KU LEUVEN

Robust Statistics: PROJECT-001

Alternatives to MAD estimator

Al-awa Sarah
QingYuan Xue
Ricardo Castañeda
Zachary Jones

Introduction

This study wants to continue the discussion in the paper ***Alternatives to the Median Absolute Deviation*** by Peter J. Rousseeuw and Christophe Croux. The methods discussed in the paper are further evaluated **to confirm that the scale estimator MAD is not the best option for a robust scale estimator**, and to detect whether there are new findings.

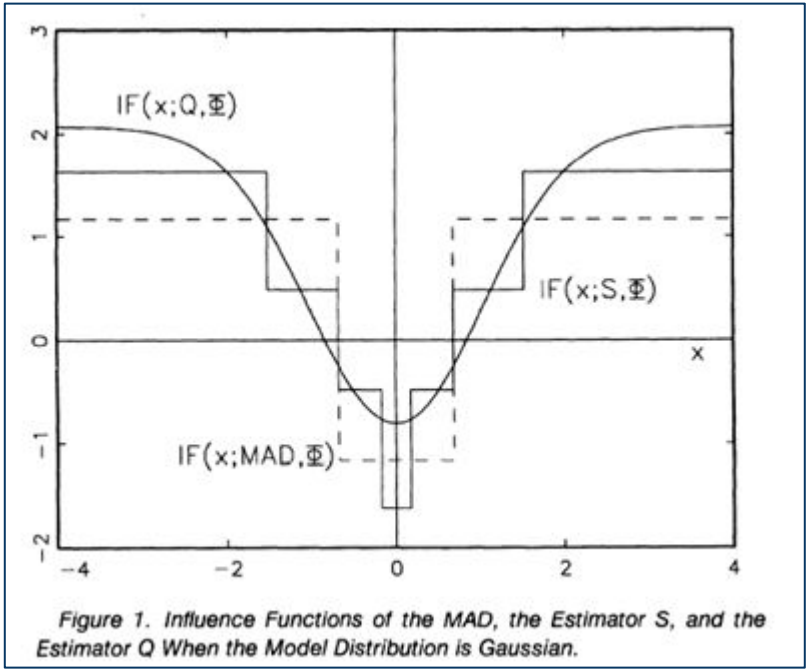


Figure 1. Influence Functions of the MAD, the Estimator S, and the Estimator Q When the Model Distribution is Gaussian.

Methods Discussed in the paper

- **Average Deviation:** Wrongly considered a robust estimator, but it is unbounded. Meaning that it is highly influenced by large observations, so according to the influence of each outlier the estimator may change significantly. Nearly 88% of efficiency.

$$ave|\ x\text{-}ave(x)|$$

- **MADn:** Indirectly assumes the underlying distribution is symmetric which may not be satisfied often, and its asymptotic efficiency is 37%.

$$MAD_{(X)} = b * Med_{(X)}\ |\ x\text{-}Med(x)|$$

$$MADn_{(X)} = \frac{1}{\phi^{-1}(0{,}75)} MAD_{(X)} = \frac{MAD_{(X)}}{0{,}6745}$$

- **Note:** The report mentions Xi, Xj, and Xn but all of them refer to the values from the vector X in the case of average deviation and MADn. Elsewhere, “xi, and xj” represent pairwise distances.

- **Sn Estimator:** It is affine equivariant, has the highest possible break down value, is Fisher consistent, and has an efficiency of 58,23%. Moreover, simulating the standardized variance showed a good approximation to the asymptotic variance of Sn which was given by 0.8573 in eq. 2.9[6].

$$Sn_{(X)} = c * Med_{(i)}|Med_{(j)}(x_i - x_j)|$$

Methods Discussed in the paper

- **Qn:** It needs $O(n^2)$ space and $0(n^2)$ time. But Croux and Rousseeuw (1992b) have constructed an algorithm for computing Qn with O(n) space and O(n log n) time. It has a 50% breakdown point and almost 82% of efficiency[6]. For consistency at Normal dist. d=2.21914.

$$Qn = d\{|x_i - x_j|; i < j\}_{(k)}$$

d is a constant factor and $K = \binom{h}{2}$; *where* $h = [n/2] + 1$

- **LMSn:** It's a location-free estimator so it works well in the case of asymmetrical distributions. 50% breakdown point, boundedness, an efficiency of 36,74%. c'= 0:7413, which achieves consistency at Gaussian distributions and h=[n/2]+1 a half sample.

$$LMSn = c' \, min_{(i)} |x_{(i+h-1)} - x_i|$$

- **The Bickel-Lehmann estimator:** Boundedness, 29% breakdown point, and Gaussian efficiency of about 86%. u =1.0483 for consistency under Gaussian distribution.

$$BL = u * med(|x_{(i)} - x_{(j)}|; i < j)$$

Estimators Features

Under Normal Distribution

- MAD shows lower efficiency (IF and sensitivity curves differ from SD) than Sn and Qn.
- Gross error sensitivity of 1.167, 1.625, 2.069 for MAD, Sn, and Qn respectively.
- Explosion, Bias curve Sn is almost as good as MAD.
- Implosion, Bias curve MAD is slightly better than Sn.
- Implosion and Explosion, Bias curve Qn is better than MAD and Sn.
- LMSn which in the case of normality has an influence function (IF) equivalent to the MADn.

Under Exponential and Cauchy Distribution

- In Cauchy, the Gross error for Qn = 2.2214.
- In Cauchy distribution, the absolute asymptotic efficiencies become e(MADn) = 81%, e(Sn) = 95%, and e(Qn) = 98%.
- LMSn is less efficient than both Sn or Qn.

Sensitivity Curves For the Finite sample size of 100 (One Run)

- Consistency factors are estimated first in all three distributions (500 trails with the finite sample size of 1000).
- Average Deviation does not have the consistency factor as it is not Fisher consistent that satisfies(the mean for Cauchy is undefined): $\hat{T}_\infty(F_\theta) = \theta$

Consistency Factors Estimated in Gaussian, Exponential, and Cauchy distribution

	Average Deviation	MAD	Sn	Bickel-Lehmann estimator	Qn	LMSn
Gaussian Distribution	1.2556	1.4944	1.1934	1.0490	2.2128	0.7623
Exponential Distribution	1.3603	2.0768	1.6974	1.441	3.4596	1.4438
Cauchy distribution	*	0.9970	0.70541	0.498618	1.19928824	0.50768

Sensitivity Curves Gaussian Distribution

1. The Average Deviation does not show robustness and is unbounded
2. Qn estimator is bounded and has the third smallest gross error sensitivity, but it is not significantly different from the gross error for Sn and BL, indicating strong robustness.
3. The Sn estimator shows adequate and higher efficiency than the Qn estimator according to the curve smoothness
4. MAD estimator appears to be robust due to its boundedness, it has a largest gross error sensitivity when the outliers are taken negative, while the second smallest when x>0. Its shape is not smooth, and its efficiency is significantly low as well. These facts are reflected when comparing the curve with the SD curve.

