_{0.95}$  shown in Figure 4b. Note that the intersection of this confidence region with a vertical line  $x = x_0$  is not a 95% probability interval for an *observation*  $y$  at  $x_0$ . It is the interval spanned by the *fitted values*  $\hat{y} = \theta_1 x_0 + \theta_2$  for  $(\theta_1, \theta_2)^t$  in a 95% confidence region for  $(\tilde{\theta}_1, \tilde{\theta}_2)^t$ .

An example of a confidence region in higher dimensions is shown in Figure 3. It shows the 3-dimensional stock return dataset with its deepest regression plane, obtained with the MEDSWEEP algorithm. The 95% confidence region shown in Figure 3 was based on  $m = 1,000$  bootstrap samples.

(a)



(b)

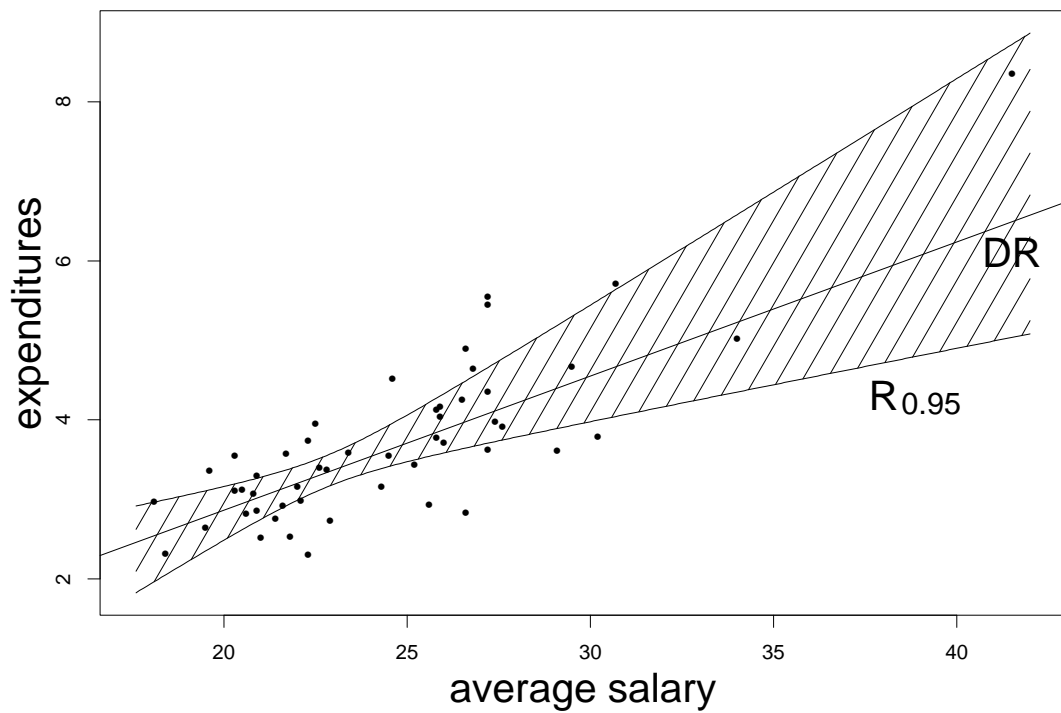


Figure 4: (a)  $DR$  estimates of 1,000 bootstrap samples from the Educational Spending data with the 95% confidence ellipse  $E_{0.95}$ ; (b) Plot of the data with the  $DR$  line and the 95% confidence region  $R_{0.95}$  for the fitted value.

Table 4: Windmill data: Maximal rdepth  $k$  with corresponding  $p$ -value  $p(k)$  obtained by simulation using a gaussian, Cauchy, and exponential error distribution.

| k  | p(k)     |        |             |
|----|----------|--------|-------------|
|    | gaussian | Cauchy | exponential |
| 12 | 1        | 1      | 1           |
| 11 | 0.7334   | 0.7235 | 0.7250      |
| 10 | 0.0481   | 0.0470 | 0.0487      |
| 9  | 0.0001   | 0      | 0           |

### 4.3 Test for linearity in simple regression

If the observations of the bivariate dataset  $Z_n$  lie exactly on a straight line, then  $\max_{\boldsymbol{\theta}} \text{rdepth}(\boldsymbol{\theta}, Z_n) = n$  is the highest possible value. On the other hand, if  $Z_n$  lies exactly on a strictly convex or strictly concave curve, then  $\max_{\boldsymbol{\theta}} \text{rdepth}(\boldsymbol{\theta}, Z_n) \approx n/3$  is at its lowest (Rousseeuw and Hubert [20, Theorem 2]). Therefore,  $\max_{\boldsymbol{\theta}} \text{rdepth}(\boldsymbol{\theta}, Z_n)$  can be seen as a measure of linearity for the dataset  $Z_n$ . Note that the alternative of convexity/concavity is very general, in contrast with other linearity tests (such as the F-test) where a more specific alternative, e.g. a quadratic term, is needed.

Note that this lower bound does not depend on the *amount* of curvature when the  $(x_i, y_i)$  lie *exactly* on the curve. However, as soon as there is noise (i.e. nearly always), the relative sizes of the error scale and the curvature come into play.

The null hypothesis assumes that the dataset  $Z_n$  follows the linear model

$$H_0 : y_i = \tilde{\theta}_1 x_i + \tilde{\theta}_2 + e_i \quad i = 1, \dots, n$$

for some  $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \tilde{\theta}_2)^t$  and with independent errors  $e_i$  each having a distribution with zero median. To determine the corresponding  $p$ -value we generate  $m = 10,000$  samples  $Z^{(j)} = \{(x_i, e_i); i = 1, \dots, n\}$  with the same  $x$ -values as in the dataset  $Z_n$  and with standard gaussian  $e_i$ . For each  $j$  we compute the maximal regression depth of the dataset  $Z^{(j)}$ . Then for each value  $k$  we approximate the  $p$ -value  $P(\text{maxrdepth} \leq k | H_0)$  by

$$p(k) = \#\{j; \text{maxrdepth}(Z^{(j)}) \leq k\} / m. \quad (15)$$

For example, let us consider the windmill dataset (Hand et al. [10]) which consists of 25 measures of wind velocity with corresponding direct current output, as shown in Figure 5.

Table 5: Maximal rdepth  $k$  with corresponding  $p$ -values  $p(k)$  for 100 equispaced  $x_i$  in  $[0, 1]$ . The  $p$ -values are based on 10,000 replications using a gaussian, Cauchy, and exponential error distribution.

| k  | p(k)     |        |             |
|----|----------|--------|-------------|
|    | gaussian | Cauchy | exponential |
| 49 | 1        | 1      | 1           |
| 48 | 0.9995   | 0.9997 | 0.9997      |
| 47 | 0.9446   | 0.9477 | 0.9488      |
| 46 | 0.5985   | 0.6030 | 0.6045      |
| 45 | 0.2086   | 0.2088 | 0.2128      |
| 44 | 0.0405   | 0.0399 | 0.0408      |
| 43 | 0.0063   | 0.0048 | 0.0054      |
| 42 | 0.0002   | 0.0007 | 0.0007      |
| 41 | 0        | 0      | 0.0001      |

The first column of  $p$ -values in Table 4 was obtained from (15) using gaussian errors (we put  $\mu = 0$  and  $\sigma = 1$  without loss of generality). For the actual  $\max rdepth(Z_n) = 10$  we obtain  $p(10) = 0.0481$ , so we reject the linearity at the 5% level. To illustrate that this  $p$ -value does not depend much on the type of error distribution, we also computed  $p$ -values using Cauchy distributed errors (second column) as an example of a very long-tailed error distribution. As an example of a very asymmetric error distribution, we generated errors according to  $e_i = u_i - 1$  with  $u_i$  exponentially distributed (third column). From Table 4 we see that the resulting  $p$ -values are all very similar.

