

AD-A049 701

PRINCETON UNIV N J DEPT OF STATISTICS
FURTHER PROGRESS ON ROBUST/RESISTANT WIDTHERS.(U)

F/G 12/1

UNCLASSIFIED

AUG 77 J W TUKEY, H I BRAUN, M SCHWARZSCHILD
TR-129-SER-2

DAA629-76-G-0298
ARO-14244.1-M
NL

| OF |
AD
A049701



END

DATE
FILMED

3-78

DDC

ARO 14244.1-M

**READ INSTRUCTIONS
BEFORE COMPLETING FORM**

AD A049701

Approved for public release; distribution unlimited.

DDC
RECEIVED
FEB 8 1978
F

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

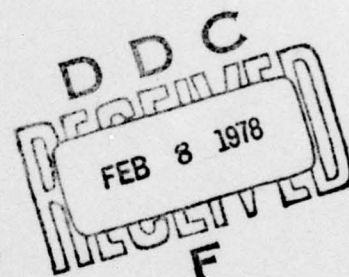
Estimators	Efficiency
Width	Computation
Monte Carlo technique	

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 68 IS OBSOLETE fact apparent triefficiencies of 89% were obtained
406873 *Ince* Unclassified for sample size 10. ☒
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

FURTHER PROGRESS ON ROBUST/RESISTANT WIDTHERS

by

John W. Tukey
Henry I. Braun
Michael Schwarzschild



Technical Report No. 129, Series 2
Department of Statistics
Princeton University
August 1977

Research sponsored by a contract with U. S. Army Research
Office No. DAAG29-76-G-0298, awarded to the Department of
Statistics, Princeton University.

FURTHER PROGRESS ON ROBUST/RESISTANT WIDTHERS

1. Introduction.

This report describes work carried out during the 1975-77 academic years by H. Braun, M. Schwarzschild and J. W. Tukey on the development of robust estimators of width. An earlier technical report, "An interim report of a Monte Carlo study of robust estimators of width," by David A. Lax, describes previous work undertaken in connection with this project. The interested reader should consult Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W., 1972. Robust Estimates of Location: Survey and Advances, Princeton University Press, Princeton, N. J., for references and introductory material, as well as the results of the early Monte Carlo investigations.

Essentially a single family of estimators (together with a few obvious variations) was investigated for sample sizes $n = 10$ and $n = 20$, using Monte Carlo methods. The authors were motivated by considerations of conceptual and computational simplicity as well as a desire to achieve high robustness of efficiency. Moderate success was achieved on both counts. In fact efficiencies apparent, as explained below (see footnotes to page 4 and Table 1), of 89% were obtained for sample size 10.

As in previous work, efficiencies of estimators were calculated for three different underlying distributions. At sample size 20, these are Gaussian (0, 1) Slash (a Gaussian divided by an independent uniform (0, 1) variate), and Wild (19 Gaussian (0, 1) 1 Gaussian (0, 100)). The investigation for sample size 10 was begun with the first two being the same as before but the third being 9 Gaussian (0, 1) and 1 Gaussian (0, 9), for which samples used in the Princeton

Robustness Study were available. Two arguments have been suggested for possibly preferring this to our standard one wild, which would here by:

9 from Gau (0, 1) and 1 from Gau (0, 100)

namely:

a) with the minimum actual non-zero contamination of any sample now 10%, the variance of 100 for the wild value seems possibly excessive (counter argument: for our standard one wild at $n = 20$, $5/6$ of the variance of the mean comes from the wild one, increasing this to $10/11$ does not seem excessive).

b) when the "wild one" comes outside of ± 5 , we surely ought to be able to recognize which observation is wild; for our standard one-wild this fails to happen about $2/3$ ds of the time. If marginally wild values are the surest test (are they?) then we might do better with a variance of less than 100 (maximum probabilities near 3 and 4 come from variances of 9 and 16 respectively, (counter argument: the general experience, in the Princeton Robustness Study and its extensions, is that Gau (0, 9) contamination is much easier to deal with than Gau (0, 100)).

Furthermore, it turns out that the triefficiencies involving

9 from Gau (0, 1) and 1 from Gau (0, 9)

are greater than or equal to those involving

9 from Gau (0, 1) and 1 from Gau (0, 100).

