

We would like to thank the discussants for their wide-ranging comments, criticisms, and questions. Most discussants focus on the deepest regression (DR) method.

Since nearly everybody made a comment about computation time, we'll address this point up front. In simple regression ( $p = 2$ ), the naive  $O(n^3)$ -time algorithm for the DR line left a lot of room for improvement. In early 1998, Clive Loader (personal communication) suggested and implemented an approximate algorithm for the DR line in  $O(n^2)$  time. Later, collaborative work with several specialists of computational geometry yielded an exact algorithm of complexity  $O(n \log^2 n)$ , i.e. little more than linear time (van Kreveld et al. 1999).

For higher dimensions ( $p \geq 3$ ) we also mentioned a naive algorithm for DR in our paper, to get started. The exact algorithm constructed in (Rousseeuw and Struyf 1998) for computing the rdepth of a given  $\theta$  in  $\mathbb{R}^p$  has complexity  $O(n^{p-1} \log n)$ . Fortunately, the same paper also contains a much faster approximate algorithm. Also, we would not try out all elemental fits  $\theta_J$  where  $J$  is any  $p$ -subset of  $Z_n$ . We are currently working on a fast approximate algorithm for DR. Intuitively it should be possible to approximate the DR in less time than other positive-breakdown estimators, that are faced with local optima which are hard to overcome. The DR behaves like a median and is 'nearly monotone' to movements in the data, so iteration schemes based on some type of local improvement steps might do the job.

## 1. HE

Professor He makes several constructive suggestions. His first section explains that the depth envelope  $E_k$  can be seen as a simultaneous confidence band for  $\tilde{y} = \tilde{\theta}_1 x + \tilde{\theta}_2$  where  $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)$  is the unknown true parameter, i.e. the conditional median line. Under assumption (2.5) of our paper, 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  $F_n(k) := P(\text{rdepth}(\tilde{\theta}, Z'_n) \leq k)$  where  $Z'_n$  has the same  $\{x_1, \dots, x_n\}$  as the actual data set

$Z_n$ . By invariance properties,

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

where the  $e_i$  are i.i.d. from (say) the standard gaussian. Thus we can compute  $F_n(k)$  by simulating (1), as proposed by He. When there are *no ties* among the  $x_i$  we can compute  $F_n(k)$  exactly from formula (4.4) in Daniels (1954), yielding

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

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 (2) can easily be extended to large  $n$ . Note that (2) also holds for location depth: it gives  $P(\text{ldepth}(\tilde{\theta}, X_n) \leq k)$  where  $X_n$  comes from a distribution that is angularly symmetric about  $\tilde{\theta}$  and has a density. We have implemented (2) as an S-Plus function, so there is no need for tables any more. But note that (2) is restricted to the *bivariate* case without ties in  $\{x_1, \dots, x_n\}$ , so in all other situations we use (1).

For the Skeena River data, He's Figure 1 shows the depth envelope  $E_7$ . Since there are no ties among these  $x_i$  we can apply (2) to compute the confidence as  $P(\text{rdepth} \geq 7) = 1 - F_{28}(6) = 1 - 0.045 = 95.5\%$ . (We can easily list the confidence  $1 - F_n(k - 1)$  for each  $E_k$  of a given data set.) Note that  $E_7$  looks rather wide, but in fact corresponds to a relatively small confidence region for  $\tilde{\theta}$  in parameter space. Also note that the intersection  $E_7$  with a vertical line  $x = x_0$  is not a 95.5% 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  $(\theta_1, \theta_2)$  in a 95.5% confidence region for  $(\theta_1, \theta_2)$ .

Regression depth also allows us to test one or several regression coefficients. To test the significance of the slope ( $H_0 : \tilde{\theta}_1 = 0$ ) we compute  $\max \text{rdepth}((\theta, \theta_2), Z_n)$  over  $\theta_2$ . This is easy, because we only have to compute the rdepth of all horizontal lines passing through an observation. For the Skeena data, the maximal  $\text{rdepth}((\theta, \theta_2), Z_n)$  equals 9

Stefan Van Aelst is Research Assistant with the Fonds voor Wetenschappelijk Onderzoek (FWO), Belgium. He joined the team after the original version of our paper was submitted, and has done a lot of work on this subject since. The authors are grateful to the Editor for organizing this thorough discussion.

(and is attained at  $\theta_2 = 1381$ ). Therefore the corresponding  $p$ -value is  $P(\text{rdepth}(\tilde{\theta}, Z'_n) \leq 9) = F_{28}(9) = 0.51$  so  $H_0$  is not rejected. This  $p$ -value 0.51 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  $\theta_2 = 0$  we compute the maximal  $\text{rdepth}((\theta_1, 0), Z_n) = 10$  by considering all lines through the origin and an observation, yielding the  $p$ -value  $F_{28}(10) = 0.78$  which is not significant either. However, the combined null hypothesis  $(\theta_1, \theta_2) = (0, 0)$  yields  $\text{rdepth}((0, 0), Z_n) = 0$  with  $p$ -value  $F_{28}(0) < 0.0001$  which is highly significant.

These confidence regions and tests extend to higher dimensions. The Nuclear Power data (Figure 6 in our paper) has  $n = 32$  and  $p = 3$ , so we can compute  $F_n(k)$  from (1). The indicated envelope  $E_6$  has confidence  $1 - F_n(5) = 98.5\%$ . (The smaller envelope  $E_7$  has confidence 94.2% here.) For testing the null hypothesis  $(\tilde{\theta}_1, \tilde{\theta}_2) = (0, 0)$  that both slopes are zero (this would be done by  $R^2$  in LS regression) we compute the maximal  $\text{rdepth}((0, 0, \theta_3), Z_n)$  over all  $\theta_3$  (i.e. over all  $y_i$  in the data set). By computing the exact  $\text{rdepth}$  of these 32 horizontal planes (by the fast algorithm of Rousseeuw and Struyf 1998) we obtain 9, hence the  $p$ -value is  $F_n(9) = 0.18$ , not significant.

Note that the envelopes  $E_k$  are somewhat conservative because the exact confidence region in dual space need not be convex, so mapping it to the data space yields  $E_k$  minus perhaps some ‘holes’ at the side. We prefer to plot the entire  $E_k$  which is easier to interpret, and anyway the difference between the conservative coverage and the exact coverage goes down quickly with  $n$ .

We are grateful to Xuming He for suggesting that we compare the  $E_k$  to bootstrapped regression confidence regions, corresponding to the work of Yeh and Singh (1997, page 648) for bootstrapped location estimates. Each of the 500 dots in Figure 1 is the deepest regression (DR) estimate  $(\hat{\theta}_1, \hat{\theta}_2)$  of a data set obtained by randomly drawing 28 observations, with replacement, from the Skeena data.

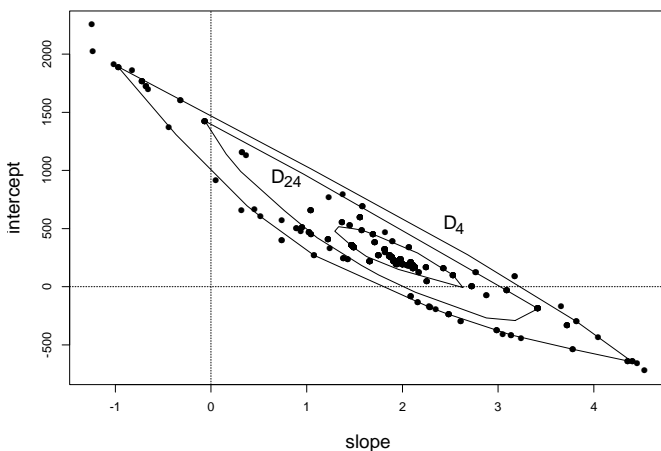


Figure 1. Deepest regression estimates of 500 bootstrap samples from the Skeena data, with the region  $D_4$  of location depth  $\geq 4$  which contains 95.8% of the estimates.

Note that there are a few outlying estimates  $(\hat{\theta}_1, \hat{\theta}_2)$  in Figure 1, which is natural for the bootstrap since some bootstrap samples contain relatively many coinciding observations. Therefore we don't draw the usual ellipse based on the classical mean and covariance matrix, but we use the more robust location depth. Using the algorithm of Struyf and Rousseeuw (1998) we find the deepest location in Figure 1 to be (1.96, 215.31) which corresponds well to the DR fit  $y = 1.99x + 217.46$  obtained in our paper. As a confidence region for  $(\tilde{\theta}_1, \tilde{\theta}_2)$  we take a location depth contour  $D_k$  relative to the 500 estimates such that  $D_k$  contains roughly 95% of them. This yielded  $k = 4$  with coverage 95.8%. It turns out that the corresponding confidence band

$$\{(x, y) \in \mathbb{R}^2; \min(\theta_1 x + \theta_2) \leq y \leq \max(\theta_1 x + \theta_2)\} \text{ where } (\theta_1, \theta_2) \in D_k\}$$

is somewhat narrower than  $E_7$ . In Figure 1 we also see that the lines  $\theta_1 = 0$  and  $\theta_2 = 0$  amply intersect  $D_4$ , so neither  $\theta_1$  or  $\theta_2$  are significant at the 5% level. (The line  $\theta_1 = 0$  barely intersects  $D_{24}$ , and the line  $\theta_2 = 0$  intersects  $\tilde{D}_{57}$ .) The point (0, 0) however does lie outside  $D_4$  hence  $(\tilde{\theta}_1, \tilde{\theta}_2)$  is again significantly different from (0, 0) at the 5% level.

He's second section proposes a faster exact algorithm for the DR fit. It starts by enumerating all the  $N = \binom{n}{p}$  elemental fits, but unlike the naive algorithm it does not compute the  $\text{rdepth}$  of all of them. Instead it uses the fact that  $\text{rdepth}(\theta) \leq \min\{d_1(\theta), \dots, d_p(\theta)\}$  where each  $d_j(\theta)$  can be computed in  $O(n)$  time. After computing the actual  $\text{rdepth}$  of the first set of  $N_0 < N$  elemental fits and denoting the largest of them by  $d_0$ , only the remaining elemental fits  $\theta$  with  $\min\{d_1(\theta), \dots, d_p(\theta)\} > d_0$  need to be considered. Although this approach is faster than the naive algorithm (as illustrated by the Nuclear Power example with  $n = 32$  and  $p = 3$ ), it is not clear that the worst-case time complexity decreases (this depends on how often one has to compute  $\text{rdepth}(\theta)$  when  $n$  increases). In any case, the complexity is still at least  $O(n^{p+1})$  due to the  $O(n^p)$  elemental subsets and the  $O(n)$  time to compute  $d_1(\theta)$ . The *approximate* algorithm proposed by He is also at least  $O(n^{p+1})$ , but it is only intended for data from the semiparametric linear model because it relies on  $\max \text{rdepth}(\theta, Z_n) \approx n/2$  (whereas for curved data this value may be much lower). We hope that further work, possibly using He's ideas, will yield algorithms with lower time complexity.