To further investigate the effect of the type of error distribution on the resulting  $p$ -values, we performed the following simulation. We took 100 equispaced  $x_i = (i - 1/2)/100$  in  $[0, 1]$  and generated errors according to the three error distributions given above. The corresponding  $p$ -values are shown in Table 5. As in Table 4 we see that the  $p$ -values change very little with the type of error distribution used in (15). Note that in all three cases we reject linearity at the 5% level if the maximal depth  $k \leq 44$ , and we reject linearity at the 1% level if  $k \leq 43$ .

## 5 Nonlinear models

### 5.1 Depth of a general function

By definition, the regression depth of a fit  $\boldsymbol{\theta}$  relative to a dataset  $Z_n = \{(\mathbf{x}_i, y_i); i = 1, \dots, n\}$  only depends on the  $\mathbf{x}_i$  and the signs of the residuals. Therefore this definition can also be applied to more general models. For example, suppose we have a regression fit of the form

$$y = f(\mathbf{x}) = f(x_1, \dots, x_{p-1}) \quad (16)$$

for some real function  $f$ . Denote the residuals  $r_i(f) = y_i - f(x_{i1}, \dots, x_{i,p-1})$ . Then the regression depth of  $f$  is defined as follows.

**Definition 3.** For any data set  $Z_n = \{(\mathbf{x}_i, y_i); i = 1, \dots, n\}$  and any real function  $f$  on  $\mathbb{R}^{p-1}$  the regression depth of  $f$  is defined as

$$rdepth(f, Z_n) = \min_{\mathbf{u}, v} \{ \#(r_i(f) \geq 0 \text{ and } \mathbf{x}_i^t \mathbf{u} < v) + \#(r_i(f) \leq 0 \text{ and } \mathbf{x}_i^t \mathbf{u} > v) \} \quad (17)$$

where (as in Definition 1) the minimum is over all unit vectors  $\mathbf{u} = (u_1, \dots, u_{p-1})^t \in \mathbb{R}^{p-1}$  and all  $v \in \mathbb{R}$  with  $\mathbf{x}_i^t \mathbf{u} \neq v$  for all  $(\mathbf{x}_i^t, y_i) \in Z_n$ .

The regression depth has the following monotone invariance property.

**Proposition 1.** Suppose we have a dataset  $Z_n = \{(\mathbf{x}_i, y_i); i = 1, \dots, n\}$  and a strictly monotone real function  $g$ . Denote  $y'_i = g(y_i)$  and  $Z'_n = \{(\mathbf{x}_i, y'_i); i = 1, \dots, n\}$ . Then it holds for any function  $f$  that

$$rdepth(f, Z_n) = rdepth(g(f), Z'_n).$$

This property of regression depth allows us to deal with several interesting models, as shown in the following examples.

### 5.2 Generalized linear models

Suppose we want a regression fit of the form

$$y = g(\theta_1 x_1 + \dots + \theta_{p-1} x_{p-1} + \theta_p) \quad (18)$$

with  $g$  a link function. Denote  $r_i(g_{\boldsymbol{\theta}}) = y_i - g(\theta_1 x_{i1} + \dots + \theta_{p-1} x_{i,p-1} + \theta_p)$ , then the regression depth of the (nonlinear) fit  $g_{\boldsymbol{\theta}}$  is given by Definition 3. Using this definition of depth, we

now compute the deepest generalized linear regression as in (2) and denote it by  $DR_g(Z_n)$ . If several fits  $g_{\theta}$  have the same (maximal) regression depth, then we take the average of all those  $\theta$ .

From the monotone invariance of the regression depth (Proposition 1) it follows that the deepest regression has a monotone equivariance property which allows for monotone transformations of the response  $y_i$ . This monotone equivariance does not hold for  $L^1$  or other estimators such as least squares, least trimmed squares (Rousseeuw [18]) or S-estimators (Rousseeuw and Yohai [24]).

**Proposition 2.** *Take  $Z_n = \{(\mathbf{x}_i, y_i); i = 1, \dots, n\}$  and a strictly monotone link function  $g$ . Put  $\tilde{y}_i = g^{-1}(y_i)$  and denote the deepest linear regression of the transformed data  $(\mathbf{x}_i^t, \tilde{y}_i)$  as  $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_p)^t$ . Then the deepest generalized linear regression to the original data is*

$$DR_g(Z_n) = g\hat{\theta}.$$

Typical examples of  $g$  include the logarithmic, the exponential, the square root, the square and the reciprocal transformation.

### 5.3 Polynomial regression

Consider a dataset  $Z_n = \{(x_i, y_i); i = 1, \dots, n\} \subset \mathbb{R}^2$ . Polynomial regression wants to fit the data by  $y = \theta_1 x + \theta_2 x^2 + \dots + \theta_k x^k + \theta_{k+1}$  where  $k$  is called the degree of the polynomial. The residuals of  $Z_n$  relative to the fit  $\theta = (\theta_1, \dots, \theta_{k+1})^t$  are denoted as  $r_i = r_i(\theta) = y_i - \theta_1 x - \dots - \theta_k x^k - \theta_{k+1}$ .

We could consider this to be a multiple linear regression problem with regressors  $\mathbf{x} = (x, x^2, \dots, x^k)^t$  and determine the depth of a fit  $\theta$  as in Definition 1. But we know that the joint distribution of  $(\mathbf{x}^t, y)$  is degenerate (i.e. it does not have a density), so many properties of the deepest regression, such as Theorem 1 about the breakdown value, would not hold in this case. In fact, the set of possible  $\mathbf{x}$  forms a so-called *moment curve* in  $\mathbb{R}^k$ , so it is inherently one-dimensional.

A better way to define the depth of a polynomial fit  $f(x) = \theta_1 x + \theta_2 x^2 + \dots + \theta_k x^k + \theta_{k+1}$  is given by Definition 3 of the regression depth of general functions. Note that in this case  $x$  is univariate. We denote the corresponding deepest polynomial of degree  $k$  by  $DR_k(Z_n)$ . The following theorem shows that with this definition of depth the deepest polynomial regression has a positive breakdown value of approximately 1/3, so it is robust to vertical outliers as well as to leverage points.

**Theorem 3.** For any dataset  $Z_n = \{(x_i, y_i); i = 1, \dots, n\} \in \mathbb{R}^2$  with distinct  $x_i$  the deepest polynomial regression of degree  $k$  with  $3k < n$  has breakdown value

$$\varepsilon_n^*(DR_k, Z_n) \geq \frac{n - 3k}{3n - 3k} \approx \frac{1}{3}. \quad (19)$$

In Section 4.3 we rejected the linearity of the windmill data. Let us now fit a quadratic model to this data, i.e.  $y = \theta_1 x + \theta_2 x^2 + \theta_3$ . Figure 5 shows the windmill data with the deepest quadratic fit, which has depth 12.