Thus the latter are, in fact, tetraefficiencies for

ACCESSION	
NTIS	Wile Section ☒
DDC	B H Section ☐
UNANNOUNCED	☐
JUSTIFICATION	
BY	
DISTRIBUTION/AVAILABILITY CODES	
Dist.	SPECIAL
A	

10 from Gau

9 from Gau (0, 1) and 1 from Gau (0, 9)

9 from Gau (0, 1) and 1 from Gau (0, 100)

10 from slash

More specifically, the best groups of estimators have, at the larger values of c , triefficiencies equal to the efficiency at the wild corner. (cf. Table VII).

2. Preliminaries.

As described in Lax, David A. (1975), "An interim report of a Monte Carlo Study of Robust Estimators of Width", TR #93, Series 2, Department of Statistics, Princeton University, several groups of estimators were originally investigated, the most successful being of the ASYMPV type. In particular the ones corresponding to the bisquare -- $\psi(u) = u(1-u^2)^2$ -- were among the top performers. On the basis of these results, one of us suggested exploring a "W-version" of the estimator involving some 7 free parameters. The estimator was of the form

$$||^2 = \frac{\sum (x - \hat{x})^2 (1-V)^a}{[\sum (1-V)^b (1-5V)^d] [-e + \sum (1-V)^b (1-5V)^d]} \quad (1)$$

where

x = sample or "sample" value

$$\hat{x} = \begin{cases} \bar{x}, & \text{if } g = 0 \\ \text{one step biweight (using } \bar{x}, \text{ MAD)}, & \text{if } g = 1 \end{cases}$$

$$V = \begin{cases} 0 & , \text{ if } |u| \leq f/(1+f) \\ ((1+f)|u| - f)^2 & , \text{ if } f/(1+f) < |u| < 1 \\ 1 & , \text{ if } 1 \leq |u| \end{cases}$$

$$u = \frac{x - \hat{x}}{cS} \quad , \quad S = \text{MAD} .$$

The seven free parameters are labelled a, b, c, d, e, f, g, though not all appear explicitly in formula (1). Here "f" and "g" determine the construction of the normalized residuals, while "a" "b" , and "d" determine how they will be used in the weighting scheme; finally "c" and "e" are meant to be responsive to the long-tailedness of the sample. The variance of the logarithm of the estimator was the measure of performance, and the efficiency was computed relative to the smallest known attainable variance for the given situation.*

A rough plan of the investigation follows below.

- (a) early runs concerned estimators of the form
(1) with $c = 9$
- (b) the form (1) was simplified by removing one of the factors in the denominator
- (c) by allowing e to vary, the denominator occasionally became negative. A modification was introduced to prevent this occurrence
- (d) selected families of estimators (both of (a) and (b) forms) were investigated at different combinations of c and e .

One remark on notation is in order. In what follows, estimators will be identified by their parameter combination (abd - efg), e. g. (411-100). Families of estimators will be denoted by a convenient shorthand. For example (411-100, 200, 300)

*Thus all efficiencies for "slash" or "wild" are apparent efficiencies. (The minimum variance is known for the Gaussian case.)

denotes the family of estimators with $(a,b,d) = (4, 1, 1)$ e varying over $\{1,2,3\}$ and (f,g) held fixed at $(0,0)$. The value of the remaining parameter, c , is given separately.

3. First Monte Carlos.

Run 1 investigated estimators based on (1) at sample size 20. It included the following parameter combinations:

$a = 2,4$
 $(b,d) = (1,1), (4,0), (6,0)$
 $c = 9$
 $e = 0,1$
 $f = 0,1$
 $g = 0$

$a = 1,2,3,4$
 $(b,d) = (1,0), (2,0), (2,1), (4,0), (6,0)$
 $c = 9$
 $e = 0$
 $f = 0$
 $g = 0$

$a = 2,4$
 $(b,d) = (1,1), (4,0), (6,0)$
 $c = 9$
 $e = 0$
 $f = 0,1$
 $g = 1$

The results of Run 1 indicated that one could achieve computational simplicity without the usual corresponding loss of efficiency by setting the parameters f and g equal to zero. It also suggested that it would be fruitful to explore a wider range of values for the parameters a,b,d , and c . Run 2 was executed accordingly and showed some improvements.