We are grateful to He and Portnoy (1998) and Bai and He (1998) for obtaining the asymptotics of the DR method. Xuming He mentions efficiency comparisons between DR and other estimators at various distributions of  $(x, y)$ . It would be useful if one could compute (e.g. numerically) the covariance matrix of the limiting distribution, e.g. when  $x$  is gaussian. We are aware that this is not yet part of the empirical process toolbox, but it would be a most important addition to it. It has hampered our research a few times already that the second moment of limiting distributions of this type is not known (e.g. for the LMS estimator, even though it is known that all its moments exist). This stands in the way of making efficiency comparisons,

tests, and confidence regions. It would also be useful to have probability contours or other multivariate quantiles of this type of distributions.

## 2. KOENKER

We are encouraged by Professor Koenker's warm welcome of the notion of regression depth, which he calls 'an even more robust strategy for accomplishing the quantile regression task'.

Koenker's Section 2 is about the efficiency of the deep regression (DR) relative to that of  $L^1$ . For univariate data the ARE is 100% because both estimates equal the median, and for simple regression with gaussian  $x_i$  the slope estimates satisfy  $\text{ARE}(\text{DR}, L^1) \approx 88\%$  (He and Portnoy 1998). In his comment, Roger Koenker considers the submodel of simple regression through the origin, where  $\text{ARE}(\text{DR}, L^1)$  is 64% for gaussian  $x_i$  and less for longer-tailed  $x$ -distributions. (Of course, one could always increase the efficiency of the DR by following it by one or a few M-steps or similar techniques.) But from the robustness point of view the DR slope is hard to beat in this model, where it is the minimax bias estimator and hence its breakdown value is 50%. The  $L^1$  slope has breakdown value 0% according to the usual definition which allows to replace a point  $(x_i, y_i)$  by any point  $(\tilde{x}_i, \tilde{y}_i)$ . The breakdown values in Koenker's Table 1 are higher, and correspond to replacing  $(x_i, y_i)$  by  $(x_i, \tilde{y}_i)$ .

Prof. Koenker then mentions that for any *fixed* design for which the DR slope is consistent also the  $L^1$  slope is consistent, but that the converse is not true in some pathological examples such as  $x_i = 1/\sqrt{i}$ . The latter design is indeed pathological, because the closer  $x_i$  comes to zero the less information it carries (with  $x_i = 0$  carrying *no* information). For  $x_i = 1/i$  neither the DR nor the  $L^1$  slope are consistent. Of course no statistician would construct a design with  $x_i \rightarrow 0$  when there are so many better designs, but Roger's point is that we might come in a situation where someone else has made the measurements according to such a pathological design, and we have to analyze that data.

On the other hand, we wonder whether there are situations with *random*  $\mathbf{x}_i$  where DR is consistent but  $L^1$  is not. Let us take a framework where we all agree on what should be estimated, the linear conditional median. Our semiparametric model  $\mathcal{H}$  consists of all distributions  $H$  on

$\mathbb{R}^p$  with a strictly positive density  $h$  and such that there exists  $\tilde{\boldsymbol{\theta}}$  such that

$$\text{med}[y - (\mathbf{x}, 1)\tilde{\boldsymbol{\theta}}' | \mathbf{x}] = 0. \quad (3)$$

The assumption  $h(\mathbf{x}, y) > 0$  is to make the conditional median (3) unique; otherwise the parameter  $\tilde{\boldsymbol{\theta}}$  might not be identifiable, i.e. (3) could be satisfied by some  $\boldsymbol{\theta}^* \neq \tilde{\boldsymbol{\theta}}$  as well. From our Theorem 7(c) we know that

$$r\text{depth}(\boldsymbol{\theta}, H) = \inf_{\substack{\mathbf{u} \in S \\ t \in \mathbb{R}}} E_H[\text{sgn}(y - (\mathbf{x}, 1)\boldsymbol{\theta}') \text{sgn}(\mathbf{x}\mathbf{u}' - t)] \quad (4)$$

(where  $S$  is the unit sphere in  $\mathbb{R}^{p-1}$ ) attains its upper bound  $1/2$  at  $\tilde{\boldsymbol{\theta}}$ . (For regression through the origin we replace  $(\mathbf{x}, 1)\boldsymbol{\theta}'$  by  $\mathbf{x}\boldsymbol{\theta}'$  everywhere, and put  $t = 0$ .) Van Aelst and Rousseeuw (1998) prove that any other  $\boldsymbol{\theta} \neq \tilde{\boldsymbol{\theta}}$  has  $r\text{depth}(\boldsymbol{\theta}) < 1/2$  hence  $\text{DR}(H) = \tilde{\boldsymbol{\theta}}$ . This implies that the DR is always Fisher-consistent.

Note that  $|\text{sgn}(y - (\mathbf{x}, 1)\boldsymbol{\theta}') \text{sgn}(\mathbf{x}\mathbf{u}' - t)| \leq 1$  so the objective (4) exists for *any* distribution  $H$  on  $\mathbb{R}^p$ . On the other hand, the  $L^1$  is defined by minimizing

$$A(\boldsymbol{\theta}, H) = E_H[|y - (\mathbf{x}, 1)\boldsymbol{\theta}'|] \quad (5)$$

which does not exist when the error term does not have a first moment. In the case of  $L^1$  location this is adequately solved by minimizing  $E_H[\|\mathbf{x} - \boldsymbol{\theta}\| - \|\mathbf{x}\|]$  instead of  $E_H[\|\mathbf{x} - \boldsymbol{\theta}\|]$ , since  $\|\mathbf{x} - \boldsymbol{\theta}\| - \|\mathbf{x}\| \leq \|\boldsymbol{\theta}\| < \infty$  for all  $\boldsymbol{\theta}$ . Trying the same here, we would minimize

$$B(\boldsymbol{\theta}, H) = E_H[|y - (\mathbf{x}, 1)\boldsymbol{\theta}'| - |y|] \quad (6)$$

but now  $||y - (\mathbf{x}, 1)\boldsymbol{\theta}'| - |y|| \leq |(\mathbf{x}, 1)\boldsymbol{\theta}'|$  so we need  $E[\|\mathbf{x}\|]$  to exist! Another approach is to define  $L^1$  by setting

$$C(\boldsymbol{\theta}, H) = E_H[(\mathbf{x}, 1)\text{sgn}(y - (\mathbf{x}, 1)\boldsymbol{\theta}')] \quad (7)$$

equal to  $\mathbf{0}$ , but then we also need  $E[\|\mathbf{x}\|]$  to exist (like for any other zero-breakdown M-, L-, or R-estimator). It remains an open question whether  $L^1$  regression can be written as a functional in a way that it still exists for distributions where the errors and the  $\mathbf{x}$  have no moments. In this sense, the DR is a more 'median-like' method than the  $L^1$ , since the DR exists at any distribution and is Fisher-consistent at the entire model  $\mathcal{H}$ .

For a specific example, let us return to simple regression through the origin. Consider three independent standard gaussian variables  $z_1, z_2$  and  $z_3$ . Then  $H$  is defined as the distribution of  $(x, y)$  where  $x = (z_1/z_3)^2 \text{sgn}(z_1/z_3)$  and  $y = (z_2/z_3)^2 \text{sgn}(z_2/z_3)$ . Note that  $H$  is related to the spherical bivariate Cauchy distribution (see Johnson and Kotz 1972, pages 133-134). The marginal distributions of  $x$  and  $y$  have the same distribution function  $F(x) = 1/2 + \text{sgn}(x)(1/\pi)\text{Arctan}(\sqrt{|x|})$  and symmetric density

$$f(x) = \frac{1}{2\pi\sqrt{|x|}} \frac{1}{1 + |x|}, \quad (8)$$

which has the same tails as the Pareto distribution with exponent  $\alpha = 1/2$ . This bivariate  $H$  belongs to the model

Table 1. Dispersion of the sampling distribution of the DR method and the  $L^1$  method, applied to  $m = 10,000$  samples from  $H$  for various sample sizes  $n$ .

$n$	$\text{med}_{j=1}^m  \text{DR}^{(j)} $	$\text{med}_{j=1}^m  L_1^{(j)} $
100	.0120	.217
200	.00576	.209
300	.00378	.214
400	.00277	.202
500	.00234	.209
1,000	.00111	.218
5,000	.000230	.200
10,000	.000113	.206

$\mathcal{H}$  with  $\tilde{\theta} = 0$ , so  $f(x)$  in (8) is also the density of the error term  $e = y - \tilde{\theta}x$ . Since DR is Fisher-consistent,  $\text{DR}(H) = \tilde{\theta} = 0$ . It is also consistent in the usual sense, because  $\text{DR}_n = \text{med}_{i=1}^n(y_i/x_i) = \text{med}_{i=1}^n(t_i)$  where the  $t_i$  are i.i.d. according to the same distribution (8) with  $f(0) = \infty$ . Note that the distribution of  $\text{DR}_n$  is symmetric and converges at the rate  $1/n$  (which is much faster than the usual rate  $1/\sqrt{n}$ ) because of the long-tailed  $x$ -distribution. Table 1 is the result of simulating  $m = 10,000$  samples from this bivariate distribution  $H$ , for different sample sizes  $n$ . It lists  $\text{med}_{j=1}^m |\text{DR}_n^{(j)}|$  where  $\text{DR}_n^{(j)}$  is the DR estimate applied to sample  $j$ . The consistency is reflected in the fact that this dispersion goes to zero (like  $1/n$ ) for increasing  $n$ .

We don't know whether one can write down an  $L^1$  functional that exists at this  $H$ . Even if such a functional would not exist, this does not preclude the finite-sample  $L^1$  estimator to be consistent for  $\tilde{\theta} = 0$ . But we doubt it, because in Table 1 the dispersion of the sampling distribution of the  $L^1$  estimates shows no sign of decreasing with  $n$ .