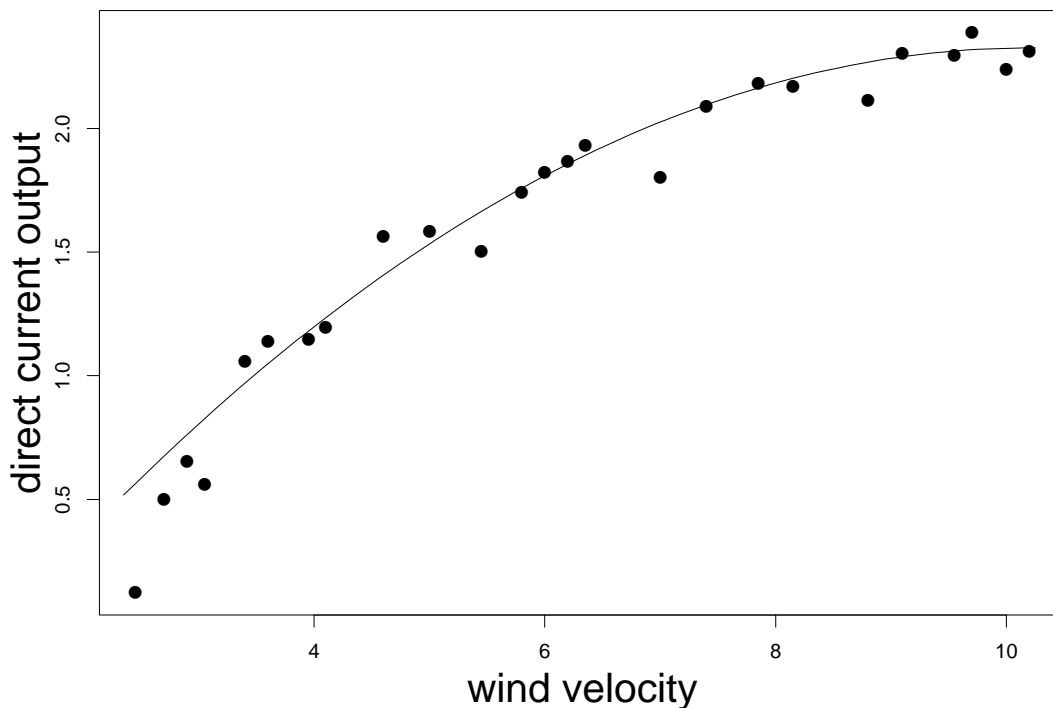


Figure 5: Windmill data with the deepest quadratic fit, which has regression depth 12.

## 5.4 Michaelis-Menten model

In the field of enzyme kinetics, the steady-state kinetics of the great majority of the enzyme-catalyzed reactions that have been studied are adequately described by a hyperbolic relationship between the concentration  $s$  of a substrate and the steady-state velocity  $v$ . This relationship is expressed by the Michaelis-Menten equation

$$v = \frac{v_{\max} s}{K_m + s} \quad (20)$$

where the constant  $v_{\max}$  is the maximum velocity and  $K_m$  is the Michaelis constant. The Michaelis-Menten equation is nonlinear, and has been linearized by rewriting it in the following three ways:

$$\frac{v}{s} = \frac{v_{\max}}{K_m} - \frac{1}{K_m}v \quad (21)$$

$$\frac{1}{v} = \frac{1}{v_{\max}} + \frac{K_m}{v_{\max}} \frac{1}{s} \quad (22)$$

$$\frac{s}{v} = \frac{K_m}{v_{\max}} + \frac{1}{v_{\max}}s \quad (23)$$

which are known as the Scatchard equation (Scatchard [25]), the double reciprocal equation (Lineweaver and Burke [13]) and the Woolf equation (Haldane [9]). Each of the three relations (21), (22), (23) can be used to estimate the constants  $v_{\max}$  and  $K_m$ . In general the three relations yield different estimates for the constants  $v_{\max}$  and  $K_m$ , because the error terms are also transformed in a nonlinear way. Cressie and Keightley [5] compared these three linearizations of the Michaelis-Menten relation in the context of hormone-receptor assays, and concluded that for well-behaved data the Woolf equation (23) works best, but for data containing outliers the double reciprocal equation (22) with robust regression gives better results.

Theorem 4 shows that applying the deepest regression to the Woolf equation (23) yields the same estimates for  $v_{\max}$  and  $K_m$  as the deepest regression applied to the double reciprocal equation (22). This resolves the ambiguity.

**Theorem 4.** *Let  $Z_n = \{(s_i, v_i); i = 1, \dots, n\} \in \mathbb{R}^2$  with  $s_i > 0$  for all  $i = 1, \dots, n$ , and denote the DR fit of the double reciprocal equation as  $DR(\{(\frac{1}{s_i}, \frac{1}{v_i}); i = 1, \dots, n\}) = (\hat{\theta}_1, \hat{\theta}_2)^t$ . Then the DR fit of the Woolf equation satisfies*

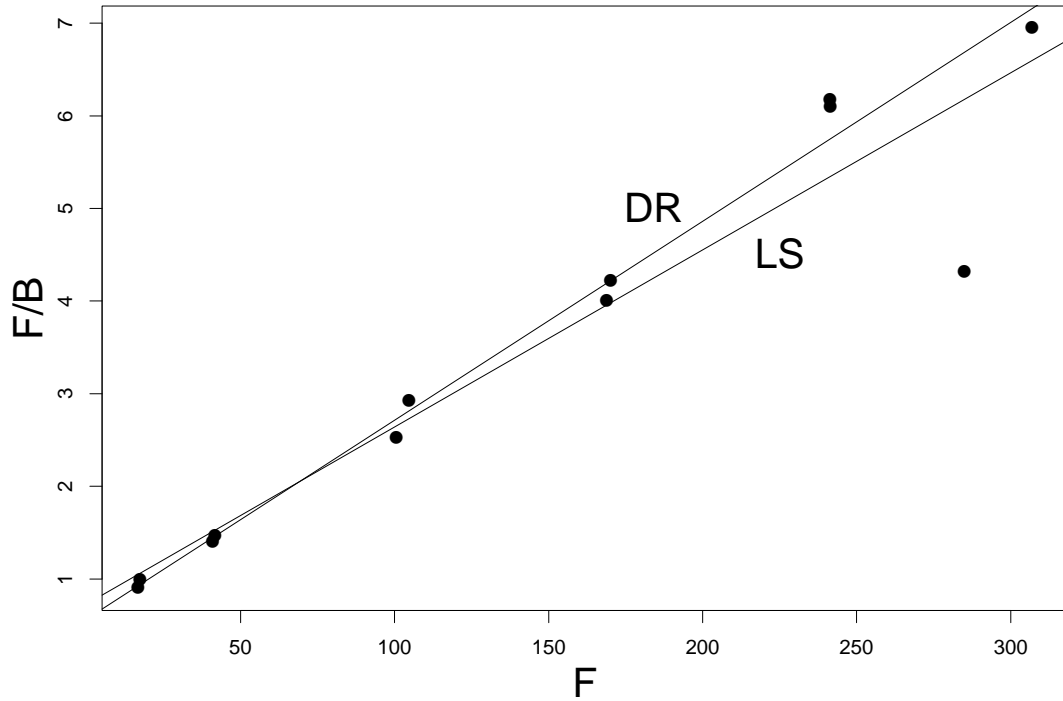
$$DR(\{(s_i, \frac{s_i}{v_i}); i = 1, \dots, n\}) = (\hat{\theta}_2, \hat{\theta}_1)^t.$$

*In both cases we obtain the same  $\hat{v}_{\max} = 1/\hat{\theta}_2$  and  $\hat{K}_m = \hat{\theta}_1/\hat{\theta}_2$ .*