A new approach was suggested by recognizing that the denominator in (1) was analogous in form to the factor " $N(-1 + N)$ " common to variance estimates. However, it was not clear a priori

whether the factor "N" was superfluous in some sense and that a new estimator of the form

$$W^{-2} = \frac{\sum (x - \hat{x})^2 (1-V)^a}{[-e + \sum (1-V)^b (1-5V)^d]} \quad (2)$$

might do better as one tried sample sizes different from 20. There would, of course, be an advantage in simplicity as well. Estimators based on (2) are said to be in the "short-form". Estimators using the same set of parameters, but based on (1), are said to be in the "long-form". Run 3 accordingly investigated the short-form of many estimators considered in the previous two runs. Run 4 and 5 investigated the long- and short-form estimators for sample size 10.

A sample of the results obtained in these five runs is given in Table I. For both sample sizes 10 and 20, the long-form estimators are in general superior to the short-form versions. A few groups of estimators seem to stand out. For example at $N = 20$ the long-form estimators (411-000, 100, 200) and (331-100, 200, 300) do very well. Since the efficiencies of these estimators at the three corners are roughly equal, it is apparent that little improvement can be obtained by varying the value of c . However the picture is quite different at $N = 10$, where the triefficiency of the same estimators is limited by the efficiency in the Gaussian corner. Thus, there is some chance that varying the value of c will bring some improvement here. (See Table IV for further results in this direction). Similar considerations apply to estimators that performed best in one category or another. It was clear that we had to experiment with various values of c and an eye towards

perhaps eventually allowing the value of c to vary with sample size, while keeping the other parameters fixed.

One difficulty with using values of " e " greater than one is that occasionally the denominator becomes negative. In order to avoid this unpleasantness, the offending factor was replaced by

$$\max \{1, [-e + \varepsilon(1-V)^b (1-5V)^d]\}$$

in both long and short forms; which are now said to be in modified form. Runs 6 and 7 investigated modified long- and short-form estimators respectively for sample size 10 with values of c equal to 10, 12, and 14. Long-form estimators achieved efficiencies of 89.6% (220-500) and short-form estimators achieved efficiencies of 89.25% (220-600). It was refreshing to see these top estimators coming from the same family. Also the choice $c = 10$ pretty well dominated $c = 12$ or 14.

Finally, using information gathered above, Runs 8 and 9 were planned to give a more complete picture of what had been found. Table II contains the triefficiencies of long-form estimators cross tabulated by parameters e and c for various combinations of parameters a and b . It seems clear that new insights would be needed to bring the triefficiencies of these estimators above 90%. Table III presents the same information for short-form estimators. Run 8 also investigated the performance of two groups of estimators (331-efg and 411-efg) which had done best for $N = 20$, but had only had triefficiencies of 79% for $N = 10$. By trying $c = 11$, triefficiency was raised to above 88% (411-200 ($c = 11$)) which,

of course, is close to the best obtained. These results were encouraging because they pointed to the possibility that a single family of estimators using a value of c chosen to depend on sample size only would attain uniformly high triefficiencies. Run 9 further pursued the performance of certain long-form estimators at $N = 20$ but produced no new stars.

To complete this phase of the work Runs 10 and 11 investigated the three most promising families of estimators 211-(short form), 220(long-form), and 411-(long-form) for different combinations of c and e , at sample sizes 20 and 10. The results are displayed in Tables V and VI.

One point to note is that the value of e denoted by an asterisk "*" is an adaptively chosen quantity equal to

$$N/\Sigma(1-v)^b(1-5v)^d \geq 1.$$

As is readily seen, this choice of e behaved very much like $e = 1$. Furthermore, it is evident that while $c = 9$ is the best choice when $n = 20$, $c = 11$ is the best choice when $n = 10$. What is somewhat surprising is that the above statement holds simultaneously for all three families. We are led to conclude then that we can, with a single estimator, obtain triefficiencies between 85% and 90% for sample sizes $n = 10$ and $n = 20$, requiring only a choice of c dependent on sample size. Further study would undoubtedly disclose how c should, in general, be chosen. Clearly sample sizes such as 5, 7, 40 must be investigated to accomplish the above objective.