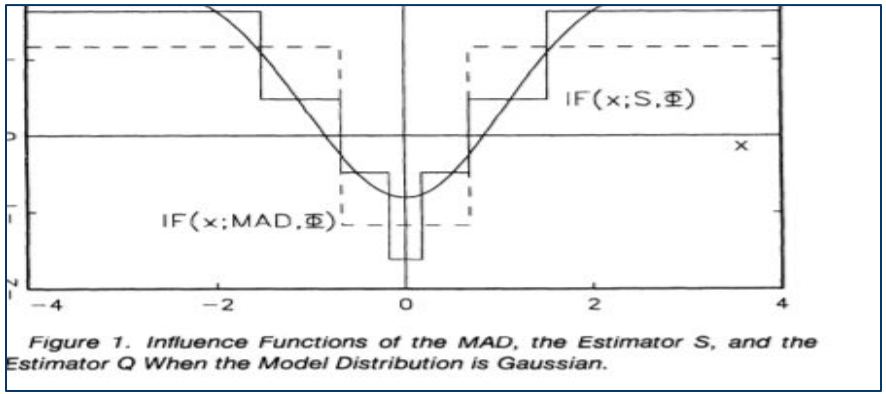


Figure 1. Influence Functions of the MAD, the Estimator S, and the Estimator Q When the Model Distribution is Gaussian.

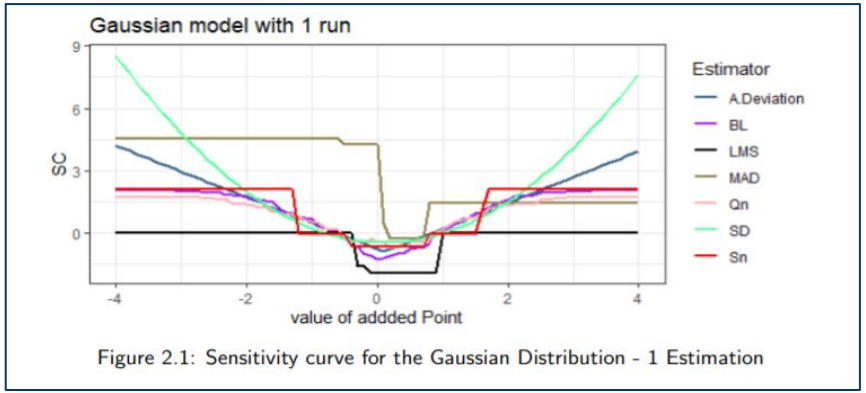


Figure 2.1: Sensitivity curve for the Gaussian Distribution - 1 Estimation

Cauchy Distribution

- MADn, Qn , and Sn estimators are all bounded, while their efficiencies are shown to be different: Qn>Sn>MAD.
- There are inconsistencies when comparing each of the SC to their corresponding influence functions(IF); their curves' approximation(convergences) to the IF are dependent on the size of the samples for different values of outliers.
- The Average Deviation estimator seems to be bounded with the lowest gross error. However, the interpretation of such result should be limited, as it is contradictory to the theory's suggestion.

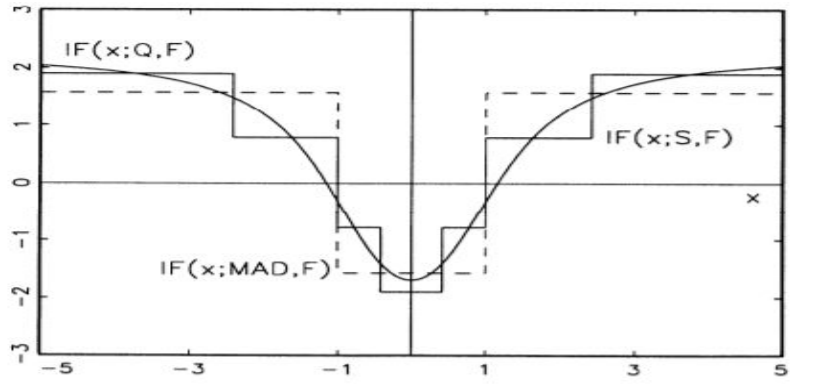


Figure 3. Influence Functions of the MAD, S, and Q at the Cauchy Distribution.

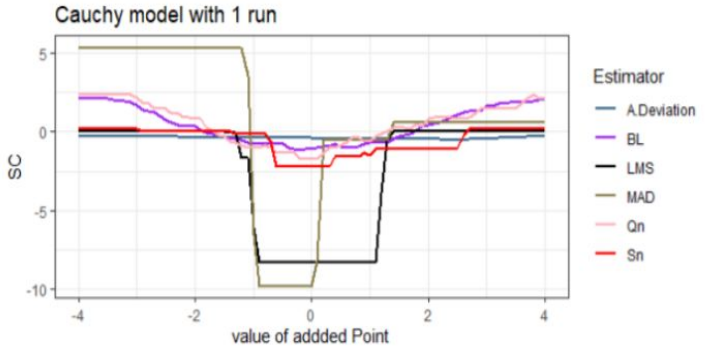


Figure 2.2: Sensitivity curve for the Cauchy Distribution - 1 Estimation

Exponential Distribution

- The sensitivity curves observed of all estimators, except for SD and Mean deviation, seem to be bounded in the graph.
- Sn, Qn, and Mad by the shape of their curves show the similar efficiency.
- Mad has the lowest gross error with the negative outliers, while LMS overtake MAD when outliers are positive.