**Example:** In assays for hormone receptors the Michaelis-Menten equation describes the relationship between the amount  $B$  of hormone bound to receptor and the amount  $F$  of hormone not bound to receptor. These assays are used e.g. to determine the cancer treatment method (see Cressie and Keightley [5]). Equation (20) now becomes

$$B = \frac{B_{\max}F}{K_D + F}.$$

(a)



(b)

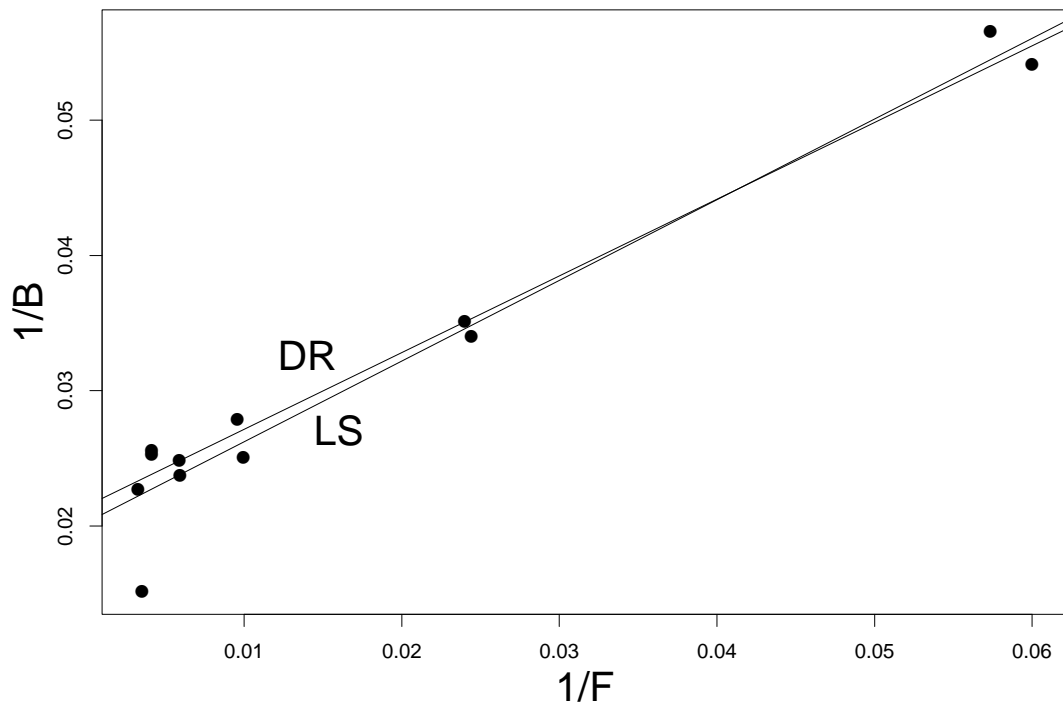


Figure 6: (a) Woolf plot of the Cressie and Keightley data with the deepest regression line  $DR = (0.0215, 0.567)^t$  and the least squares fit  $LS = (0.0191, 0.728)^t$ ; (b) double reciprocal plot of the data with the deepest regression line  $DR = (0.567, 0.0215)^t$  and the least squares fit  $LS = (0.596, 0.0203)^t$ .

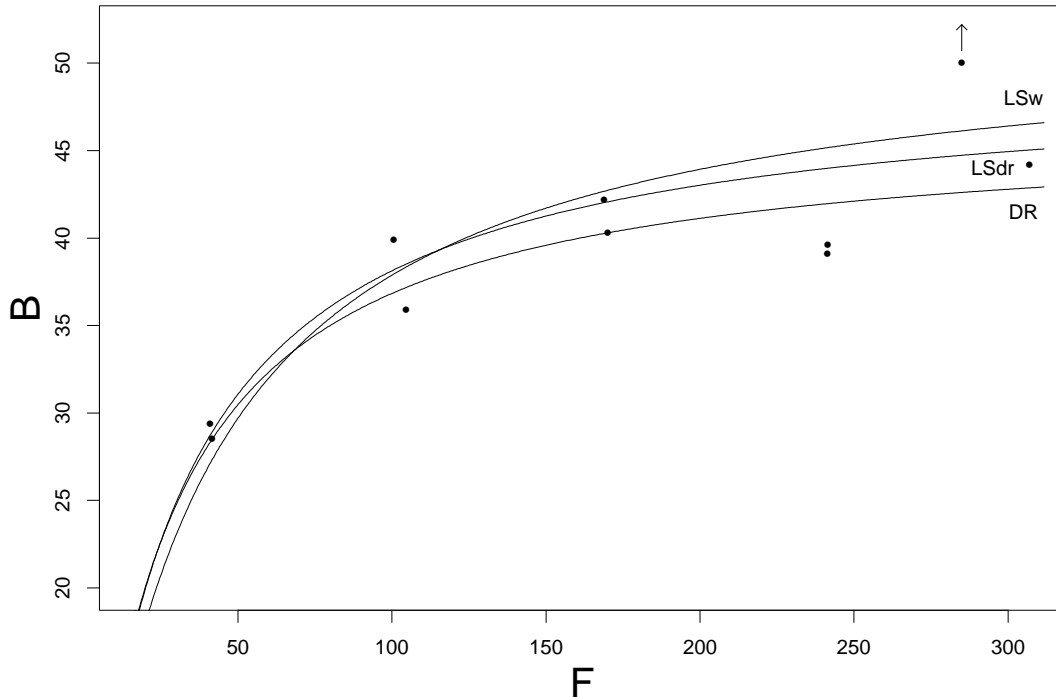


Figure 7: Plot of the Cressie and Keightley data. Superimposed are the deepest regression (DR), the transformed LS fit according to the Woolf equation (LSw) and the transformed LS fit according to the double reciprocal equation (LSdr).

The parameters of interest are the concentration  $B_{\max}$  of binding sites and the dissociation constant  $K_D$  for the hormone-receptor interaction. Figure 6a shows the Woolf plot of data from an estrogen receptor assay obtained by Cressie and Keightley [4]. Note that this dataset clearly contains one outlier. In the plot we indicated the deepest regression line  $DR(\{(F_i, \frac{F_i}{B_i}); i = 1, \dots, n\}) = (0.0215, 0.567)^t$ . In Figure 6b we show the double reciprocal plot with its deepest regression line  $DR(\{(\frac{1}{s_i}, \frac{1}{v_i}); i = 1, \dots, n\}) = (0.567, 0.0215)^t$ . Thus in both cases we obtain the same estimated values  $\hat{B}_{\max} = 46.556$  and  $\hat{K}_D = 26.398$ , which are comparable to the least squares estimates  $\hat{B}_{\max} = 45.610$  and  $\hat{K}_D = 24.079$  obtained from the Woolf equation based on all data except the outlier. On the other hand, least squares applied to the full data gives  $\hat{B}_{\max} = 52.290$  and  $\hat{K}_D = 38.048$  based on the Woolf equation, and  $\hat{B}_{\max} = 49.354$  and  $\hat{K}_D = 29.436$  based on the double reciprocal equation. These estimates are quite different. This can also be seen in Figure 7, which shows the data points and the estimated curves in the original (F,B) space. Here, *LSw* resp. *LSdr* stand

for the LS fit according to the Woolf equation and the double reciprocal equation. With least squares, we see that in both cases the estimates  $\hat{B}_{\max}$  and  $\hat{K}_D$  are highly influenced by the outlying observation (both  $\hat{B}_{\max}$  and  $\hat{K}_D$  come out too high) which may lead to wrong conclusions, e.g. when determining a cancer treatment method. Note that the outlier lies outside the plot region in the direction of the arrow.