Of course, data from such a Pareto-tailed bivariate distribution are about as unlikely to occur in practice as data that were obtained with the pathological design  $x_i = 1/\sqrt{i}$ . Still, such examples teach us a lot about the mathematical fundamentals and applicability limits of a method.

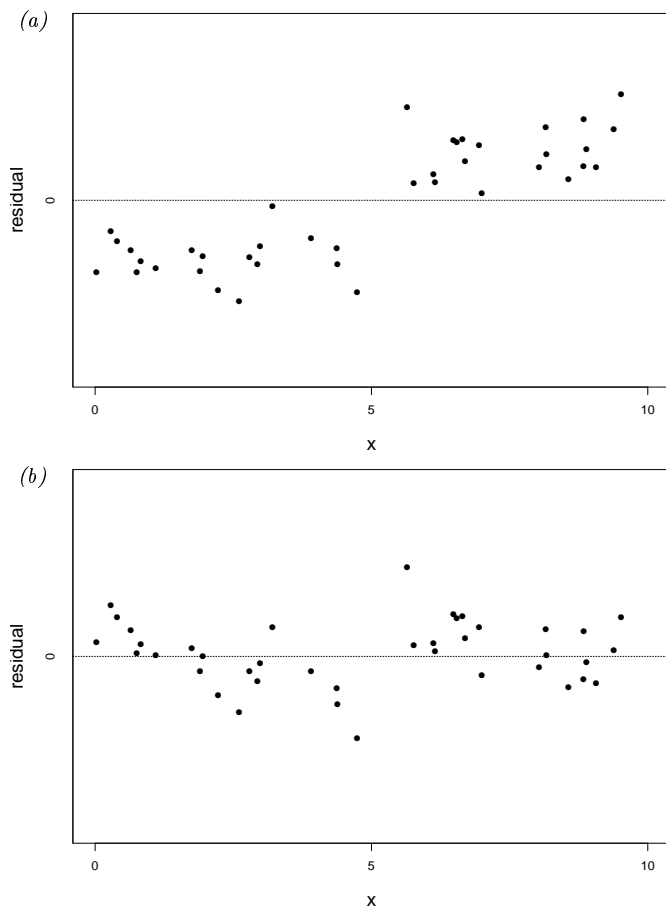


Figure 2. Data of Example 1. Residuals of (a) line 1 with  $rdepth=0$ , and (b) the DR line with  $rdepth=16$ .

In Section 1, Roger Koenker focuses on the monotone equivariance of the DR. To further clear up the paradox, we note that the stated monotone equivariance of the  $L^1$  method concerns the  $L^1$  functional (where it exists) at distributions  $H$  in the semiparametric model  $\mathcal{H}$ , but that it does not hold at distributions outside  $\mathcal{H}$ . For instance, the empirical distribution  $H_n$  of a finite sample does not belong to  $\mathcal{H}$ . On the other hand, the monotone equivariance of the DR functional holds at *any* distribution  $H$ , including  $H_n$ .

We were very impressed with the gigantic speedup of the  $L^1$  regression computation in Portnoy and Koenker (1998), and we hope that similar ideas may benefit the DR as well. It has taken a long time from the inception of  $L^1$  (around 1750) to the current fast algorithm for its computation, and the DR has just started, but we hope to learn from other people's great ideas!

### 3. LIU AND SINGH

Professors Liu and Singh convincingly argue that the time has come to see depth as a practical tool for data analysis. For this they have compiled a substantial list of applications of location depth, ranging from exploratory statistics to rank tests and multivariate quality control, and we are eagerly awaiting their review paper (Liu, Parelius and Singh 1997) for more.

Regina and Kesar are for the most part positive about the new notion of regression depth, and (like us) they appreciate the similarity between the properties of  $ldepth$  and  $rdepth$ . For instance, the notion of angular symmetry plays the same role for  $ldepth$  as our condition (3.7) in Theorem 7(c), which we might call *conditional median linearity*, does for  $rdepth$ . It turns out that there is a fundamental connection between these notions (Rousseeuw 1999).

On the other hand, Regina and Kesar give three examples where they feel that  $rdepth$  does not behave the way it should. We are convinced that these are misunderstandings due to parts of our paper that were perhaps unclear, and which therefore many readers may have problems with.

**Example 1.** In Figure 1 of Regina and Kesar we indeed see a roughly linear pattern with a small error scale. But when looking at the lines 1, 2 and 3 in their Figure 2, the intercept shift becomes clearly visible. To us line 1 does not seem like a good fit, because for  $x_i < 5$  all its residuals are strictly negative and for  $x_i > 5$  all  $r_i > 0$ . Looking at its residual plot in Figure 2a here, we would not conclude that the fit is good. Of course, this corresponds exactly to Definition 2 of a nonfit in our paper. The discussants' preference for line 1 may be due to thinking in terms of an objective function based on the absolute magnitudes of the residuals, whereas  $rdepth$  is concerned with their signs. For line 3 the residual plot (not shown here) looks strange too, because all  $r_i > 0$  except for two points with a negative residual close to zero (hence its  $rdepth$  is 2). If line 3 would lie just a tiny bit lower, all its  $r_i > 0$  so it would be a nonfit too. Line 2 is somewhat more balanced

since it has some positive and negative residuals for  $x_i < 5$  (but still only positive residuals for  $x_i > 5$ ). The deepest regression (DR) line is by definition the most balanced for a given data set, as you can see from its residual plot in Figure 2b here, with  $rdepth=16$ . (Since we didn't have Regina and Kesar's data, our data was simulated according to their description.) In the scatterplot of the data, the DR line lies in essentially the same position as the LS line in the discussants' Figure 3. The slightly higher slope they reported seems to be due to taking only one data set; when we generated many data sets like this we found that DR and LS had the same slope on average, which is understandable because there are no outliers here.

**Example 2.** The discussants consider uniformly distributed data on the unit disk in the plane, with the model of regression through the origin. They note that the DR is the horizontal line  $y = 0x$ , whereas the  $rdepth$  of a line  $y = \theta x$  goes to zero as the line rotates towards the vertical. This is exactly how it should be, since the data distribution has a linear conditional median  $\text{med}[y|x] = 0$  (hence, also the  $L^1$  fit is the horizontal line). The conditional expectation  $E[y|x]$  is zero too, hence the LS fit is the horizontal line as well. In fact, most regression methods yield the same result. The discussants argue that all lines through the origin are equally inadequate because the (population) distribution is invariant for rotations of  $(x, y)$  around  $(0, 0)$ . But in the regression setting the vertical direction is special, as can be seen from the use of the conditional distribution of  $y$  given  $x$ , and the fact that residuals are measured vertically.

Perhaps this misunderstanding was due to our statement that  $rdepth$  is affine invariant and therefore DR is affine equivariant. To be precise, affine equivariance (see Rousseeuw and Leroy 1987, page 116) does not refer to linear transformations (e.g. rotations) of  $(\mathbf{x}, y)$  in  $\mathbb{R}^p$ , but to transformations of the type  $(\mathbf{x}, y) \rightarrow (A(\mathbf{x}, 1)', y)$  where  $A$  is a nonsingular matrix. [For regression through the origin, just replace  $(\mathbf{x}, 1)$  by  $\mathbf{x}$  throughout.] The other two equivariances shared by most regression methods correspond to  $(\mathbf{x}, y) \rightarrow (\mathbf{x}, cy)$  where  $c$  is a constant (*scale equivariance*), and  $(\mathbf{x}, y) \rightarrow (\mathbf{x}, y + (\mathbf{x}, 1)\mathbf{v}')$  where  $\mathbf{v}$  is any vector (*regression equivariance*). Using affine equivariance (on  $\mathbf{x}$ ) or scale equivariance (on  $y$ ) we can transform the disk of

Example 2 to a nearly flat ellipse, to which the horizontal fit may seem more natural. For arbitrary dimensions and regression with intercept, any elliptical distribution determines a (hyper)plane given by  $\text{med}[y|\mathbf{x}]$ , which coincides with  $E[y|\mathbf{x}]$  if the latter exists, and the DR method always obtains this hyperplane by Theorem 7(c).

General rotations of  $(\mathbf{x}, y)$  are allowed in orthogonal regression, which corresponds to a model where residuals are measured in the direction orthogonal to the fitted hyperplane. Rotations are also allowed if we wish to describe the round shape of the data cloud by means of contours, e.g. those obtained by location depth.

**Example 3.** In our paper we said that if the bivariate data set  $Z_n$  lies exactly on a straight line then  $\max_{\theta} rdepth(\theta, Z_n) = n$  is the highest possible. If  $Z_n$  lies exactly on a strictly convex curve, then (Theorem 2) the  $\max rdepth$  is at its lowest. Regina and Kesar were not convinced by our statement that  $\max rdepth(Z_n)$  can be seen as a measure of linearity of  $Z_n$  because its lower bound does not depend on the *amount* of curvature, which they illustrate by three different curves in their Figure 4. However, Theorem 2 is restricted to the extreme case where the  $(x_i, y_i)$  lie *exactly* on the curve. As soon as there is noise (i.e. nearly always), the relative sizes of the error scale and the curvature come into play. To illustrate this, we took 50 equispaced  $x_i = (i - 1/2)/50$  in  $[0, 1]$ . For several values of  $\sigma$  we generated  $y_i = f_j(x_i) + e_i$  where  $e_i \sim N(0, \sigma^2)$  and where  $f_1(x) = x$ ,  $f_2(x) = (1/21)(x^2 + 20x)$ ,  $f_3(x) = (1/7)(x^2 + 6x)$ , and  $f_4(x) = x^4$  as in Regina and Kesar's comment. In the special case  $\sigma = 0$  all  $e_i = 0$  so we have 50 points on the straight line  $f_1$  hence  $\max rdepth(Z_n) = n = 50$ , or 50 points on one of the convex functions  $f_2$ ,  $f_3$  or  $f_4$  with for each  $\max rdepth(Z_n) = \lceil (n+2)/3 \rceil = 18$  according to Theorem 2.