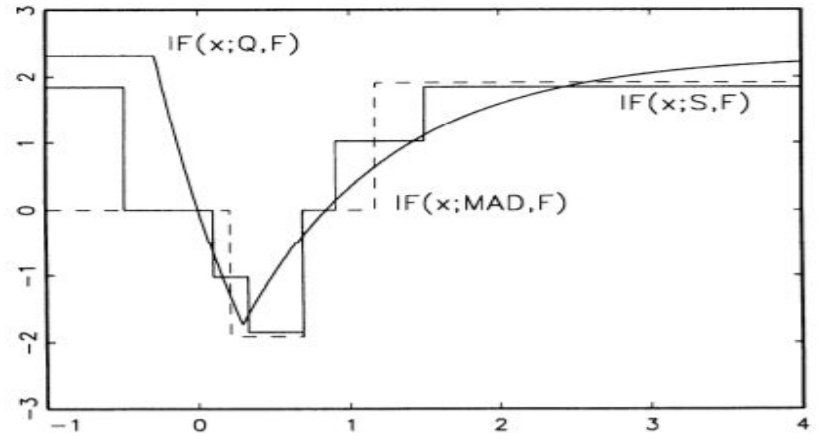


Figure 4. Influence Functions of the MAD, S, and Q at the Exponential Distribution.

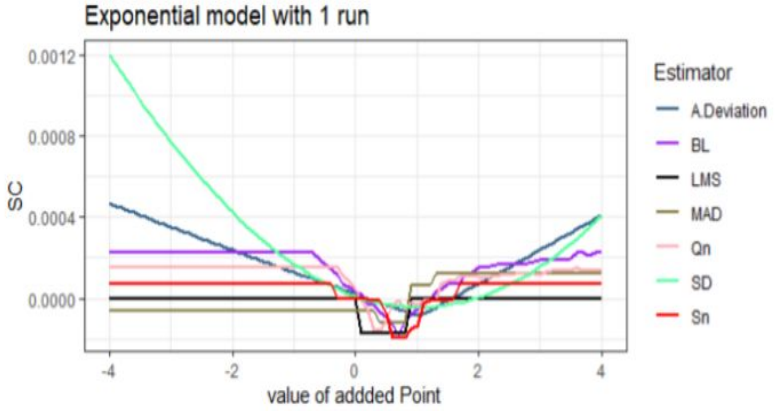
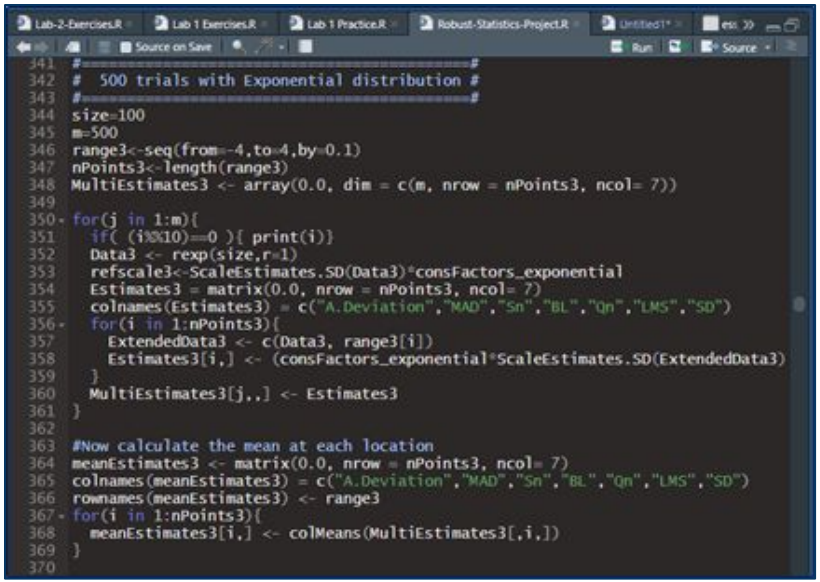


Figure 2.3: Sensitivity curve for the Exponential Distribution - 1 Estimation

Sensitivity curves using 500 runs

Based on the obtained sensitivity curves and to obtain smoother curves, the derivation of the estimators was repeated 500 computation times (m=500) and then the curves were calculated again after averaging the 500 results.



```
#=========================================#
#  500 trials with Exponential distribution #
#=========================================#
size=100
m=500
range3<-seq(from=-4,to=4,by=0.1)
nPoints3<-length(range3)
MultiEstimates3 <- array(0.0, dim = c(m, nrow = nPoints3, ncol= 7))

for(j in 1:m){
  if( (j%%10)==0 ){ print(j)}
  Data3 <- rexp(size,r=1)
  refscale3<-ScaleEstimates.SD(Data3)*consFactors_exponential
  Estimates3 = matrix(0.0, nrow = nPoints3, ncol= 7)
  colnames(Estimates3) = c("A.Deviation","MAD","Sn","BL","Qn","LMS","SD")
  for(i in 1:nPoints3){
    ExtendedData3 <- c(Data3, range3[i])
    Estimates3[i,] <- (consFactors_exponential*ScaleEstimates.SD(ExtendedData3)
  }
  MultiEstimates3[j,,] <- Estimates3
}

#Now calculate the mean at each location
meanEstimates3 <- matrix(0.0, nrow = nPoints3, ncol= 7)
colnames(meanEstimates3) = c("A.Deviation","MAD","Sn","BL","Qn","LMS","SD")
rownames(meanEstimates3) <- range3
for(i in 1:nPoints3){
  meanEstimates3[i,] <- colMeans(MultiEstimates3[,i,])
}
```

Normal distribution using 500 runs

- Significant improvement in the curves' smoothness, as the average of 500 runs tends to show a more accurate value for true sensitivity.
- The BL is more efficient (closer to SD curve) than the Sn estimator.
- The MAD sensitivity curve went smoother, shows more efficiency compared to the experiment with 1 run, and the gross error seems a whole lot smaller.
- The Sn decreased its efficiency but remains robust.