Lastly, the three families of estimators discussed in the

previous paragraph, were tried out on samples of size 10 from the wild distribution (cf. discussion in Section 1). At $c = 9$, the triefficiencies were, as before, limited by the efficiency at the Gaussian corner. At the larger values of c , they tended to be limited by the efficiency at the wild corner. The results are displayed in Table VII and should be compared with those in Table VI.

4. Summary.

The previous section detailed the rather high triefficiencies (over 85%) achieved by estimators that are fairly easy to understand and computationally simple. Particular profit was realized by allowing the value of c to vary with sample size. But the optimal relationship between them can not be known until more sample sizes have been investigated. For the moment, we may use $c = 7 + 40/n$ as a rough guide. The importance of the parameter c is not unexpected as it determines the scaling of the normalized residuals and hence how vigorous the down-weighting.

Overall, the estimators of choice would seem to be 411-100, 300, *00 with $c = 9$ or $c = 11$ (according to sample size). They seem to be about the best as well as the most stable in performance.

Dramatic improvements will be hard to come by and fresh insights will be required. One approach that is being currently investigated is to tailor the estimator to the sample, i.e. using a more adaptive method. The problem here is to seize the appropriate sample characteristics on which to base the adaptive nature of the estimator.

TABLE I
1
Triefficiencies of certain estimators

Estimator abd-e fg c	N = 10		N = 20	
	Short	Long	Short	Long
211-0 00 9	78.35	<u>82.34</u>	79.69	76.35
211-1	79.79	81.08	80.01	75.77
211-2	<u>81.41</u>	79.07	80.34	
211-3	<u>82.60</u>	75.21	80.66	
311-0	71.01	80.26	80.22	
311-1	73.03	81.03	80.82	
311-2	75.65	81.60	<u>81.48</u>	
311-3	78.87		<u>82.21</u>	
411-0		76.59		85.8
411-1		78.05		86.2
411-2		79.70		<u>86.59</u>
240-1	74.62	80.67	78.89	83.67
240-2	75.88	81.64	79.29	
240-3	77.58	<u>82.89</u>	79.73	
260-1	74.93	80.35	79.55	84.06
260-2	76.22	81.15	79.99	
260-3	77.96	<u>82.11</u>	80.47	
331-0	67.09	74.04	78.18	85.43
331-1	68.43	74.79	78.92	85.66
331-2	70.20	75.64	79.74	85.88
331-3	72.54	75.44	80.66	<u>86.07</u>

¹ Apparent. Since this report was prepared, the best variance at the slash corner has been reduced by 2.5%. This will temporarily reduce triefficiencies by the same amount. (Long run reductions of perhaps 1.1% can perhaps be compensated for by other improvements not reported here.

For this table, 100% efficiency corresponds to variances of 0.0610420 (Gauss), 0.0932260 (wild), 0.204067 (slash), for N = 10 and 0.0262696 (Gauss), 0.0294203 (wild), 0.0992588 (slash), for N = 20.

TABLE II

N = 10
 Triefficiencies¹ of certain estimators (modified, long-form)

Estimator abd - e	c	9	10	11	12
220 - 0		78.41	83.67	88.01	
1		79.16	84.28	88.50	
2		80.12	85.07	89.11	
3		81.41	86.10	<u>89.62</u>	86.87
4		83.23	87.55	<u>89.29</u>	86.50
5		85.96	<u>89.61</u>	88.39	85.61
6		<u>89.38</u>	<u>89.02</u>	86.36	83.71
240 - 0		79.9	85.22	87.54	
1		80.67	85.84	87.34	
2		81.64	86.61	87.03	
3		82.89		86.54	
4		84.55	88.35	85.63	82.88
5		86.65	86.42	83.71	81.01
6		84.03	81.87	79.49	77.17
250 - 0		79.90	85.29	86.82	84.17
1		80.62	85.86	86.57	83.88
2		81.51	86.55	86.18	83.47
3		82.63	87.38	85.56	82.82
4		84.01	87.15	84.46	81.71
5		85.38	84.77	82.14	79.51
6		82.06	79.70	77.28	74.99

¹ Apparent. (See footnote to Table I)