Table 2 in this rejoinder shows what happens for other  $\sigma$ . As soon as  $\sigma > 0$ , the 50 points generated around the line  $f_1$  no longer fit exactly, and their  $\max rdepth$  goes down to 23 which is roughly what one would expect since the population distribution then has a linear conditional median, so  $\max rdepth(Z_n) \approx n/2$  by Theorem 1(d). The  $\max rdepth$  of the data sets generated around  $f_2$ ,  $f_3$  and  $f_4$  remain highly significant (i.e. at the 0.1% level) for  $\sigma = .001$  and  $.005$ . But for  $\sigma \geq .01$  the  $\max rdepth$  of  $Z_n \sim f_2$  is no longer significant. This is because  $f_2$  lies close to  $f_1$  in Figure 4 of the comment, so a data set generated from  $f_2$  with  $\sigma \geq .01$  no longer has a visible curvature. Since  $f_3$  is more curved than  $f_2$  this happens later, at  $\sigma \geq .05$ . And for the highly curved  $f_4$  it takes  $\sigma \geq .5$  to wash out the curvature.

The significance levels in Table 2 were obtained by simulation. We took a set of 50 equispaced  $x_i$  and generated  $m = 10,000$  samples  $Z^{(j)} = \{(x_i, e_i); 1 \leq i \leq 50\}$  with standard gaussian  $e_i$ . For each  $j$  we computed  $\max rdepth(Z^{(j)})$ . This yields the  $p$ -values  $p(19) = P(\max rdepth(Z_n) \leq 19) = .000$ ,  $p(20) = .002$ ,  $p(21) = 0.041$ ,  $p(22) = .355$ ,  $p(23) = .927$  and  $p(24) = 1.000$ .

At the end of their comment, Regina and Kesar make a suggestion for using location depth in regression. In the  $p$ -variate case, their proposal is to consider subsets

Table 2. Maximal  $rdepth$  of data sets generated by the linear function  $f_1$  and the convex functions  $f_2$ ,  $f_3$  and  $f_4$  plus a gaussian error term with standard deviation  $\sigma$ . Here (\*) means significant at the 5% level, (\*\*) at the 1% level, and (\*\*\*) at the 0.1% level.

$\sigma$	maximal $rdepth$			
	$f_1$	$f_2$	$f_3$	$f_4$
.000	50	18(***)	18(***)	18(***)
.001	23	18(***)	18(***)	18(***)
.005	23	19(***)	18(***)	18(***)
.010	23	22	19(***)	18(***)
.020	23	23	21(*)	18(***)
.050	23	23	23	18(***)
.100	23	23	23	19(***)
.150	23	23	23	20(**)
.200	23	23	23	21(*)
.500	23	23	23	23

$Z_J = \{(\mathbf{x}_j, y_j); j \in J\}$  for all  $J \subset \{1, \dots, n\}$  with  $|J| = p$ . To each  $Z_J$  corresponds some  $\theta_J$  which fits these points perfectly. The idea is then to compute the deepest point  $\theta^*$  among the  $\binom{n}{p}$  points  $\theta_J$  in  $\mathbb{R}^p$ . This reminds us of a proposal by Oja and Niinimaa (1984) of this type, where the Oja median was used. We discussed this proposal in Rousseeuw and Leroy (1987, pages 146-148) where we pointed out that the resulting regression estimator has a low breakdown value  $\varepsilon_{reg}^*$ . In fact, when  $\varepsilon_{loc}^*$  is the breakdown value of the location estimator, it holds that  $\varepsilon_{reg}^* = 1 - (1 - \varepsilon_{loc}^*)^{1/p}$  which goes down rapidly with  $p$ . For instance, if one uses the Tukey median ( $\varepsilon_{loc}^* = 1/3$ ) then  $\varepsilon_{reg}^* = 7.8\%$  for  $p = 5$  and  $\varepsilon_{reg}^* = 1.7\%$  for  $p = 10$ . Even if one uses a location estimator with the highest possible  $\varepsilon_{loc}^* = 1/2$  (e.g. the MCD location, with depth function given by the MCD-based robust distance), one obtains  $\varepsilon_{reg}^* = 13\%$  for  $p = 5$  and  $\varepsilon_{reg}^* = 7\%$  for  $p = 10$ . For  $p = 2$  and the coordinatewise median location estimator, we recover the slope estimator of Theil (1950) and Sen (1968) with  $\varepsilon_{reg}^* = 1 - (1/2)^{1/2} = 29\%$ .

#### 4. CARROLL, RUPPERT AND STEFANSKI

Professors Carroll, Ruppert and Stefanski appreciate the geometric flavor of our paper, but say they have become skeptical about robust methods in general. This came as a surprise to us, in view of their publication record. We think that science and technology advance by trying out new ideas, comparing them with existing approaches, and gaining more insight along the way. Why should the existence of many methods be considered a disadvantage in robust statistics when it isn't in nonparametric density estimation or time series models? We agree with Roger Koenker's remark, who defends the availability of several regression estimators by the analogy of replacing all food types by a single nutritive tablet.

Let us point out that several robust methods did make it into commercial software packages. The  $L^1$  method and some M-estimators are in many packages, but also positive-breakdown methods have been in S-Plus for some time and are now in the new SAS/IML 7.01. Approximate inference has always been available in robust statistics, either through the asymptotics of the estimator (as in Huber 1981) or through reweighted least squares (as in Rousseeuw and Leroy 1987). Exact inference is harder when the data contain outliers, but we do have small-sample asymptotics for M-estimators (Field and Ronchetti 1990) and the bootstrap (Efron and Tibshirani 1993, chapter 9). The new approach based on regression depth is especially well-suited for finite-sample inference, as we saw in Sections 1 and 3 of this rejoinder.

Since the discussants explicitly ask which methods one of us (PR) uses in practice, the next paragraphs will be in the first person singular. Because of the possibility of  $x$ -outliers, which may be hard to find if  $p \geq 3$ , I use a positive-breakdown method. Both LMS and LTS were proposed in (Rousseeuw 1984), and soon generalized to S-estimators for regression (Rousseeuw and Yohai 1984). In my opinion the LMS is now superseded by the LTS.

The only concrete advantage of the LMS is that it has minimax bias among all residual-based estimators (Martin, Yohai and Zamar 1989) but this does not outweigh the many advantages of the LTS due to its smoother objective function which gives it a much higher efficiency. My default choice of the LTS coverage is now 75%, yielding a 25% breakdown value. The LTS can currently be computed much faster than the others by the FAST-LTS algorithm (Rousseeuw and Van Driessen 1999). The reasons why the LMS is still around are mostly 'traditional': the name has stuck and the principle is easy to explain. The LTS is only slightly harder to explain, but most non-statisticians will see S-estimators and later developments as a 'black box,' which they often resist.

The LMS, LTS and S are 'mode-seeking' methods that search for a concentrated linear cloud with the majority of the data. In many applications this is what we want, but not always. As I pointed out in (Rousseeuw and Leroy 1987, page 241) there may be two intersecting linear structures, and then these methods will hesitate which one to choose. This has been rediscovered and published repeatedly by critics of these methods. In many applications one would rather want a more predictable 'median-type' regression method, that moves more gradually and in a monotone way. Searching for such a method led me to the deepest regression (DR) method proposed in the present paper. In practice, when analyzing regression data I nowadays run both DR and LTS and compare the results. Both methods can easily withstand up to 25% of outliers, which is typically enough to obtain a resistant fit and to *detect* the outliers, whether there are 1% of them, 7%, 12%, or whatever.

For robustly estimating the location and scatter matrix of a point cloud, a similar situation holds. Both the Minimum Volume Ellipsoid estimator (MVE) and the Minimum Covariance Determinant estimator (MCD) were proposed in (Rousseeuw 1984, 1985) and followed by multivariate S-estimators (Rousseeuw and Leroy 1987) and more refined estimators. I currently use the MCD, which can be computed quickly by the FAST-MCD algorithm of Rousseeuw and Van Driessen (1998). The MCD thus has a much higher efficiency than the MVE, both computationally and statistically, and my default coverage is again 75%. The FAST-MCD algorithm also allows to speed up the promising hybrid method of Rocke and Woodruff (1996), in which MCD is an essential component. The MCD is again a mode-type estimator, whereas the deepest location is a median-type estimator with a comparable breakdown value, but currently needing substantially more computation time (Struyf and Rousseeuw 1998).

I disagree with the general sentiment that 'robust statistics lacks success stories'. There are many, but they have not been widely publicized. My problem is that writing review papers and doing new research do not both fit in the available time. But let me at least give some references to substantial applications of positive-breakdown regression methods. The LMS/LTS methods are in use in chemometrics since (Massart et al 1986). In econometrics, earnings functions have been estimated in different sec-

tors and countries (Rousseeuw and Wagner 1994, Hubert and Rousseeuw 1997). The protocol for connecting optical fiber cables for data transmission designed at NIST (Wang et al 1997) is based on LMS. For an application to management science, see Seaver and Triantis (1995). There are also many applications to computer vision, where LMS/LTS have been used for image recovery (Meer et al 1991, Koivunen 1995), surface reconstruction (Sinha and Schunck 1992), identifying shape from color images (Drew 1994, 1996), robot positioning (Kumar and Hanson 1994), extracting geometric primitives (Roth and Levine 1993), range data (Stewart 1995), and to detect moving objects in video from a mobile camera (Abdel-Mottaleb et al 1993, Thompson et al 1993), and constructive fitting (Veelaert 1997). For some applications in artificial intelligence based process control see Karr et al (1995).

The MVE and MCD estimators for a robust location and scatter matrix have been used for detecting  $x$ -outliers in regression by Rousseeuw and van Zomeren (1990), who proposed a diagnostic display of robust residuals (say, obtained by LTS) versus robust distances (say, obtained by MCD). In a geochemistry application (Chork 1990), MVE-based robust distances were used to detect mineralizations hidden beneath the surface. The MVE/MCD can also be used to robustify multivariate techniques like principal components or discriminant analysis (Chork and Rousseeuw 1992, Hawkins and McLachlan 1997). Applications in computer vision include image segmentation (Jolion et al 1993). In astronomy, Plets and Vynckier (1999) used the MVE to detect a special type of stars, the so-called ‘Vega phenomenon’. A major application is to real-time estimation of the state variables to control an electric power network (Mili et al 1991, 1994, 1996) which is currently in use by electric power utilities in the US and Switzerland.