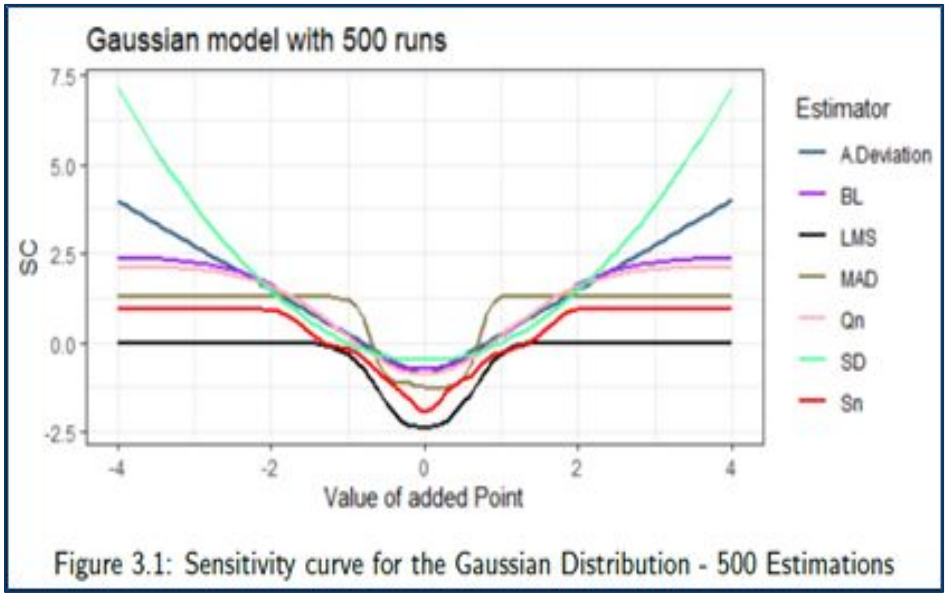


Figure 3.1: Sensitivity curve for the Gaussian Distribution - 500 Estimations

Cauchy Distribution using 500 runs

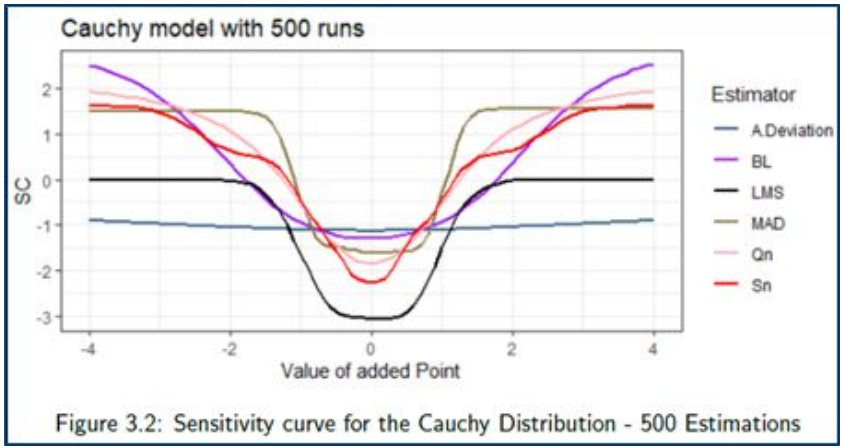


Figure 3.2: Sensitivity curve for the Cauchy Distribution - 500 Estimations

- MAD depicted smoother curvature, symmetric boundaries, and improved its efficiency and gross error.
- Sn has a smoother curve, lower efficiency, and a higher gross error.
- Qn and BLn estimators appear to be no longer bounded and their gross error has a sharp increase.

Exponential Distribution using 500 runs

- **MAD has the lowest gross error** for negative outliers as well as LMSn which seems identical to the one for MAD.
- LMSn estimator surpasses MAD as the most robust for positive outliers.
- BL estimator has shown slightly better efficiency but the lowest robustness compared to other robust estimators.

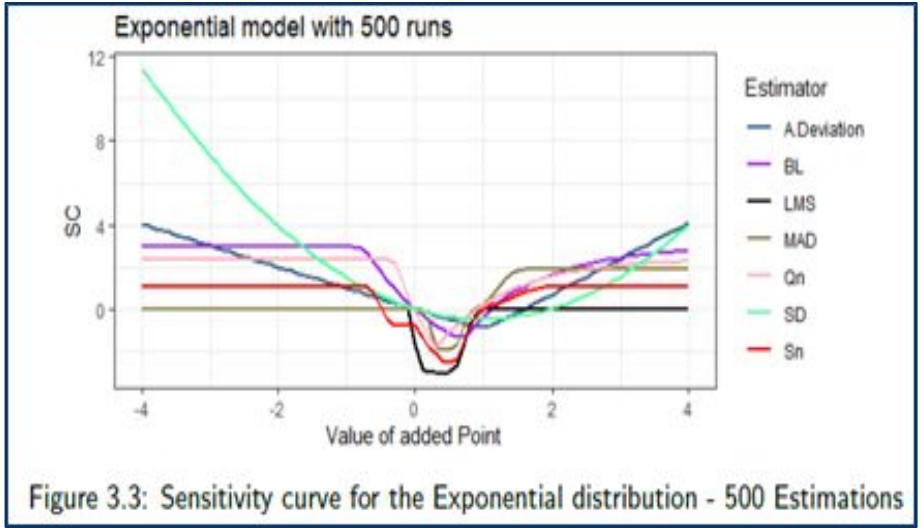


Figure 3.3: Sensitivity curve for the Exponential distribution - 500 Estimations

Gleaning Asymptotic properties on Gaussian Data

- For n not too small standardized variance is a good approximation of asymptotic standardized variance:
  - STDVAR = m var(Tn(F)) / (Tn(F))2
- More robust MADn LMSn Sn Qn not as efficient as mean ADn BLn.
- LMSn and MADn show same efficiency.
- Qn small sample size bias, more efficient than Sn and MADn.

Average Values

n	$MeanAD_n$	MAD_n	S_n	BL_n	Q_n	LMS_n
10	0.948(0.628)	0.918(1.342)	0.998(1.106)	1.047(0.738)	1.393(0.91)	0.767(1.254)
20	0.971(0.613)	0.956(1.393)	0.996(1.022)	1.018(0.666)	1.183(0.8)	0.815(1.239)
40	0.989(0.584)	0.98(1.348)	1.002(0.908)	1.012(0.609)	1.094(0.699)	0.869(1.182)
60	0.991(0.584)	0.99(1.369)	1.002(0.878)	1.007(0.614)	1.061(0.683)	0.893(1.193)
80	0.994(0.566)	0.992(1.349)	1.001(0.849)	1.006(0.588)	1.047(0.647)	0.91(1.164)
100	0.994(0.565)	0.991(1.328)	0.999(0.846)	1.003(0.579)	1.036(0.637)	0.918(1.178)
200	0.997(0.566)	0.995(1.345)	0.999(0.851)	1.001(0.584)	1.018(0.628)	0.945(1.206)
1000	0.999(0.573)	0.999(1.347)	0.999(0.857)	1(0.583)	1.003(0.618)	0.979(1.229)

Table 4.1: Average values (averaged standardized variance) pairs for the six estimators taken over 5000 random trials with increasing sample size from a standard Gaussian distribution

Asymptotic Properties of Cauchy Data

- Not all symmetric distributions are Gaussian. Often times real distributions have 'fat tails'. Archetypal of a Cauchy distribution.
- Mean AD does **not converge.**
- Estimators which were less efficient on Gaussian data (MADn, Sn, Qn, and LMSn) now more efficient than their counterparts.
- Still prevalent: Small sample size bias of Qn, MADn and LMSn similar efficiencies!