TABLE III

N = 10

Triefficiencies¹ of certain estimators (modified, short-form)

Estimator \ c (abd - e)	9	10	12	14
220 - 3		81.41	87.36	83.93
4		83.08	87.41	83.79
5		85.60	87.20	83.36
6		<u>89.25</u>	86.30	82.24
240 - 4		85.17	85.17	81.55
5		<u>88.00</u>	84.04	80.29
6		85.94	81.11	77.66
250 - 3		83.51	85.03	81.57
4		85.48	84.42	80.79
5		<u>88.14</u>	82.97	79.21
6		84.08	79.25	75.28

¹Apparent. (See footnote to Table I)

TABLE IV
 $N = 10$
 Triefficiencies¹ of certain estimators (modified, long-form)

Estimator (abd - e) c	10	11
331 - 1	80.73	85.55
2	81.40	86.05
3	81.98	85.81
4	81.87	83.46
5	78.63	78.72
6	71.95	75.48
411 - 1	83.04	87.22
2	84.55	88.56
3	85.98	86.59
4	83.76	82.84
5	78.80	78.04
6	78.55	76.19

¹ Apparent. (See footnote to Table I)

TABLE V
 Triefficiencies¹ of certain estimators
 (N = 20)

c =	7	8	9	11
211 (short)				
1	76.30	82.59	80.01	71.03
*	76.68	82.72	80.07	71.05
3	78.60	83.69	80.66	71.26
6	80.28	79.55	77.89	71.51
220 (long)				
1	73.64	81.56	82.33	74.78
*	73.81	81.69	82.38	74.81
3	74.75	82.41	82.92	75.18
6	77.09	84.16	84.12	76.01
411 (long)				
1	75.63	82.55	86.20	82.93
*	75.93	82.82	86.27	82.97
3	77.38	84.13	86.59	83.32
6	77.53	82.95	82.31	80.28

¹ Apparent. (See footnote to Table I)

TABLE VI
 Triefficiencies¹ of certain estimators
 (N = 10)

	c =	9	10	11	12	13
211 (short)						
1		79.79		84.01	81.82	80.45
*		79.95		83.57	81.45	80.09
3		82.66		80.77	78.81	77.53
6		73.24		70.92	68.36	66.93
220 (long)						
1		79.16	84.28	88.50	87.15	85.23
*		79.25		88.55	87.12	85.18
3		81.41	86.10	89.62	86.87	84.86
6		89.38	89.02	86.36	83.71	81.63
411 (long)						
1		78.05	83.04	87.22	88.58	87.30
*		78.21	(83±)	87.35	88.22	86.94
3		81.05	85.98	86.59	85.54	84.33
6		73.26	78.55	76.19	74.46	73.20

¹ Apparent. (See footnote to Table I)

TABLE VII
Triefficiencies¹ of certain estimators
(N = 10)

	c = 9	11	12	13
211 (short)				
1	79.55	71.20	66.19	61.59
*	79.76	71.26	66.21	61.60
3	80.88	71.43	66.22	61.50
6	71.14	64.43	60.58	56.81
220 (long)				
1	79.16 (85.85)	77.95[88.50]	72.75	67.77
*	79.25 (86.10)	78.12[88.55]	72.84	67.84
3	81.41 (88.25)	79.57[89.91]	74.11	68.92
6	89.38 (96.79)	85.35[94.58]	78.99	73.07
411 (long)				
1	78.05 (88.40)	85.64[87.22]	82.41	78.79
*	78.21 (88.73)	85.79[87.35]	82.47	78.83
3	81.05 (90.75)	86.59[89.88]	83.03	79.23
6	73.26 (79.99)	76.19(73.89)	74.46 (76.81)	73.20 (74.11)

¹ Apparent. (See footnote to Table I)
These are triefficiencies with one of the distributions being the one-wild (with 1 Gau (0,100)). Numbers in () are efficiencies at the one-wild corner when it differs from the triefficiency. Numbers in [] are efficiencies at the Gaussian (which determines the triefficiency here for c = 9).

Also note that 100% efficiency for the one-wild used here corresponds to a variance of .0888569. (The 100% efficiency variance for the other two corners are the same as in the previous tables.)