Carroll, Ruppert and Stefanski say that the examples in the literature do not show that robustness gains you anything over least squares-plus-diagnostics, which they call relatively simple. The simplest and most well-known diagnostics are of the ‘leave-one-out’ type, but surely Carroll et al cannot be referring to them because the statistical literature is full of examples where they suffer from masking (i.e. where they do not detect outliers because there are two or more of them). Since the eighties, the original proponents of leave-one-out have shifted their research and publications toward multiple outlier diagnostics to avoid the masking effect. (Meanwhile, many non-statisticians are not aware of this and continue to apply leave-one-out diagnostics because they are now used to them - these things take time.) When commenting on one of our examples, the discussants mention ‘modern diagnostics’ by which they probably mean multiple outlier diagnostics. It is interesting, and deserves to be stressed here, that many of those modern diagnostics actually *use* positive-breakdown estimators. For instance, Atkinson (1988, 1994) uses LMS and MVE, Cook and Nachtsheim (1994) and Fung (1993) apply the MVE, and Hadi (1992) constructs an approximation of the MVE. Also, algorithms for the LMS, LTS, MVE and

MCD were proposed in Atkinson and Weisberg (1991), Cook, Hawkins and Weisberg (1993), and Hawkins (1994) by people who would traditionally be considered members of the ‘diagnostics school’. In fact, the goals and techniques of modern diagnostics and positive-breakdown estimators are closely related. In the long run, robustness and diagnostics will be considered as the same subject.

It strikes us that Carroll et al feel there is not enough work about inference in robust regression, but don’t ask the same from diagnostic tools. There the idea is to detect the outliers, after which they are removed or the model is changed, and then the LS inference is applied. When we did the very same thing by applying reweighted LS (e.g. in Rousseeuw and Leroy 1987) we were criticized by people who said that our confidence intervals and tests were ‘only approximations’. Similarly, nobody criticizes the diagnostics literature about ‘efficiency losses’, whether few or many observations are removed. Work under the ‘robustness’ and ‘diagnostics’ labels is thus being compared to different criteria.

So, why are these closely related fields still held to different standards? One reason may be the naming, since ‘diagnostics’ evokes ‘diagnosing an infection,’ whereas ‘robustness’ evokes ‘preventing (the ill effects of) an infection’. Another reason may be that pioneering diagnostics work such as Cook (1977, in *Technometrics*) was written from an applied viewpoint, whereas the paper of Huber (1964, in *Annals of Mathematical Statistics*) was much more theoretically inclined. Our viewpoint is in between: we think that detecting deviations in the data works best when comparing them to the fit one would have found without the deviations, and we have a high respect for practical theory.

Data sets can usually be analyzed in different ways, and the discussants point out that our examples were no exceptions. Of course the cost data are curved (see Section 3 above for detecting curvature using *rdepth*). We chose that example to illustrate the monotone equivariance of deepest regression (DR) and to show the DR curve. So we’re not opposed to transformations, we’ve *used* one. Similarly, we often apply Q-Q plots, Box-Cox transformations, variable selection, and a lot of other data analysis techniques that we didn’t mention in this particular paper. It is also possible to choose transformations based on robust methods (Atkinson 1988). For the Skeena River data the depth-based inference in Section 1 above indicated that the intercept is not significant ( $p = 0.78$ ), so we may switch to the smaller model of regression through the origin, thereby complying with the biological truth that zero spawners produce zero recruits. In that model, the depth envelope  $E_k$  is the region between two rays emanating from  $(0, 0)$ , the lower one with slope  $(y_i/x_i)_{k:n}$  and the upper one with slope  $(y_i/x_i)_{n-k+1:n}$ . This envelope  $E_k$  does not contain any negative recruits. In our paper we have used the Skeena data as an example, without any criticism on the techniques of Carroll and Ruppert (1988). We feel misunderstood, because when we write ‘here we do this’ people sometimes read ‘you have to do this, and nothing else’.

The DR method has not yet been applied to many other models because it is so recent: this is the first paper we wrote about it. It is difficult to foresee the ultimate utility of a newborn child (although the initial DNA tests gave promising results). But a start is being made: Zhang (1998) looks at depth-based covariance matrices, Christmann and Rousseeuw (1999) apply rdepth to logistic regression, and Mizera (1998) sketches how the DR could be generalized to nonlinear models.

We agree that data mining offers great opportunities for robust methods. At the moment DR cannot yet be computed that fast, but other robust methods can and have been applied to large data sets. Rousseeuw and Van Driessen (1998) ran the MCD on an astronomy data set with  $n = 132,000$  celestial objects and  $p = 28$  variables, which took 18 minutes on our SUN Ultra 2170 machine. Our analysis led the astronomers to modify their classification of objects as stars and galaxies. We intend to perform robust regression on data sets of similar size as well.

## 5. MCKEAN AND SHEATHER

This discussion raises a lot of points for us to reply to. It begins by saying that, apart from the deepest regression (DR), there are already other regression methods which are not affected by skewness and heteroskedasticity. But when estimating the intercept in our semiparametric model, there seem to be no alternatives yet to the  $L^1$  and the DR to do it consistently. For R-estimators one can choose the right scores only if one *knows* the form of the error distribution in advance (and similarly for the  $\psi$ -function of an M-estimator). For the slope it is possible to obtain consistent M-estimators, but this is not true for GM-estimators other than the Mallows type, as shown by Carroll and Welsh (1988). And while it is true that heteroskedasticity can be modelled from the data, this is not the same as having an estimator like DR based on a model which already *includes* heteroskedasticity. Also note that

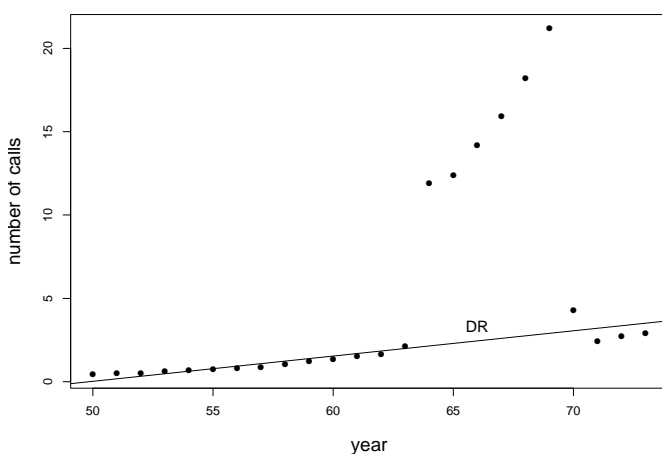


Figure 3. The Belgian phone calls data with its DR line.

our model allows for a different type of error distribution with arbitrary dispersion at any  $x$ , a tough combination.

As we pointed out in our paper, there may be several  $\theta$  with the maximal rdepth. Therefore we defined the DR as the average of those  $\theta$ , to make it unique and to increase the finite-sample efficiency (by a few percent). We didn't take their median, to avoid the extra computation of a multivariate median. The average solution is fine, because the breakdown value holds uniformly for all solutions. Our definition is similar to that of Donoho and Gasko (1992) for the deepest location.

The discussants have not used our definition, but instead modified our code to give all the lines with maximal depth, and then complain about the variability between them. For this they have used two carefully selected examples. Their Figure 2 was one of 1,000 samples  $\{(x_i, y_i); i = 1, \dots, 30\}$  with  $x_i \sim (3/4)N(0, 1) + (1/4)N(0, 16)$  and  $y_i \sim N(0, 1)$ . We have repeated this simulation and found very few samples where the variability was as large as in Figure 2, so the discussants have chosen their example well. They mention 'chasing both the low and high outliers,' but this is an interpretation not supported by the data, where the  $x_i$  are between  $-4$  and  $+4$  (quite uncommon for this simulation setup) and the  $y_i$  are  $N(0, 1)$ . In our simulation the variability between the solutions was typically much lower. Going further, we repeated the simulation for sample size  $n = 50, n = 100, n = 200, \dots$  and found that the variability between the slopes goes down a lot faster than  $1/n$ . Therefore this is a typical 'small sample granularity effect'. In our definition of the DR and its implementation, this effect is already averaged out.

The discussants' Figure 1 is not the 'Belgian calls data set' from Rousseeuw and Leroy (1987, p. 25) but a particular subset of it, where they removed the first five years 1950,  $\dots$ , 1954 in which there were no outliers. Our Figure 3 shows the complete data set with its DR fit (the variability between solutions was negligible).

When we redid the analysis by deleting the year 1950, and then also 1951,  $\dots$  we found that the DR remained in the right place all the time, until we got to the subset plotted by the discussants! Since we proved in our paper (Theorem 4) that the breakdown value of the DR line is  $1/3$ , and the Belgian phone calls data has  $n = 24$  observations with 6 of them being outliers, it is easy to see that one can make the DR break down by removing 5 good observations. So, the discussants selected the largest subset in which DR gives the wrong answer. Their Figure 1 has two linear modes, with the DR in between.

Their Figure 3 is based on the same principle. For a regression line model through the origin, the DR has breakdown value  $1/2$ . It was proved in Rousseeuw (1984, page 879) that no regression equivariant estimator can have a higher breakdown value. The artificial data in the discussants' Figure 3 consists of two groups of 10 points each, the first group having higher  $y_i/x_i$  than the second. The discussants keep calling the first group 'bad' and the second group 'good', but who is to tell? Based on the data, this is not decidable: One could also have started with 20 points like the first group and have replaced 10 of them to form

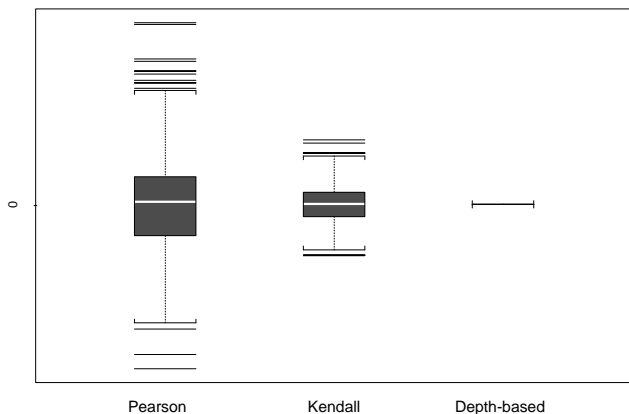


Figure 4. Boxplots of  $\text{corr}(x, r)$  using Pearson's  $\rho$ , Kendall's  $\tau$ , and the depth-based  $\eta$ , for the DR residuals in 1,000 simulated samples.

the second group. There are no 'good' and 'bad' points here, so the discussion about which method best identifies the 'good' points is without meaning. (Both residual plots show that there are two groups.) The discussants may rally support for their preferred line because it is close to where the LS line would be, and many people are so used to LS that this corresponds to their acquired intuition. But in reality there are two groups, and by taking the slopes  $t_i = y_i/x_i$  the situation is equivalent to trying to estimate the 'central location' of perfectly bimodal univariate data.