Average Values

n	$MeanAD_n$	MAD_n	S_n	BL_n	Q_n	LMS_n
10	27.777(6384635.32)	1.112(4.746)	1.148(4.252)	1.341(12.671)	1.659(4.89)	0.96(5.005)
20	10.049(73629.79)	1.048(3.372)	1.05(2.878)	1.114(4.121)	1.288(3.072)	0.918(3.433)
40	16.032(2173368.85)	1.02(2.86)	1.017(2.372)	1.052(3.32)	1.132(2.5)	0.919(2.75)
60	11.451(683683)	1.013(2.684)	1.01(2.242)	1.031(2.942)	1.085(2.303)	0.929(2.607)
80	11.756(288309.26)	1.011(2.626)	1.008(2.179)	1.024(2.856)	1.065(2.229)	0.937(2.491)
100	11.637(521695.09)	1.009(2.648)	1.007(2.27)	1.021(2.832)	1.052(2.251)	0.943(2.537)
200	12.543(3287763.95)	1.005(2.456)	1.004(2.103)	1.01(2.618)	1.027(2.062)	0.96(2.315)
1000	16.963(48276696.54)	1.003(2.455)	1.003(2.124)	1.005(2.672)	1.007(2.081)	0.985(2.337)

Table 4.2: Average values (averaged standardized variance) pairs for the six estimators taken over 5000 random trials with increasing sample size from a standard Cauchy distribution

Consistency factors can also be derived!

The consistency factors can be found experimentally by observing **the asymptotic limit for the value of the scale estimate**. We cannot do this so we take the large n limit. These factors are approximations (and good ones!) of true asymptotic factors.

- BLn can be found asymptotically as BL(F) = cH-1F(½)
  - HF = L(| X - Y |), and L(| X - Y |) = 2FZ(z) - 1 for Z = X - Y, z >= 0 Z ~ Cauchy(0,2)

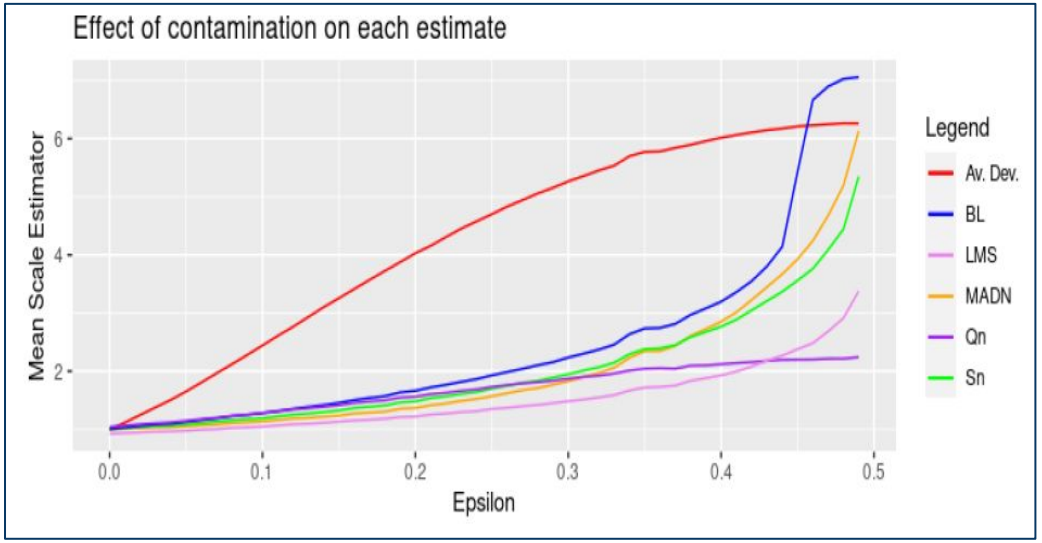
- LMSn : The asymptotic consistency factor becomes the same as the IQR
  - Intuitively, for a symmetric unimodal distribution the shortest interval which contains 50% of the data is the IQR as a majority of the probability mass is located near the median.
  - This would not hold for exponentially distributed data, which is asymmetric.

Studying the effect of asymmetric corrupt data

Setting:

- Running a simulation with n = 100. Where,
  - Fraction **(1 − ε) of clean data** follow a standard normal **N(0, 1)**
  - Fraction **ε of contaminated data** are from **N(a, 1).**
- Make two plots:
  - One with the average estimated value for fixed a = 10 where ε ranges from zero to 0.49,
  - The other with fixed ε = 0.2 where a ranges from zero to 10.
  - Meaning: Examine the estimators behavior with varying proportions of corrupted data and large deviations versus a fixed proportion of corruption and varying deviations.