## Appendix

**Proof of Theorem 1.** To prove Theorem 1 we need the following lemmas.

**Lemma 1.** *If the  $\mathbf{x}_i$  are in general position, then  $\{\boldsymbol{\theta}; rdepth(\boldsymbol{\theta}, Z_n) \geq p\}$  is bounded.*

**Proof:** For  $J \subset \{1, \dots, n\}$  with  $\#J = p$  we denote the fit that passes through the  $p$  observations  $\{(\mathbf{x}_{i_1}^t, y_{i_1}), \dots, (\mathbf{x}_{i_p}^t, y_{i_p}); i_j \in J\}$  by  $\boldsymbol{\theta}^J$ . Since the  $\mathbf{x}_i$  are in general position, such a fit  $\boldsymbol{\theta}^J$  will be non-vertical (for any  $J$ ). Therefore,  $\text{conv}\{\boldsymbol{\theta}^J; J \subset \{1, \dots, n\}, \#J = p\} = \text{conv}\{\boldsymbol{\theta}^J, rdepth(\boldsymbol{\theta}^J, Z_n) \geq p\} = \text{conv}\{\boldsymbol{\theta}, rdepth(\boldsymbol{\theta}, Z_n) \geq p\}$  is bounded, hence also  $\{\boldsymbol{\theta}, rdepth(\boldsymbol{\theta}, Z_n) \geq p\}$  is bounded. (Here, ‘conv’ stands for the convex hull.)  $\square$

Note that the bound in Lemma 1 depends on the sample  $Z_n$ .

**Lemma 2.** *For any dataset  $Z_n \in \mathbb{R}^p$  with the  $\mathbf{x}_i$  in general position (i.e. no more than  $p-1$  of the  $\mathbf{x}_i$  lie in any  $(p-2)$ -dimensional affine subspace of  $\mathbb{R}^{p-1}$ ) the deepest regression has breakdown value*

$$\varepsilon_n^*(DR, Z_n) \geq \frac{n - p^2 + 1}{n(p+1) - p^2 + 1} \approx \frac{1}{p+1}. \quad (24)$$

**Proof:** Let  $\boldsymbol{\eta}$  denote a fit with maximal regression depth relative to  $Z_n$ . Since  $DR$  is the average of the fits  $\boldsymbol{\eta}$ , it follows that  $\varepsilon_n^*(DR(Z_n), Z_n) = \varepsilon_n^*(\boldsymbol{\eta}, Z_n)$  for any  $\boldsymbol{\eta}$ . By Lemma 1 we know that  $\{\boldsymbol{\theta}, rdepth(\boldsymbol{\theta}, Z_n) \geq p\}$  is bounded. Therefore to break down the estimator we must add at least  $m$  observations such that  $rdepth(\boldsymbol{\xi}, Z_n) \leq p-1$  for any fit  $\boldsymbol{\xi}$  with maximal rdepth relative to  $Z_{n+m}$ . Since for all  $Z_n$  and  $\boldsymbol{\eta}$  it holds that  $rdepth(\boldsymbol{\eta}, Z_n) \geq \lceil n/(p+1) \rceil$  (see Mizera [15], Amenta et al. [1]), we obtain

$$\left\lceil \frac{n+m}{p+1} \right\rceil \leq rdepth(\boldsymbol{\xi}, Z_{n+m}) \leq p-1+m$$

for any  $\boldsymbol{\xi}$ , which yields  $m \geq \frac{n-p^2+1}{p}$ . It follows that

$$\varepsilon_n^*(DR(Z_n), Z_n) = \varepsilon_n^*(\boldsymbol{\eta}, Z_n) \geq \frac{m}{n+m} \geq \frac{n-p^2+1}{n(p+1)-p^2+1} \quad (25)$$

for any  $\eta$ .

**Lemma 3.** *If  $Z_n \subset Z_{n+m}$  then*

$$\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_n) \leq \max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_{n+m}).$$

**Proof.** Since  $Z_n \subset Z_{n+m}$  it holds that  $rdepth(\boldsymbol{\theta}, Z_n) \leq rdepth(\boldsymbol{\theta}, Z_{n+m})$  for all  $\boldsymbol{\theta}$ . Thus for all  $\boldsymbol{\theta}$  it holds that  $rdepth(\boldsymbol{\theta}, Z_n) \leq \max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_{n+m})$ , hence  $\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_n) \leq \max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_{n+m})$ .  $\square$

**Lemma 4.** *Under the conditions of Theorem 1 we have that*

$$\frac{\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_n)}{n} \xrightarrow[n \rightarrow \infty]{a.s.} \frac{1}{2}. \quad (26)$$

**Proof.** Let us consider the dual space, i.e. the  $p$ -dimensional space of all possible fits  $\boldsymbol{\theta}$ . Dualization transforms a hyperplane  $H_{\boldsymbol{\theta}} : y = \theta_1 x_1 + \dots + \theta_{p-1} x_{p-1} + \theta_p$  to the point  $\boldsymbol{\theta}$ . An observation  $\mathbf{z}_i = (x_{i,1}, \dots, x_{i,p-1}, y_i)$  is mapped to the set  $D(\mathbf{z}_i)$  of all  $\boldsymbol{\theta}$  that pass through  $\mathbf{z}_i$ , so  $D(\mathbf{z}_i)$  is the hyperplane  $H_i$  given by  $\theta_p = -x_{i,1}\theta_1 - \dots - x_{i,p-1}\theta_{p-1} + y_i$ . In dual space, the regression depth of a fit  $\boldsymbol{\theta}$  corresponds to the minimal number of hyperplanes  $H_i$  intersected by any halfline  $[\boldsymbol{\theta}, \boldsymbol{\theta} + \mathbf{u} >$  where  $\|\mathbf{u}\| = 1$ .

A unit vector  $\mathbf{u}$  in dual space thus corresponds to an affine hyperplane  $V$  in  $\mathbf{x}$ -space and a direction in which to tilt  $\boldsymbol{\theta}$  until it is vertical. For each unit vector  $\mathbf{u}$ , we therefore define the wedge-shaped region  $A_{\boldsymbol{\theta}, \mathbf{u}}$  in primal space, formed by tilting  $\boldsymbol{\theta}$  around  $V$  (in the direction of  $\mathbf{u}$ ) until the fit becomes vertical. Further denote  $H_n$  the empirical distribution of the observed data  $Z_n$ . Define the metric

$$\mu_H(H_n, H) = \sup_{\boldsymbol{\theta}, \|\mathbf{u}\|=1} |H_n(A_{\boldsymbol{\theta}, \mathbf{u}}) - H(A_{\boldsymbol{\theta}, \mathbf{u}})|.$$

If  $Z_n$  is sampled from  $H$ , then it holds that

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

This follows from the generalization of the Glivenko-Cantelli theorem formulated in (Pollard [17, Theorem 14]) and proved by (Pollard [17, Lemma 18]) and the fact that  $A_{\boldsymbol{\theta}, \mathbf{u}}$  can be constructed by taking a finite number of unions and intersections of half-spaces. Now define  $\Pi(\boldsymbol{\theta}) = \inf_{\mathbf{u}} H(A_{\boldsymbol{\theta}, \mathbf{u}})$  and its empirical version  $\Pi_n(\boldsymbol{\theta}) = \inf_{\mathbf{u}} H_n(A_{\boldsymbol{\theta}, \mathbf{u}})$ . It is clear that