In their Figure 6, the discussants show boxplots of the Pearson correlation  $\rho(\hat{y}, r)$  where  $\hat{y}_i$  and  $r_i$  are obtained by the DR line in each of the 1,000 simulations of their model with long-tailed  $x_i$  and gaussian  $y_i$ . The boxplot is biased towards negative correlations, which can be explained as follows. When the DR slope  $\hat{\theta}_1$  is positive, it often happens that  $\text{corr}(x, r) < 0$  because the LS slope tends to be closer to zero due to some  $x_i$ -outliers with normal  $y_i$ . When  $\hat{\theta}_1 < 0$  we often have the opposite,  $\text{corr}(x, r) > 0$ , for the same reason. But the discussants plot  $\text{corr}(\hat{y}, r)$  instead of  $\text{corr}(x, r)$ , and  $\text{corr}(x, \hat{y}) = \text{sgn}(\hat{\theta}_1) = -1$ , so also in those cases  $\text{corr}(\hat{y}, r) < 0$ . To avoid this bias, we plot  $\rho = \text{corr}(x, r)$  in our Figure 4.

In Figure 4 we see that the boxplot of the Pearson correlation  $\rho$  is now symmetric. As the correlations of Spearman and Kendall are closely related by construction we plotted only Kendall's  $\tau$ , which has a lower variability because it is more robust. Since the LS fit is defined by requiring that  $\rho(x, r) = 0$  [in multiple LS regression this must hold for each coordinate of  $\mathbf{x}$ , and for all linear combinations of these coordinates], the Pearson correlation will look best at the LS fit. Kendall's  $\tau$  will be identically zero at the fit of Theil (1950) and Sen (1968), not shown here. We would like to have a correlation coefficient corresponding to the DR fit. A simple idea is to put

$$\eta(x, y) = \text{sgn}(\hat{\theta}_1) \min\{1, |\hat{\theta}_1| \text{mad}(x) / \text{mad}(y)\} \quad (9)$$

where  $\hat{\theta}_1$  is the DR slope and  $\text{mad}(x) = \text{med}_{i=1}^n |x_i - \text{med}_{j=1}^n(x_j)|$ .

If  $\hat{\theta}_1 = 0$  (e.g. when the linear conditional median is horizontal) we find  $\eta(x, y) = 0$ . If all the  $(x_i, y_i)$  fall on a line with strictly positive (resp. negative) slope, we find  $\eta(x, y) = 1$  (resp.  $-1$ ). By construction, the DR fit always satisfies  $\eta(x, r) = 0$  as in Figure 4.

Figure 7 in the discussion shows one of the 1,000 samples of their simulation. It is stated that the DR fit is poor because the true slope is zero. However, if one generates 1,000 samples from a model with slope zero, the question to be asked is whether the average of the DR slopes is zero (which is the case), and what is their variance. By selecting one sample out of 1,000 by an additional criterion (e.g. taking the one with largest  $|\hat{\theta}_1|$ ), this sample is no longer representative for the model. If the sample size  $n$  is small, and the number  $m$  of simulated samples is large, you can always select a sample that could equally well have been generated by a rather different model. To us, the residual plot of DR and Wilcoxon look equally non-random. The suggestion that one could fit a decreasing line to the DR plot is due to LS intuition: the LS would fit such a line because of the leverage points in the lower right, but not without them. By regression equivariance, the DR fit to this plot would be the horizontal line. The two vertical outliers being mentioned are observations with  $y_i$  between 1.5 and 2.0, so they are boundary cases.

Concerning the discussants' preference for a zero-breakdown regression method like a plain M- or R-estimator (here, the one with Wilcoxon scores), let us point out that we don't need to generate many samples to find one where the result looks counterintuitive. It suffices to take any data set in any dimension and to add one bad leverage point to see the Wilcoxon method break down. In our view this amply outweighs the high efficiency of this R-estimator in the absence of outliers. We can achieve the same efficiency by computing the DR followed by a one-step M-estimator, without losing the breakdown robustness of the DR.

For the curved examples, the discussants have used the model  $y_i = 5.5|x_i| - 0.6|x_i|^2 + e_i$  with the same  $x_i$  as before. This is a particularly confusing choice, because most of the  $|x_i|$  are near zero where the parabola is nearly linear, and the few large  $|x_i|$  look like they could be outliers, e.g. in the scatterplot of situation C. This is an age-old philosophical question, as it was e.g. described by Huber (1981, page 154): should we declare those two points as outliers or fit a quadratic model? We are told we know the latter is the right choice because of the way the simulations were generated, but this does not prove much. One could just as well generate 1,000 samples of a linear model with the same  $x$ -distribution and some  $y$ -outliers, and find situation C among them. Based on the data alone, the model choice questions cannot be settled with any reliability, so in practice one would like to combine the statistical analysis with subject-matter knowledge.

Situations B and C are similar except that the  $x_i$  are more regularly distributed in B, where the scatterplot gives more evidence of a quadratic model. The discussants ref-

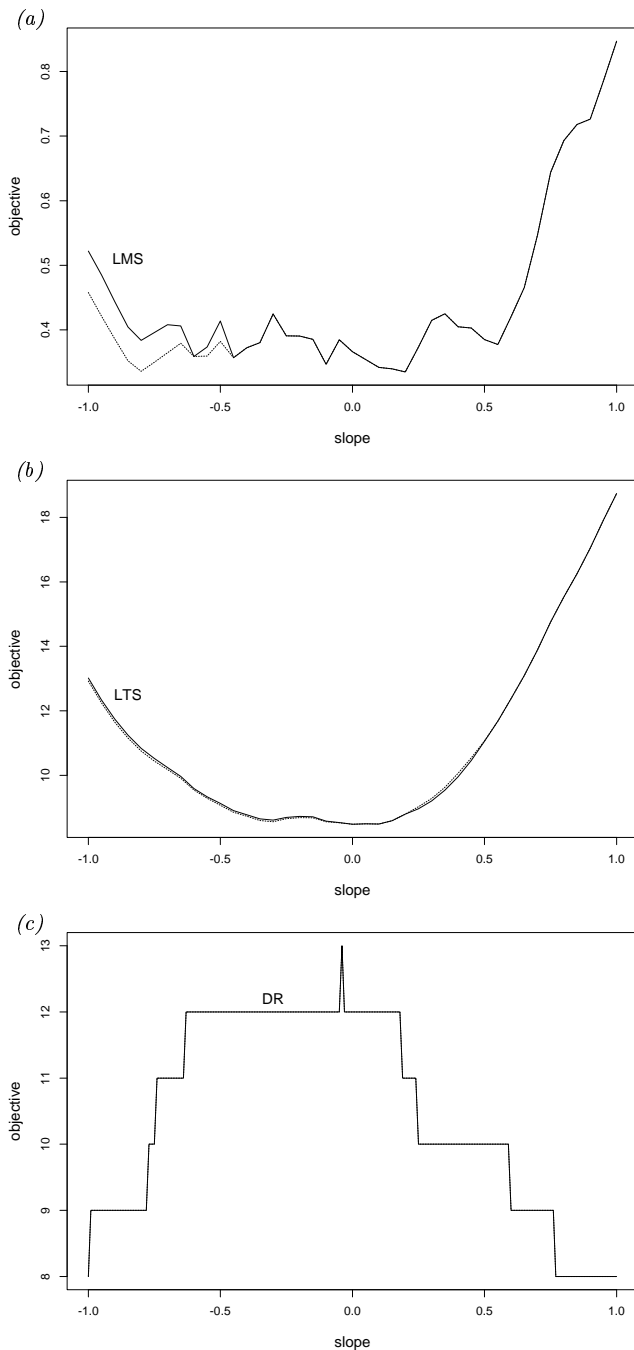


Figure 5. Data of Figure 16: objective functions of (a) the LMS, (b) the LTS, and (c) the DR. The solid curves are for the original data, and the dashed curves for the modified data.

erence the papers of Cook, Hawkins and Weisberg (1992) and McKean, Sheather and Hettmansperger (1990, 1993) where it is said that residual plots of zero-breakdown estimators of M- and R-type can be interpreted in much the same way as LS plots (perhaps because the influence functions of the latter resemble the LS normal equations). These authors have also claimed that curvature is harder to detect in residuals from positive-breakdown estimators (and in small samples!), although we never saw the difference, and neither did Davies (1994, pages 365-366) who also felt that the LMS residual plot in Cook et al (1992) did reveal curvature, and went on to analyze it.

Looking at the discussion's residual plots of DR residuals and Wilcoxon residuals (Figures 11-12) we still see the curvature just as easily in both. We do not see why one of them would support a quadratic model while the other wouldn't. If we carry out an F-test for adding a quadratic term, the test statistics will be *identical* for both because of all the invariance properties (regression invariance, scale invariance, affine invariance) which allows one to go from one plot to the other. The same holds if we would look for curvature in these plots by means of the test based on the maxdepth as in Table 2 of Section 3 of this rejoinder, since rdepth has the same invariance properties. So mathematically, there *is* no difference!

Fortunately, we finally figured out what has been bothering all these people. If you are used to LS plots, you expect *both* arms of a concave curve to stick out at the bottom of the plot, because the linear LS fit always makes an 'average compromise'. Not wearing these LS-colored glasses, we see the curvature just as well when only *one* arm of the curve sticks out, as in Figure 11 of the discussion. It never occurred to us that the much-heralded problem of detecting curvature in positive-breakdown fits could be something that simple.

Using the example in their Figure 16, the discussants then again raise the issue of local instability of the LMS, as they did already in Hettmansperger and Sheather (1992) and Sheather, McKean and Hettmansperger (1997). We knew about this effect much earlier (Rousseeuw and Leroy 1987, page 241). As we explained in Section 4 above, this is one of the main reasons why we have replaced the LMS by the LTS, with the default coverage of 75%. The LTS has a much smoother objective function and is far less prone to local instability. The DR is even better at this, because it is a 'median-like' regression method whereas LTS is more 'mode-like'. Since the discussants had applied our DR algorithm to all other examples in their contribution, it seemed remarkable that in Figure 16 they had not. We asked the data of their example and applied DR to it, and found *exactly* the same DR fit in the data before and after moving the circled point.