Studying the effect of asymmetric corrupt data

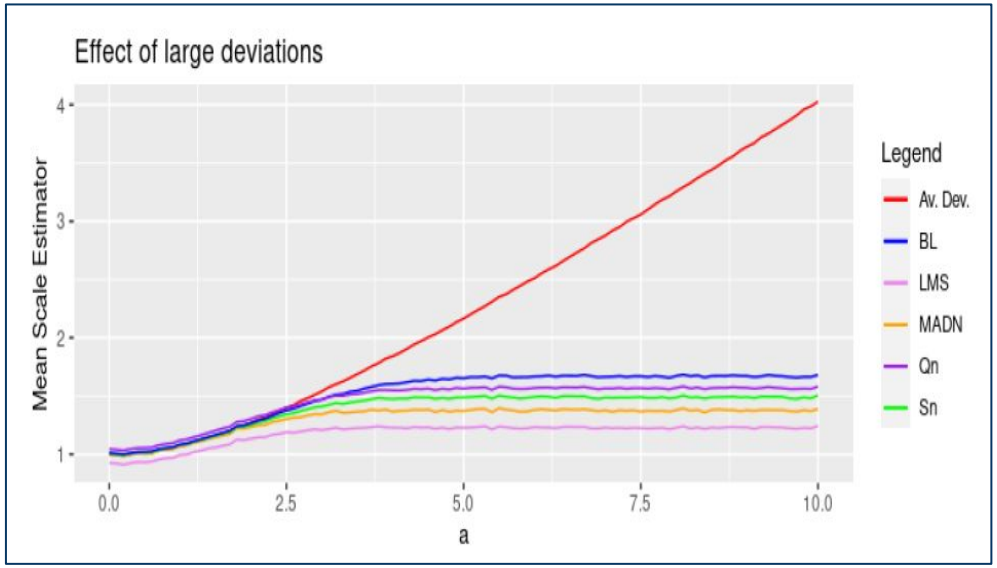


- LMSn, MADn, Qn and Sn all behave similarly.
- Of which, MADn is least robust., however it performs similarly to Sn.
- **Qn and LMSn perform with the least change** in value.

- Enables a **qualitative comparison of the robustness** of the various estimators.
  - **Desirable** => when estimator is resistant to increasing number of corrupted data, so the estimators whose value changes most rapidly are least robust.
- The Mean Average Deviation i**s the least robust** as evidenced by the rapid increase in value.
- The **BLn estimator is fairly robust up to a point,** but not as robust as LMSn, Qn, Sn, and MADn. This behavior corresponds well to the breakdown value of 29% versus the 50% breakdown value of the other estimators. (see past 29% , BLn surpasses Mean average deviation at 46%).

Studying the effect of asymmetric corrupt data

We visualize the estimates behaviors for large deviations of the corrupted data, for a moderate proportion of corruption.



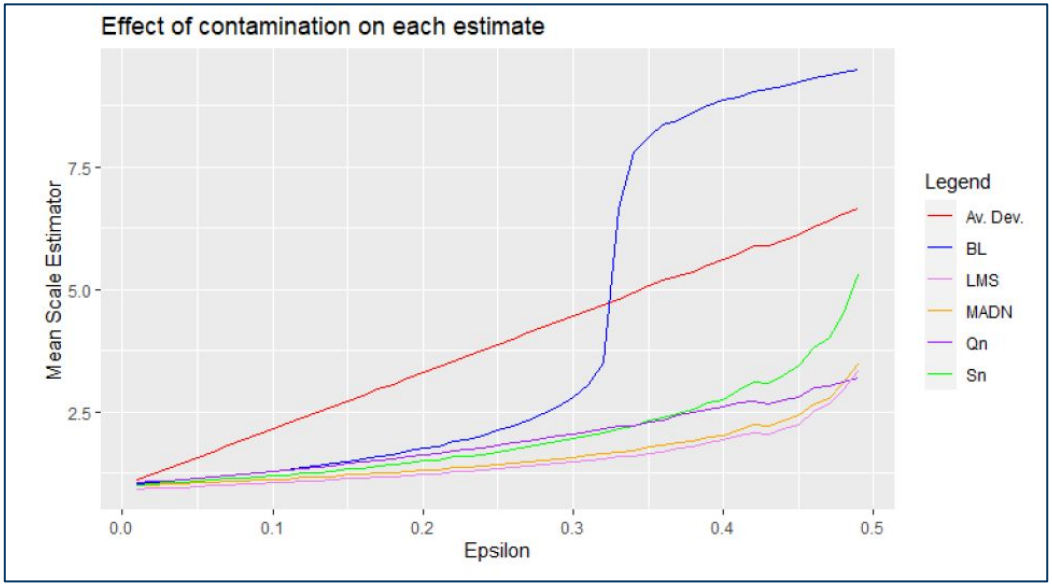
- For a moderate amount of corruption, we can see how the estimators change in value.
- Since ε = 20% <29% (the breakdown value of the BLn estimator), Therefore, **all but the mean average deviation** will have a **bounded change**.
- Furthermore, these bounds are similar to those found in questions 2 and 3 and closely mimic the results seen in a sensitivity curve analysis. To elaborate on this effect, we notice that in 5.1 the change in value due to the corrupt data is dependent on the percentage of corruption present.

Studying the effect of symmetric corrupt data

Setting:

- Running a simulation with n = 100. Where,
  - Fraction **(1 – ε) of clean data** follow a **N(0,1)**.
  - Fraction **ε of contaminated data** are from **N(-a, 1) to N(a, 1)** evenly.
- Make the same two plots
- Meaning: Are there any differences when the corrupt data is symmetric? Are there any differences in an increasing proportion of corrupted data? Does symmetric corrupt data produce any changes in behaviors of the scale estimates for large deviations?

Studying the effect of symmetric corrupt data



- The mean average deviation, is the same as in the case of asymmetric corrupt data.
- The **LMSn, the MADn and the Sn, have again similar behaviors**. The LMSn and MADn seem closer in behavior than in an asymmetric case.
- The **Qn remains seemingly the most robust** for most. Overall, approaching towards the 50 percent limit of outliers, the most robust outliers are the Qn, LMSn, MADn and the worst is the BLn estimator.
- With a symmetrical distribution of corrupt data, behaviors in **robustness of the estimators can vary** like the MADn.