$\Pi_n(\boldsymbol{\theta}) = rdepth(\boldsymbol{\theta}, Z_n)/n$  and  $\Pi(\tilde{\boldsymbol{\theta}}) = \frac{1}{2}$ . Moreover we have that  $|\Pi_n(\tilde{\boldsymbol{\theta}}) - \Pi(\tilde{\boldsymbol{\theta}})| \leq \mu_H(H_n, H)$  hence

$$\Pi_n(\tilde{\boldsymbol{\theta}}) = rdepth(\tilde{\boldsymbol{\theta}}, Z_n)/n \xrightarrow[n \rightarrow \infty]{a.s.} \frac{1}{2}. \quad (27)$$

Finally it holds that

$$\frac{rdepth(\tilde{\boldsymbol{\theta}}, Z_n)}{n} \leq \frac{\max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_n)}{n} \stackrel{a.s.}{\leq} \left\lceil \frac{n+p}{2} \right\rceil \frac{1}{n}. \quad (28)$$

The latter inequality uses Theorem 7 in (Rousseeuw and Hubert [20]) which is valid since  $Z_n$  is almost surely in general position. Taking limits in (28) and using (27) then finishes the proof.  $\square$

**Proof of Theorem 1:** We will first show that  $\liminf_n \varepsilon_n^* \geq \frac{1}{3}$  almost surely. From Lemma 1 we know that  $\{\boldsymbol{\theta}; rdepth(\boldsymbol{\theta}, Z_n) \geq p\}$  is bounded a.s. if  $Z_n \subset \mathbb{R}^p$  is sampled from  $H$ . Therefore, to break down the estimator we must add at least  $m$  observations such that  $rdepth(\boldsymbol{\xi}, Z_n) \leq p-1$  for any fit  $\boldsymbol{\xi}$  with maximal rdepth relative to  $Z_{n+m}$  (using the notation of the proof of Lemma 2). This implies that

$$r^* := rdepth(\boldsymbol{\eta}, Z_n) \leq rdepth(\boldsymbol{\xi}, Z_{n+m}) \leq p-1+m$$

where the first inequality is due to Lemma 3, and thus

$$\varepsilon_n^* \stackrel{a.s.}{\geq} \frac{m}{n+m} \geq \frac{r^* - p + 1}{n + r^* - p + 1}.$$

Finally we apply Lemma 4 to conclude that

$$\liminf_n \varepsilon_n^* \stackrel{a.s.}{\geq} \frac{1}{3} \quad (29)$$

for all fits  $\boldsymbol{\eta}$  and thus also for their average  $DR(Z_n)$ .

Next, we prove that  $\limsup_n \varepsilon_n^* \leq \frac{1}{3}$  almost surely. Since regression depth is affine invariant, we may assume that none of the  $\|\mathbf{x}_i\|$  in the dataset are zero. We will denote a hyperplane  $\boldsymbol{\theta}$  with equation  $y = \mathbf{x}^t \boldsymbol{\beta}_\theta + \alpha_\theta$  by its two components  $\boldsymbol{\beta}_\theta$  and  $\alpha_\theta$  corresponding to the slopes and the intercept. Fix two strictly positive real numbers  $\beta_0$  and  $\alpha_0$ . We now consider a point  $(\mathbf{x}^t, y)$  such that for any hyperplane  $\boldsymbol{\theta}$  passing through  $(\mathbf{x}^t, y)$  and  $p-1$  data points  $(\mathbf{x}_i^t, y_i)$  from  $Z_n$  it holds that  $\beta_0 < \|\boldsymbol{\beta}_\theta\| < \infty$  and  $\alpha_0 < |\alpha_\theta| < \infty$ . Because we assumed that none of the  $\|\mathbf{x}_i\|$  equals 0, we can always find such a point  $(\mathbf{x}^t, y)$ . Note that  $\mathbf{x} \neq \mathbf{x}_i$  for any  $i$ , otherwise  $\boldsymbol{\beta}_\theta$  would be unbounded.

The dataset  $Z_{n+m}$  is then obtained by enlarging the dataset  $Z_n$  with  $m = \lceil \frac{n+1+p}{2} \rceil - p + 2$  points equal to  $(\mathbf{x}^t, y)$ . We know that any fit  $\boldsymbol{\xi}$  with maximal rdepth must pass through at least  $p$  different observations of  $Z_{n+m}$ . Denote by  $\boldsymbol{\theta}^J$  any candidate fit with maximal rdepth. If  $\boldsymbol{\theta}^J$  passes through  $(\mathbf{x}^t, y)$  it is clear that  $rdepth(\boldsymbol{\theta}^J, Z_{n+m}) \geq m+p-1 = \lceil \frac{n+1+p}{2} \rceil + 1$ . On the other hand, any  $\boldsymbol{\theta}^J$  which passes through  $p$  data points of  $Z_n$  has  $rdepth(\boldsymbol{\theta}^J, Z_{n+m}) \leq \lceil \frac{n+1+p}{2} \rceil$ . This can be seen as follows. First consider the dataset  $Z_{n+1}$  which consists of the  $n$  original observations and one copy of the point  $(\mathbf{x}^t, y)$ . From Theorem 7 in (Rousseeuw and Hubert [20]) it follows that  $rdepth(\boldsymbol{\theta}^J, Z_{n+1}) \leq \lceil \frac{n+1+p}{2} \rceil$ . Now there always exists a unit vector  $\mathbf{u} \in \mathbb{R}^{p-1}$  and  $v \in \mathbb{R}$  such that  $\mathbf{x}^t \mathbf{u} = v$  and  $\mathbf{x}_i^t \mathbf{u} \neq v$  for all  $i = 1, \dots, n$ . Then the number of observations passed when tilting  $\boldsymbol{\theta}^J$  around  $(\mathbf{u}, v)$  as in Definition 1 plus the number of observations passed when tilting  $\boldsymbol{\theta}^J$  around  $(-\mathbf{u}, -v)$  equals  $n + 1 + p$  because the fit  $\boldsymbol{\theta}^J$  passes through exactly  $p$  observations. Therefore, we can suppose that the number of observations passed when tilting  $\boldsymbol{\theta}^J$  around  $(\mathbf{u}, v)$  is at most  $\lceil \frac{n+1+p}{2} \rceil$ . (If not, we replace  $(\mathbf{u}, v)$  by  $(-\mathbf{u}, -v)$ .) Note that the residual  $r_{(\mathbf{x}^t, y)}(\boldsymbol{\theta}^J) \neq 0$  because  $\boldsymbol{\theta}^J$  does not pass through  $(\mathbf{x}^t, y)$ . First suppose the residual  $r_{(\mathbf{x}^t, y)}(\boldsymbol{\theta}^J)$  is strictly positive. The data are in general position, therefore we can always find a value  $\varepsilon > 0$  such that for all  $i = 1, \dots, n$  it holds that  $\mathbf{x}_i^t \mathbf{u} < v - \varepsilon$  if  $\mathbf{x}_i^t \mathbf{u} < v$  and always  $\mathbf{x}_i^t \mathbf{u} > v - \varepsilon$  if  $\mathbf{x}_i^t \mathbf{u} > v$ . Since  $\mathbf{x}^t \mathbf{u} > v - \varepsilon$  and  $r_{(\mathbf{x}^t, y)}(\boldsymbol{\theta}^J) > 0$  the number of observations passed when tilting  $\boldsymbol{\theta}^J$  around  $(\mathbf{u}, v - \varepsilon)$  is the same as when tilting around  $(\mathbf{u}, v)$ . Finally, adding the other  $m - 1$  replications of  $(\mathbf{x}^t, y)$  does not change this value. Therefore  $rdepth(\boldsymbol{\theta}^J, Z_{n+m}) \leq \lceil \frac{n+1+p}{2} \rceil$ . If the residual  $r_{(\mathbf{x}^t, y)}(\boldsymbol{\theta}^J)$  is strictly negative, we replace  $v$  by  $v + \varepsilon$  in a similar way and obtain the same result.