In order to explain these differences, our Figure 5 shows the objective functions of the LMS, LTS, and DR as a function of the slope  $\theta_1$ . (For each  $\theta_1$  we determined the intercept  $\theta_2$  yielding the best objective value.) Figure 5a shows the LMS objective of the data in their Figure 16, indicated by the solid curve for the original data and by the dashed curve for the data with modified point. These curves are not smooth, and their minimum is attained at different slopes. By comparison, the objective function of LTS in Figure 5b is very smooth, and the solid and dashed curves lie very close to each other and yield essentially the same minimum. The objective in Figure 5c must be interpreted differently, because rdepth takes only integer values and we want to *maximize* it. Here the solid and dashed curves coincide exactly, so the maximum is attained at the same slope. This is not the algorithm we use in practice for finding the DR fit, but it does give an idea as to why it should be possible to compute the DR rather quickly.

In their conclusions, McKean and Sheather mention the approach of Simpson, Ruppert and Carroll (1992) which starts by LTS and follows it by one step of a GM-estimator. The discussants are now publishing their own brand of this approach in (Chang et al. 1999), which sits uneasy with their simultaneous rejection of our high-breakdown work of the past and present. For analyzing data, Rousseeuw (1984, page 875) suggested to “run both an LMS and LS regression. If they agree closely, the LS result can be trusted. If, on the other hand, there is a significant difference, then we know which observations are responsible by looking at the LMS residuals.” We are pleased that the discussants close by making the same recommendation (updated to their preferred high-breakdown and zero-breakdown fits). The novelty is that they have constructed a diagnostic to decide when the difference between both fits is significant.

## 6. OLIVE AND HAWKINS

We are grateful to Professors Olive and Hawkins for their positive comments on our work. A large part of their discussion centers around the computation time of the deepest regression (DR) method. Since this concern was shared by most discussants, we responded to it in the beginning of our rejoinder. We of course agree that exact algorithms with a computational price tag of  $O(n^p)$  or more need to be complemented by faster approximate algorithms.

As an alternative to the  $L^1$  and the DR, David and Doug then propose the Least Adaptively Trimmed sum of Absolute deviations (LATA). It is an adaptive version of the Least Trimmed Absolute Deviations (LTAD) method,

first proposed by Bassett (1991) as an  $L^1$ -like version of the LTS. Here the adaptivity is in the number of terms  $U_n$  which in turn depends on the real constant  $k \geq 1$ . The LATA( $k$ ) method has 50% breakdown value for any  $k$ , whereas it tends to the  $L^1$  for increasing  $k$ . This reminds us of a nameless proposal we made in (Rousseeuw and Leroy 1987, formula (6.11) on page 153), which consisted of minimizing the objective

$$\min\left\{\frac{1}{n} \sum_{i=1}^n r_i^2(\boldsymbol{\theta}), (k \sum_{i=1}^n |r_i|)^2\right\}. \quad (10)$$

The constant  $k$  in (10) has a similar interpretation: a low  $k$  gives the LMS, any finite  $k$  gives a breakdown value of 50%, and increasing  $k$  yields LS. This paradoxical behaviour is also encountered with  $k$ -step M-estimators starting from an initial estimator with 50% breakdown value. The explanation is the same in each case: increasing  $k$  keeps the same breakdown value but increases the maxbias curve, which goes to infinity with  $k$  (see Rousseeuw and Croux 1994). Since we are concerned with the maximal bias that can occur for a given fraction of outliers, we should keep  $k$  low in all these methods.

David and Doug also propose to define LATA-based regression quantiles by computing the  $L^1$ -based quantiles based on the  $U_n$  cases that were together responsible for minimizing (1.1). We are a bit uneasy with this suggestion because this set of cases was obtained by taking a symmetric view on residuals, whereas Koenker and Bassett (1978) used the asymmetric  $\rho_\tau(r_i) = \tau|r_i|1(r_i \geq 0) + (1-\tau)|r_i|1(r_i \leq 0)$  instead of  $|r_i|$  throughout. Therefore, we wonder about the consistency of LATA-quantiles for the true conditional  $\tau$ -quantile of  $y$  given  $x$ . Perhaps one needs to compute a different set of cases for each  $\tau$ .

## 7. PORTNOY AND MIZERA

We are glad to hear that Professors Portnoy and Mizera found our paper exciting. Although we apologize for the sleepless nights, we have to admit to similar symptoms when developing this topic. We agree that a lot of intriguing questions remain about the connections and differences between deepest regression (DR) and  $L^1$  regression.

Stephen and Ivan provide a plausible heuristic reasoning indicating that the asymptotic efficiency of DR goes down with  $p$ . We would be grateful if they could provide a technique to compute the covariance matrix  $B$ , not only to settle this conjecture but also for the benefit of inference about the parameter vector  $\boldsymbol{\theta}$ . In two dimensions we have seen (in Section 1 of this rejoinder) that we can construct confidence intervals and tests by simulating the distribution of rdepth or by bootstrapped DR estimates, but especially the latter approach becomes costly for higher dimensions.

Stephen and Ivan note that the nonfits that form the basis of our definition of regression depth can also be interpreted as ‘inadmissible fits in a data-analytic sense,’ although the latter use the *magnitude* of residuals, whereas we used only the *signs* of residuals in our paper. The work of Mizera (1998) contains a reference to (Carrizosa 1996)

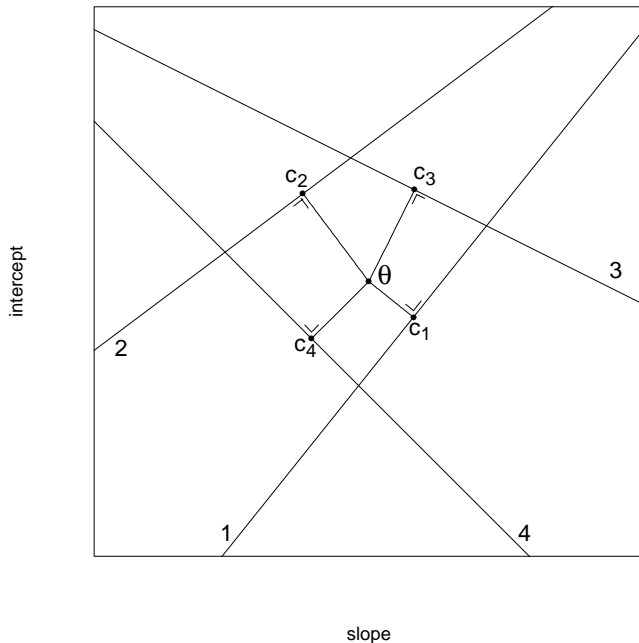


Figure 6. Dual plot of the regression fit  $\boldsymbol{\theta}$  and 4 observations. Each  $c_i$  is the orthogonal projection of  $\boldsymbol{\theta}$  on line  $i$ .

which is mainly about a distance-based characterization of halfspace depth, but at the end makes a suggestion about depth in regression. The latter uses the magnitude of the residuals in a similar way, which yields the same notion although at first sight it looks very different from our definition.

We applaud Stephen's and Ivan's notion of *tangent depth*, which facilitates the extension of depth to nonlinear situations. In retrospect, tangent depth is related to a display that we used to connect regression depth to location depth. The dual plot (Section 5 of our paper) is one way to see the parameter space, in which the original observations  $(\mathbf{x}_i, y_i)$  appear as hyperplanes. For simple regression, look at Figure 6 here with a candidate fit  $\boldsymbol{\theta} = (\theta_1, \theta_2)$  and four observations represented as lines. To compute  $rdepth$  we can project  $\boldsymbol{\theta}$  on each line, yielding  $C = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4\}$ . Note that  $\mathbf{c}_i - \boldsymbol{\theta}$  has the same direction as the gradient of Stephen and Ivan, given by  $(-x_i, -1)r_i$  in our notation, because line  $i$  has the equation  $(-x_i, -1)(\theta_1, \theta_2) = -y_i$  and hence is perpendicular to the gradient. Interestingly,  $rdepth(\boldsymbol{\theta}, Z_n) = ldepth(\boldsymbol{\theta}, C_n)$ . (In Figure 6 we have  $ldepth(\boldsymbol{\theta}, C_n) = 1$ .) The problem in using this to prove Conjecture 1 is that  $C_n$  changes whenever  $\boldsymbol{\theta}$  changes, so that a fixed-point theorem is needed.

Ivan Mizera's (1998) proof of Conjecture 1 is a masterpiece, which uses advanced techniques of mathematical analysis such as set-valued vector fields. He has proved both part (a) for arbitrary finite samples, as well as part (b) for population distributions. Spurred by our account of purely geometrical aspects of regression depth (Rousseeuw and Hubert 1998), researchers in combinatorial geometry have independently proved Conjecture 1' for finite samples (Amenta, Bern, Eppstein and Teng 1998). They used a different approach, based on  $p$ -dimensional projective space, and also provided a partial result to Conjecture 2 which they proved for  $k = \lceil n/(p(p+1)) \rceil$ . Amenta et al (1998) define the *crossing distance* between a point  $x$  and a hyperplane  $L$  as the smallest number of points that  $L$  must cross (in a continuous motion) in order to get to  $x$ . Then the location depth of a point  $\theta$  is its crossing distance to the hyperplane at infinity, and the regression depth of  $L$  is its crossing distance to the point at vertical infinity! This reveals yet another relation between  $ldepth$  and  $rdepth$ .

Looking at our geometrical results in Theorem 1' and Theorem 9, and their extensions by Mizera (1998) and Amenta et al (1998), we think this is one of the rare occasions where statistical work has contributed to pure mathematics, in this case geometry and combinatorics. It is even possible that the latter results will in turn allow to obtain new results in algebra or number theory, because of the many interrelations.

We will not quibble with Stephen and Ivan about the naturalness or otherwise of monotone equivariance. Everybody is entitled to their own opinion on this. We do want to note that after such a transformation the errors remain independent and that their medians remain zero, in accordance with our model. The errors do not need to be identically distributed in our approach, so (unlike most

methods) the DR does not *need* to compensate for heteroskedasticity.