- The **BLn estimator increases abruptly at approx. 32 % of contamination,** which surpasses the average deviation.
  That happens much earlier than in Q5, that was at approx. 45 %. Meaning that, the BLn is less robust earlier passed an increasing proportions of outliers when the contamination is symmetric. **WHY is that ? Let's take a look at the next slide**

Studying the effect of symmetric corrupt data

Understanding the intuition why BLn increases faster in a symmetric corrupt data

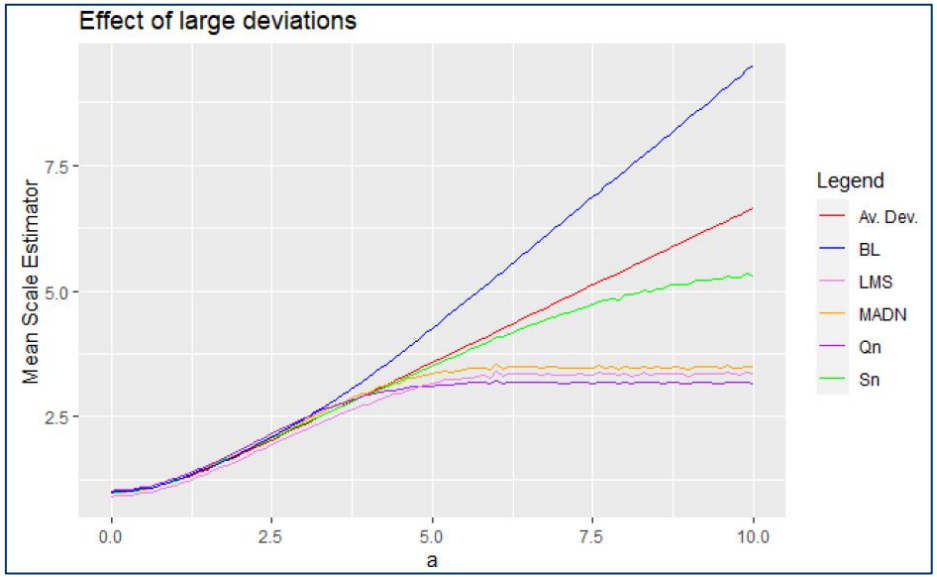
- BLn estimator is as follows: $$BL = u * med(|x_{(i)} - x_{(j)}|; i < j)$$

- We understand that BLn **computes interpoint distances.** Therefore when corrupted with a certain proportion of data:

  - Asymmetrically: The distribution centered around N(a,1), where a=10. The interpoint distances are pulled by this distribution. Therefore are **dragged to the right graphically.**

  - Symmetrically: The distribution is from N(-a,1) to N(a,1) evenly. The interpoint distances are pulled to both -a and a. The BLn estimator therefore shifts **more to the left compared to previous case**.

Studying the effect of symmetric corrupt data



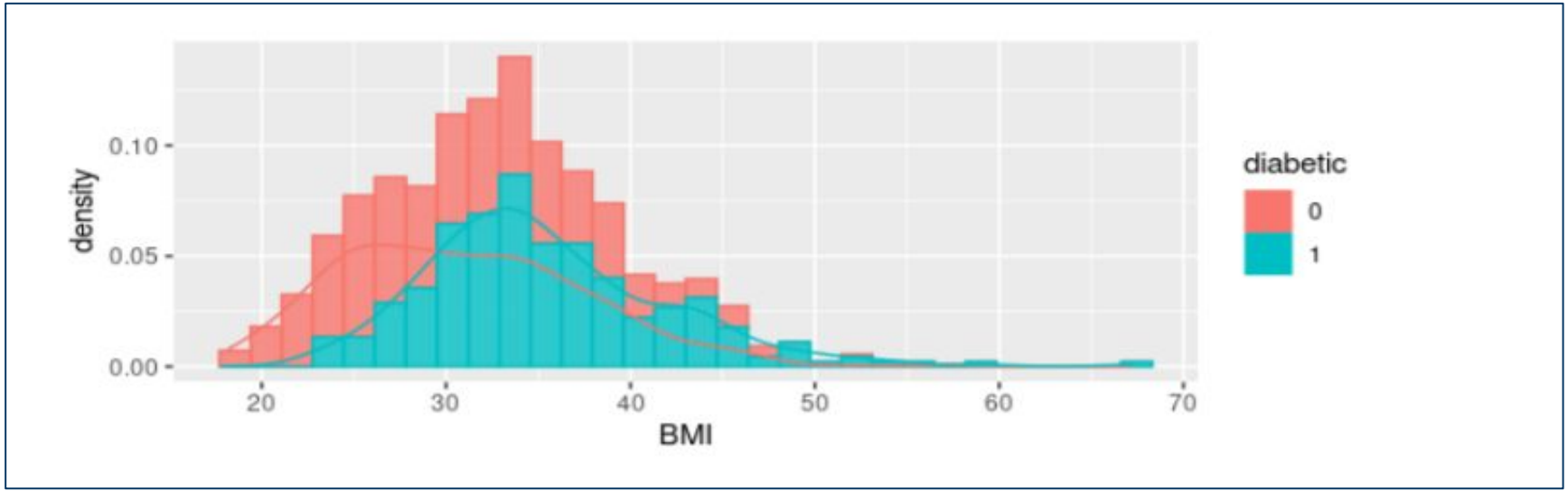
- **MADn, LMSn, and Qn are bounded.**
- BLn, Average Deviation, and Sn **seem divergent** on this plot.
- However **BLn and Sn are known to be bounded** as a fixed level of ε= 0.20. This can be explained by going back to the definitions of these estimates: In a symmetric case. The median of the interpoint distances does not differentiate between corrupt data and the correct data.
- In our understanding, there is a link between a symmetrical distribution of the corrupted data and the behavior of estimators with a magnitude of deviation.

Assessing the Estimator's Performance on Pima Indians

Dataset: The Pima Indians diabetes: Females at least 21 years of Pima Indian heritage.

Our database contains two population's body mass index(BMI), one with diabetes and another without diabetes.

The two population's BMI distribution



Detecting the Outliers

- Deriving consistency factors for the Pima Indians diabetes.
  - First normalize so that scale = 1 to find consistency factor with same scale as SD.
  - Repeatedly subsample and average results to find consistency factor.

	MeanAD	MADn	Sn	BLn	LMSn	SD
Consistency Factors	1.267	1.494	1.201	1.069	2.256	0.7770

- Checking for outliers: $z_i = \frac{x_i - \mu(X)}{\sigma(X)} \geq 3,$

Where μ is chosen to be median, and σ is varied for accessing the performance of the six estimators. Using the above equation, those BMI values between 52.9 and 58 were selected as outliers.