The above reasoning shows that any fit with maximal rdepth must pass through  $(\mathbf{x}^t, y)$  and  $p - 1$  original data points. Since we have shown that all these fits have an arbitrarily large slope and intercept, it holds that

$$\varepsilon_n^* \stackrel{a.s.}{\leq} \frac{m}{n+m} = \frac{\lceil \frac{n+1+p}{2} \rceil - p + 2}{n + \lceil \frac{n+1+p}{2} \rceil - p + 2} \xrightarrow{n \rightarrow \infty} \frac{1}{3}$$

and thus

$$\limsup_n \varepsilon_n^* \stackrel{a.s.}{\leq} \frac{1}{3}. \quad (30)$$

From (29) and (30) we finally conclude (3).  $\square$

**Proof of Theorem 2.** Consider  $\mathbf{x} \in \mathbb{R}^{p-1}$ . We will prove that the upper and lower bounds in expression (14) are the values  $y$  such that in the dual plot the hyperplane  $(\mathbf{x}^t, 1)\boldsymbol{\theta} - y = 0$  is tangent to the ellipsoid  $E_{1-\alpha} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t \hat{\Sigma}^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = c^2$  where  $c = RD(\hat{\boldsymbol{\theta}}^{(j)})_{[(1-\alpha)m]:m}$ .

First consider the special case  $\hat{\boldsymbol{\theta}} = 0$  and  $\hat{\Sigma} = I_p$  yielding the unit sphere  $\boldsymbol{\theta}^t \boldsymbol{\theta} = 1$ . The tangent hyperplane in an arbitrary point  $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \dots, \tilde{\theta}_p)^t$  on the sphere is given by  $\tilde{\boldsymbol{\theta}}^t (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) = 0$ . Therefore the hyperplane  $(\mathbf{x}^t, 1)\boldsymbol{\theta} - y = 0$  becomes a tangent hyperplane iff  $(\mathbf{x}^t, 1) = (\tilde{\boldsymbol{\theta}}^t / \tilde{\theta}_p)$  and  $y = \tilde{\boldsymbol{\theta}}^t \tilde{\boldsymbol{\theta}} / \tilde{\theta}_p$  for some  $\tilde{\boldsymbol{\theta}}$  on the sphere. Together with  $\tilde{\boldsymbol{\theta}}^t \tilde{\boldsymbol{\theta}} = 1$  this yields  $y^2 = (\mathbf{x}^t, 1)(\mathbf{x}^t, 1)^t$  giving the lower bound  $y = -\sqrt{(\mathbf{x}^t, 1)(\mathbf{x}^t, 1)^t}$  and the upper bound  $y = \sqrt{(\mathbf{x}^t, 1)(\mathbf{x}^t, 1)^t}$  corresponding to expression (14) for this case.

Consider the general case of an ellipsoid  $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t \hat{\Sigma}^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = c^2$  and denote  $\hat{\Sigma} = P\Lambda P^t$  where  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$  is the diagonal matrix of eigenvalues of  $\hat{\Sigma}$  and  $P = (\mathbf{e}_1, \dots, \mathbf{e}_p)$  is the matrix of eigenvectors of  $\hat{\Sigma}$ . We can transform this to the previous case by the transformation  $\tilde{\boldsymbol{\theta}} = c\Lambda^{-1/2}P^t(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ . The hyperplane  $(\mathbf{x}^t, 1)\boldsymbol{\theta} - y = 0$  then transforms to  $c(\mathbf{x}^t, 1)P\Lambda^{1/2}\tilde{\boldsymbol{\theta}} - (y - (\mathbf{x}^t, 1)\hat{\boldsymbol{\theta}}) = 0$  which becomes a tangent hyperplane if  $(y - (\mathbf{x}^t, 1)\hat{\boldsymbol{\theta}})^2 = (c(\mathbf{x}^t, 1)P\Lambda^{1/2})(c(\mathbf{x}^t, 1)P\Lambda^{1/2})^t = c^2(\mathbf{x}^t, 1)P\Lambda P^t(\mathbf{x}^t, 1)^t = c^2(\mathbf{x}^t, 1)\hat{\Sigma}(\mathbf{x}^t, 1)^t$ . This yields the lower bound  $y = (\mathbf{x}^t, 1)\hat{\boldsymbol{\theta}} - c\sqrt{(\mathbf{x}^t, 1)\hat{\Sigma}(\mathbf{x}^t, 1)^t}$  and the upper bound  $y = (\mathbf{x}^t, 1)\hat{\boldsymbol{\theta}} + c\sqrt{(\mathbf{x}^t, 1)\hat{\Sigma}(\mathbf{x}^t, 1)^t}$  of expression (14).  $\square$

**Proof of Proposition 1.** Denote  $r_i(f, (\mathbf{x}_i, y_i)) = y_i - f(\mathbf{x}_i)$  and  $r_i(g(f), (\mathbf{x}_i, y'_i)) = y'_i - g(f(\mathbf{x}_i))$ . From the monotonicity of  $g$  it follows that

$$\begin{aligned} \text{sign}(r_i(f, (\mathbf{x}_i, y_i))) &= \text{sign}(y_i - f(\mathbf{x}_i)) \\ &= \text{sign}(g(y_i) - g(f(\mathbf{x}_i))) \\ &= \text{sign}(r_i(g(f), (\mathbf{x}_i, y'_i))). \end{aligned}$$

Since only the responses are transformed, it follows from Definition 3 that

$$rdepth(f, Z_n) = rdepth(g(f), Z'_n). \quad \square$$

**Proof of Proposition 2.** This follows immediately from Proposition 1 which implies  $rdepth(\boldsymbol{\theta}, \tilde{Z}_n) = rdepth(g_{\boldsymbol{\theta}}, Z_n)$  for all possible fits  $\boldsymbol{\theta}$ .

**Proof of Theorem 3.** Let us denote the regression depth of a polynomial fit  $\boldsymbol{\theta}$  of degree  $k$  by  $rdepth_k(\boldsymbol{\theta}, Z_n)$ . From the depth of a polynomial fit given by Definition 3 it follows for any data set  $Z_n = \{(x_i, y_i); i = 1, \dots, n\}$  with distinct  $x_i$  as in Lemma 1 that  $\{\boldsymbol{\theta}; rdepth_k(\boldsymbol{\theta}, Z_n) \geq k + 1\}$  is bounded. Now any bivariate linear fit  $y = \tilde{\theta}_1 x_1 + \tilde{\theta}_2$  corresponds to a polynomial fit  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{k+1})^t$  with  $\theta_1 = \tilde{\theta}_1$ ,  $\theta_2 = \dots = \theta_k = 0$  and  $\theta_{k+1} = \tilde{\theta}_2$  for  $k \geq 1$ . Since it holds for any line  $\boldsymbol{\eta}$  with maximal regression depth that  $rdepth(\boldsymbol{\eta}, Z_n) = rdepth_k(\boldsymbol{\eta}, Z_n) \geq \lceil n/3 \rceil$

(Rousseeuw and Hubert [20]), it follows that any polynomial fit  $\gamma$  of degree  $k$  with maximal regression depth has  $rdepth_k(\gamma, Z_n) \geq rdepth_k(\eta, Z_n) \geq \lceil n/3 \rceil$ .