Concerning computation time, we refer to the beginning of this rejoinder. Our goal is certainly to construct algorithms for DR that are consistent for the conditional median when  $H \in \mathcal{H}$ , assuming of course that their computation time is allowed to grow with the sample size.

#### ADDITIONAL REFERENCES

- Abdel-Mottaleb, M., Chellappa, R., and Rosenfeld, A. (1993), "Binocular Motion Stereo with Independently Moving Objects," CAR-TR-679, Center for Automation Research, University of Maryland.
- Atkinson, A.C. (1988), "Transformations Unmasked," *Technometrics*, 30, 311-318.
- Atkinson, A.C. (1991), "Simulated Annealing for the Detection of Multiple Outliers using Least Squares and Least Median of Squares Fitting," in *Directions in Robust Statistics and Diagnostics, Part I*, W. Stahel and S. Weisberg, eds., New York: Springer-Verlag, 7-20.
- Atkinson, A.C. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, 89, 1329-1339.
- Bassett, G.W. Jr. (1991), "Equivariant, Monotonic, 50% Breakdown Estimators," *The American Statistician*, 45, 135-137.
- Carroll, R.J., and Welsh, A.H. (1988), "A Note on Asymmetry and Robustness in Linear Regression," *The American Statistician*, 42, 285-287.
- Chork, C.Y. (1990), "Unmasking Multivariate Anomalous Observations in Exploration Geochemical Data from Sheeted-Vein Tin Mineralization near Emmaville, N.S.W., Australia," *Journal of Geochemical Exploration*, 37, 205-223.
- Chork, C.Y., and Rousseeuw, P.J. (1992), "Integrating a High-Breakdown Option into Discriminant Analysis in Exploration Geochemistry," *Journal of Geochemical Exploration*, 43, 191-203.
- Christmann, A., and Rousseeuw, P.J. (1999), "Regression Depth and Logistic Regression," in preparation.
- Cook, R.D. (1977), "Detection of Influential Observation in Linear Regression," *Technometrics*, 19, 15-18.
- Cook, R.D., Hawkins, D.M., and Weisberg, S. (1993), "Exact Iterative Computation of the Robust Minimum Volume Ellipsoid Estimator," *Statistics and Probability Letters*, 16, 213-218.
- Cook, R.D., and Nachtshiem, C.J. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *Journal of the American Statistical Association*, 89, 592-599.
- Davies, P.L. (1994), "Desirable Properties, Breakdown and Efficiency in the Linear Regression Model," *Statistics and Probability Letters*, 19, 361-370.
- Drew, M.S. (1994), "Robust Specularity Detection from a Single Multi-Illuminant Color Image," *Computer Vision, Graphics and Image Processing: Image Understanding*, 59, 320-327.
- Drew, M.S. (1996), "Direct Solution of the Orientation-from-Color Problem Using a Modification of Pentland's Light Source Direction Estimator," *Computer Vision and Image Understanding*, 64, 286-299.
- Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.
- Field, C., and Ronchetti, E. (1990), *Small Sample Asymptotics*, Hayward, California: IMS Lecture Notes - Monograph Series No. 13.
- Fung, W.-K. (1993), "Unmasking Outliers and Leverage Points: A Confirmation," *Journal of the American Statistical Association*, 88, 515-519.
- Hadi, A.S. (1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society Series B*, 54, 761-771.
- Hawkins, D.M. (1994), "The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data," *Computational Statistics and Data Analysis*, 17, 197-210.
- Hawkins, D.M., and McLachlan, G.J. (1997), "High-Breakdown Linear Discriminant Analysis," *Journal of the American Statistical Association*, 92, 136-143.

- Huber, P.J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73-101.
- Huber, P.J. (1981), *Robust Statistics*, New York: John Wiley.
- Hubert, M., and Rousseeuw, P.J. (1997), "Robust Regression with both Continuous and Binary Regressors," *Journal of Statistical Planning and Inference*, 57, 153-163.
- Johnson, N.L., and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions*, New York: John Wiley.
- Jolion, J.-M., Meer, P., and Bataouche, S. (1991), "Robust Clustering with Applications in Computer Vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13, 791-802.
- Karr, C.L., Weck, B., Massart, D.L., and Vankeerberghen, P. (1995), "Least Median Squares Curve Fitting using a Genetic Algorithm," *Engineering Applications of Artificial Intelligence*, 8, 177-189.
- Koivunen, V. (1995), "A Robust Nonlinear Filter for Image Restoration," *IEEE Transactions on Image Processing*, 4, 569-578.
- Kumar, R., and Hanson, A.R. (1994), "Robust Methods for Estimating Pose and Sensitivity Analysis," *Computer Vision, Graphics and Image Processing: Image Understanding*, 60, 313-342.
- Martin, R.D., Yohai, V.J., and Zamar, R.H. (1989), "Min-max Bias Robust Regression," *The Annals of Statistics*, 17, 1608-1630.
- Massart, D.L., Kaufman, L., Rousseeuw, P.J., and Leroy, A. (1986), "Least Median of Squares: A Robust Method for Outlier and Model Error Detection in Regression and Calibration," *Analytica Chimica Acta*, 187, 171-179.
- Meer, P., Mintz, D., Rosenfeld, A., and Kim, D.Y. (1991), "Robust Regression Methods in Computer Vision: A Review," *International Journal of Computer Vision*, 6, 59-70.
- Mili, L., Phaniraj, V., and Rousseeuw, P. (1991), "Least Median of Squares Estimation in Power Systems" (with discussion), *IEEE Transactions on Power Systems*, 6, 511-523.
- Mili, L., Cheniae, M.G., and Rousseeuw, P.J. (1994), "Robust State Estimation of Electric Power Systems," *IEEE Transactions on Circuits and Systems - I: Fundamental Theory and Applications*, 41, 349-358.
- Mili, L., Cheniae, M.G., Vichare, N.S., and Rousseeuw, P.J. (1996), "Robust State Estimation Based on Projection Statistics," *IEEE Transactions on Power Systems*, 11, 1118-1127.
- Oja, H., and Niinimaa, A. (1984), "On Robust Estimation of Regression Coefficients," Technical Report, Department of Applied Mathematics and Statistics, University of Oulu, Finland.
- Plets, H., and Vynckier, C. (1999), "An Analysis of the Incidence of the Vega Phenomenon among Main-Sequence and Post-Main-Sequence Stars," *Astronomy and Astrophysics*, to appear.
- Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, 91, 1047-1061.
- Rousseeuw, P.J. (1985), "Multivariate Estimation with High Breakdown Point," in *Mathematical Statistics and Applications*, edited by W. Grossmann, G. Pflug, I. Vincze and W. Wertz. Dordrecht: Reidel Publishing Company, 283-297.
- Rousseeuw, P.J. (1999), "A Characterization of Angular Symmetry and of Conditional Median Linearity," in preparation.
- Rousseeuw, P.J., and Croux, C. (1994), "The Bias of k-Step M-Estimators," *Statistics and Probability Letters*, 20, 411-420.
- Rousseeuw, P.J., and Van Driessen, K. (1998), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," Technical Report, University of Antwerp, revised version, December 1998.
- Rousseeuw, P.J., and Van Driessen, K. (1999), "Computing LTS Regression for Large Data Sets," Technical Report, University of Antwerp.
- Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633-651.
- Rousseeuw, P.J., and Wagner, J. (1994), "Robust Regression with a Distributed Intercept using Least Median of Squares," *Computational Statistics and Data Analysis*, 17, 65-76.
- Rousseeuw, P.J., and Yohai, V.J. (1984), "Robust Regression by means of S-estimators". In *Robust and Nonlinear Time Series Analysis*, edited by J. Franke, W. Härdle and D. Martin, Lecture Notes in Statistics No. 26, New York: Springer, 256-272.
- Roth, G., and Levine, M.D. (1993), "Extracting Geometric Primitives," *Computer Vision, Graphics and Image Processing: Image Understanding*, 58, 1-22.
- Seaver, B.L., and Triantis, K.P. (1995), "The Impact of Outliers and Leverage Points for Technical Efficiency Measurement using High Breakdown Procedures," *Management Science*, 41, 937-956.
- Sen, P.K. (1968), "Estimates of the Regression Coefficient based on Kendall's Tau," *Journal of the American Statistical Association*, 63, 1379-1389.
- Simpson, D.G., Ruppert, D., and Carroll, R.J. (1992), "On one-step GM-estimates and Stability of Inferences in Linear Regression," *Journal of the American Statistical Association*, 87, 439-450.
- Sinha, S.S., and Schunck, B.G. (1992), "A Two-Stage Algorithm for Discontinuity-Preserving Surface Reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, 36-55.
- Stewart, C.V. (1995), "MINPRAN: A New Robust Estimator for Computer Vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17, 925-938.
- Struyf, A., and Rousseeuw, P.J. (1998), "High-dimensional Computation of the Deepest Location," Technical Report, University of Antwerp, <http://win-www.uia.ac.be/u/statis/>.
- Theil, H. (1950), "A Rank-Invariant Method of Linear and Polynomial Regression Analysis (Parts 1-3)," *Nederlandsche Akademie der Wetenschappen Proceedings Serie A*, 53, 386-392, 521-525, 1397-1412.
- Thompson, W.B., Lechleider, P., and Stuck, E.R. (1993), "Detecting Moving Objects Using the Rigidity Constraint," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 162-166.
- Van Aelst, S., and Rousseeuw, P.J. (1998), "Robustness of Deepest Regression," Technical Report, University of Antwerp.
- Van Kreveld, M., Mitchell, J.S., Rousseeuw, P.J., Sharir, M., Snoeyink, J., and Speckmann, B. (1999), "Efficient Algorithms for Maximum Regression Depth," Technical Report, University of British Columbia. Submitted.
- Veelaert, P. (1997), "Constructive Fitting and Extraction of Geometric Primitives," *Graphical Models and Image Processing*, 59, 133-251.
- Wang, C.M., Vecchia, D.F., Young, M., and Brilliant, N.A. (1997), "Robust Regression Applied to Optical-Fiber Dimensional Quality Control," *Technometrics*, 39, 25-33.
- Zhang, J. (1998), "Some Extensions of Tukey's Depth Function," Technical Report, Institute of Statistics, Universite Catholique de Louvain, Belgium.