	σ:Mean AD	σ:MADn	σ: Sn	σ: BLn	σ:SDn	σ:Qn	σ:LMSn
Number of outliers detected	6	6	6	5	5	5	6

The Interpretation

There are little differences in the results of estimators; the largest of 5% is found in SD for reasons of **minimal contamination and the distribution of the outliers is not extreme.**

The distance between outliers and regular observations is not significantly far or separated, and indeed, the percentage of outliers detected is only 0.8% of the the data set.

	mean average deviation	$MADN$	S_n	BL	Q_n	LMS	SD
6 removed	6.645356	6.6717	6.79782	6.70912	6.65742	6.59757	6.565042
5 removed	6.671014	6.74583	6.79782	6.70912	6.65742	6.59757	6.603713
Clean Data	6.845747	6.81996	6.79782	6.81395	6.879334	6.6717	6.924988

Table 7.2: In this table, rows correspond to the estimator used to remove outliers and columns correspond to the method of estimation. Values have been rounded to three significant figures. Estimators do not change much when outlying values have been removed

Trade-off Between Robustness and Efficiency

- Relatively high standardized variances on Gaussian distributed data for MAD_n, Q_n, S_n, LMS_n versus BL_n and Mean AD.
- First steps to robustify an estimator → Bounded Influence function
  - Influence Function measures the effect on an estimator of a small fraction of identical outliers at a point.

- There is a tradeoff between a low GES and a high efficiency
  - Gross Error Sensitivity: $\sup_x |IF(x,T,F)|$ can be seen as the worst approximate influence that a fixed amount of contamination can have on an estimator.
  - If GES is low, the smallest possible variance $V = \text{int } IF(x,T,F)^2 dF$ will be large, limiting the efficiency.

- Another desirable quality, a high breakdown value, is a global property, while IF is local.

Sensitivity curves and Influence Functions

- Sensitivity curve measures the effect of the location of an outlier on a statistic, Influence function measures the effect of contamination at a point.
- In the large n limit these become the same concept, a small point outlier becomes equivalent to an infinitesimal amount of corruption at a point.
- We can use the sensitivity curves as an approximation of the influence function to gain insight into the behavior of these estimators.

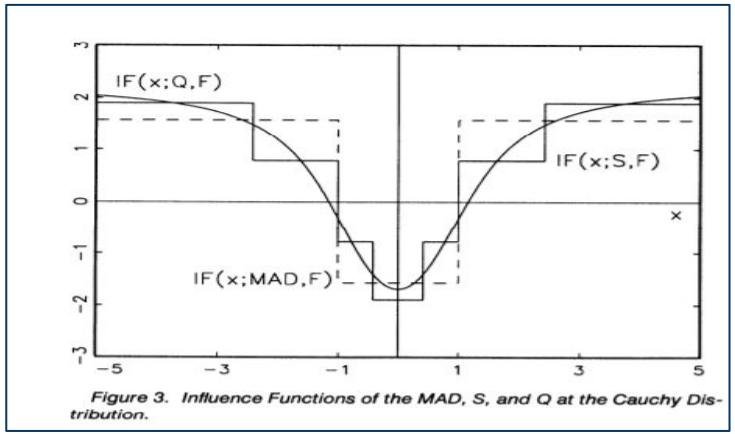


Figure 3. Influence Functions of the MAD, S, and Q at the Cauchy Distribution.

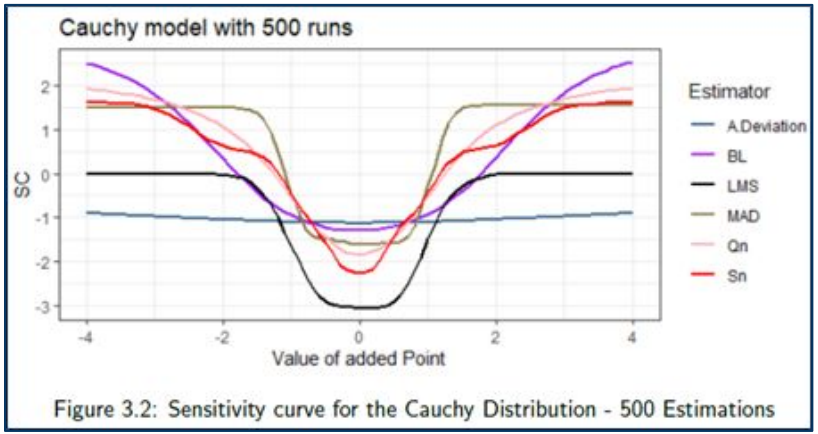


Figure 3.2: Sensitivity curve for the Cauchy Distribution - 500 Estimations

Conclusion

- Robust Estimation of scale is not always preferable:
  - Using a statistical test (ex. Shapiro-Wilk) or QQ plot one can examine data for deviations from normality - These are flawed and one can use robust estimation to check validity of standard results.
  - Avoid the tradeoff between efficiency and low gross-error sensitivity if data is ‘close enough’ and take advantage of guarantees given by ML estimators.

- Robust Estimation of scale can be **beneficial under model misspecification**
  - We saw that Q_n, S_n, MAD_n, outperform BL_n, and mean average deviation on Cauchy distributed data.

- Robust estimation of scale **comes at a cost of computation time**
  - Intuitively: Identifying points to base estimates off requires computational power and time to be spent searching.
  - Numerically: mean average deviation O(n), other estimators were O(n log(n)) or, in the case of Bln $O(n^2)$.

Bibliography

[1]R Grubel et al. “The length of the shorth”. In:The Annals of Statistics16.2 (1988), pp. 619–628.

[2]F Hampel et al. “Robust Statistics: The Approach Based On Influence Functions”. In:Wiley-Interscience502 (2005).

[3]Lecture notes in Robust Statistics. Feb. 2018.

[4]Ricardo A Maronna et al.Robust statistics: theory and methods (with R). John Wiley & Sons, 2019.

[5]F. Masci.Lecture notes in Statistics. June 2013.url:http://web.ipac.caltech.edu/staff/fmasci/home/mystats/CauchyVsGaussian.pdf.

[6]Peter Rousseeuw and Christophe Croux. “Alternatives to Median Absolute Deviation”. In:Journalof the American Statistical Association88 (Dec. 1993), pp. 1273–1283.doi:10.1080/01621459.1993.10476408.

[7]Peter J Rousseeuw and Annick M Leroy. “A robust scale estimator based on the shortest half”. In:Statistica Neerlandica42.2 (1988), pp. 103–116.

[8]J Tukey.A survey of sampling from contaminated distributions. Contributions to Probability andStatistics: Volume Dedicated to Harold Hetelling. 1960.

[9]Peter J Rousseeuw and Christophe Croux. “Explicit scale estimators with high breakdown point”. In: L1-Statistical analysis and related methods 1 (1992), pp. 77–92.