Because  $\{\theta; rdepth_k(\theta, Z_n) \geq k + 1\}$  is bounded, we must add at least  $m$  observations such that  $rdepth_k(\xi, Z_n) \leq k$  where  $\xi$  is a polynomial fit with maximal rdepth to  $Z_{n+m}$  to break down the estimator. We obtain

$$\left\lceil \frac{n+m}{3} \right\rceil \leq rdepth_k(\xi, Z_{n+m}) \leq k+m$$

from which it follows that  $m \geq (n - 3k)/2$ . This yields

$$\varepsilon_n^*(DR_k, Z_n) = \varepsilon_n^*(\gamma, Z_n) \geq \frac{m}{n+m} \geq \frac{n-3k}{3n-3k}. \quad \square$$

**Proof of Theorem 4.** We will show that when  $s_i > 0$  for all  $i = 1, \dots, n$  it holds for every  $\theta = (\theta_1, \theta_2)^t$  that  $rdepth((\theta_1, \theta_2)^t, \{(\frac{1}{s_i}, \frac{1}{v_i}); i = 1, \dots, n\}) = rdepth((\theta_2, \theta_1)^t, \{(s_i, \frac{s_i}{v_i}); i = 1, \dots, n\})$ . This follows from

$$\begin{aligned} r_i((\theta_1, \theta_2)^t, \{(\frac{1}{s_i}, \frac{1}{v_i}); i = 1, \dots, n\}) &= \frac{1}{v_i} - \theta_1 \frac{1}{s_i} - \theta_2 \\ &= \frac{1}{s_i} (\frac{s_i}{v_i} - \theta_1 - \theta_2 s_i) \\ &= \frac{1}{s_i} r_i((\theta_2, \theta_1)^t, \{(s_i, \frac{s_i}{v_i}); i = 1, \dots, n\}). \end{aligned}$$

Since  $s_i > 0$  for all  $i = 1, \dots, n$  we have

$$\text{sign}(r_j((\theta_1, \theta_2)^t, \{(\frac{1}{s_i}, \frac{1}{v_i}); i = 1, \dots, n\})) = \text{sign}(r_j((\theta_2, \theta_1)^t, \{(s_i, \frac{s_i}{v_i}); i = 1, \dots, n\}))$$

for all  $j = 1, \dots, n$ , and switching the  $x$ -values from  $\frac{1}{s_i}$  to  $s_i$  reverses their order. Therefore, according to Definition 1 both depths are the same.  $\square$

## References

1. N. Amenta, M. Bern, D. Eppstein, and S. Teng, Regression depth and center points, *Discrete and Computational Geometry*, **23** (2000), 305-323.
2. Z. Bai and X. He, Asymptotic distributions of the maximal depth estimators for regression and multivariate location, *Ann. Statist.*, **27** (1999), 1616-1637.
3. J. Benderly and B. Zwick, Inflation, real balances, output, and real stock returns, *American Economic Review* (1985), p. 1117.

4. N. A. C. Cressie and D. D. Keightley, The underlying structure of the direct linear plot with application to the analysis of hormone-receptor interactions, *Journal of Steroid Biochemistry*, **11** (1979), 1173-1180.
5. N. A. C. Cressie and D. D. Keightley, Analysing data from hormone-receptor assays, *Biometrics*, **37** (1981), 235-249.
6. H. E. Daniels, A distribution-free test for regression parameters, *Ann. Math. Statist.*, **25** (1954), 499-513.
7. D. L. Donoho and M. Gasko, Breakdown properties of location estimates based on halfspace depth and projected outlyingness, *Ann. Statist.*, **20** (1992), 1803-1827.
8. D. L. Donoho and P. J. Huber, The Notion of Breakdown Point in "A Festschrift for Erich Lehmann" (P.J. Bickel, K.A. Doksum and J.L. Hodges, Eds.), pp 157-184, Belmont, Wadsworth, 1983.
9. J. B. Haldane, Graphical methods in enzyme chemistry, *Nature*, **179** (1957), 832.
10. D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski, *A Handbook of Small Data Sets*, New York: Chapman and Hall, 1994.
11. X. He and S. Portnoy, Asymptotics of the deepest line, in "Applied Statistical Science III: Nonparametric Statistics and Related Topics" (S.E. Ahmed, M. Ahsanullah, and B.K. Sinha, Eds.), pp. 71-81, Nova Science Publishers Inc, New York, 1998.
12. S. Langerman and W. Steiger, An optimal algorithm for hyperplane depth in the plane, in "Proc. 11th Symp. Discrete Algorithms," ACM and SIAM, pp 54-59, 2000.
13. H. Lineweaver and D. Burk, The determination of enzyme dissociation constants, *Journal of the American Chemical Society*, **56** (1934), 658-666.
14. R. D. Martin, V. J. Yohai and R. H. Zamar, Min-max bias robust regression, *Ann. Statist.*, **17** (1989), 1608-1630.
15. I. Mizera, On Depth and Deep Points: a Calculus, Technical report, 1999.
16. I. Mizera and M. Volauf, Continuity of halfspace depth contours and maximum depth estimators: diagnostics of depth-related methods, Technical report, 1999.

17. D. Pollard, *Convergence of Stochastic Processes*, New York: Springer-Verlag, 1984.
18. P. J. Rousseeuw, Least median of squares regression, *J. Amer. Statist. Assoc.*, **79** (1984), 871-880.
19. P. J. Rousseeuw, Multivariate estimation with high breakdown point, in "Mathematical Statistics and Applications, Vol. B" (W. Grossmann, G. Pflug, I. Vincze and W. Wertz, Eds.), pp. 283-297, Reidel, Dordrecht, 1985.
20. P. J. Rousseeuw and M. Hubert, Regression Depth, *J. Amer. Statist. Assoc.*, **94** (1999), 388-402.
21. P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, New York: Wiley-Interscience, 1987.
22. P. J. Rousseeuw and A. Struyf, Computing location depth and regression depth in higher dimensions, *Statistics and Computing*, **8** (1998), 193-203.
23. P. J. Rousseeuw and K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41** (1999), 212-223.
24. P. J. Rousseeuw and V. J. Yohai, Robust regression by means of S-estimators, in "Robust and Nonlinear Time Series Analysis," Lecture Notes in Statistics No. 26 (J. Franke, W. Härdle and R.D. Martin, Eds.), pp. 256-272, Springer, New York, 1984.
25. G. Scatchard, The attractions of proteins for small molecules and ions, *Annals of the New York Academy of Sciences*, **51** (1949), 660-672.
26. S. Van Aelst and P. J. Rousseeuw, Robustness of deepest regression, *J. Multivariate Anal.*, **73** (2000), 82-106.
27. H. Van den Bergh, Zie de zoo, *Hamster*, **8** (1968), De vrienden van het Schoolmuseum, M. Thiery.
28. M. Van Kreveld, J. S. Mitchell, P. J. Rousseeuw, M. Sharir, J. Snoeyink, and B. Speckmann, Efficient algorithms for maximum regression depth, in "Proceedings of the 15th Symposium on Computational Geometry," ACM (Association for Computing Machines), pp. 31-40, 1999.

29. Y. Zuo, Some quantitative relationships between two types of finite sample breakdown point, Technical report, 1999.