KU Leuven

Robust Statistics

Project: Alternatives to the median absolute deviation

AL-AWA Sarah (r0688008)
Ricardo Castañeda (r0731529)
Zachary Jones (r0772334)
QingYuan Xue - (r0829510)

April 19, 2021

Contents

1.0 Q1. Describing some alternative scale estimators

1.1 Presenting Alternative Estimators

Briefly describe these methods, with the definitions and main results, but without any proofs or material from the Appendix.

1.1.1 The average deviation

Thought to be robust, the average deviation is such as

$$avedev = ave|x_i - ave_j x_j|$$

In the case of the usual parameter at a Gaussian distributions, the average deviation needs to be multiplied by 1.2533 to become consistent. The method would imply to take the averages of observations x_j, and to then take the absolute differences with regards to the observations x_i, and then average them. The average deviation has breakdown value of zero percent, and it is an unbounded location-based estimator. It is therefore not considered a robust estimator. It however presents an efficiency of approximately 88%[8]. Computing it has $O(n)$ time complexity.

1.1.2 The normalized MAD

The normalized MAD (MADN) follows the following formula:

$$MAD_n = b * med_i|x_i - med_j x_j|$$

Where b=1.4826 is the consistency factor, that makes the difference with the MAD estimator.

At a normal model,

$$MAD_n = \frac{1}{\Phi^{-1}(0.75)} MAD(X_n)$$

The method of computing the MADN (or MADn) is to instead of taking the mean like that of the average deviation we take the median. This estimation is an location-based estimator as it starts by computing a location estimate (the median) and then considers the absolute deviations from it. One of the limit of this estimator is that they attach equal importance to positive and negative deviations, which means that they indirectly assume the underlying distribution to be symmetric which may not be satisfied often. This scale estimator is an explicit formula that has a a 50 percent breakdown point, a bounded influence function and has the characteristic of being easily computable, using at most O(n log n) time and O(n) storage. Unfortunately its Gaussian efficiency is only 37 percent, which is quite low.

1.1.3 The estimator Sn

$$S_n = c * med_i(med_j|x_i - x_j|)$$

Where the c value is 1.1926 in order to be consistent under Gaussian distributions.

The Sn estimator is another explicit scale estimator in which we would compute the median of each absolute difference between x_i and x_j, where we then would take the median of the n numbers of computed medians. It has the following characteristics by being: affine equivariant, the breakdown point of Sn is the highest possible, Sn is also Fisher consistent, and has an efficiency of $58,23$ percent. It is computable by an O(n log n) time algorithm. Sn differs from the MAD as it is a location free estimator. It looks at a typical distance between observations, which is valid for asymmetric distributions. Simulating the average scale estimate on 10,000 batches of Gaussian observations. The influence functions show that Sn behaves better than MADn as it is the closest to the influence function of the classical standard deviation. Therefore, making Sn approximately unbiased for finite samples. Repeating the experiment for an increasing number of outliers, the results (bias curves) show that the asymptotic variance provides a good approximation for (not too small) finite samples, Sn remains still more efficient than MADn for small samples. A limit worth mentioning is the discontinuous nature in the influence function while the influence function of the SD is continuous and smooth.

1.1.4 The Bickel-Lehmann estimator

$$BL = u * med(|x_i - x_j|; i < j)$$

Where the consistency factor is set to 1.0483 to be achieved for σ under Gaussian distribution.

The Bickel-Lehmann estimator (BLn) is similar to Sn, However it takes a global median of the absolute distance between the points x_i and x_j, when i and j are different. The characteristics of the BL estimator are defined by: boundedness, a 29 percent breakdown point, but a quite high Gaussian efficiency of about 86 percent. In terms of computational time it need $O(n^2)$ A limit worth mentioning is its low breakdown point, that could be increased by means of another proposed estimator Qn.

1.1.5 The estimator Qn

$$Q_n = d(|x_i - X_j|; i < j)_k$$

Where it is multiplied by a constant $d = 2.21914$ in order to be consistent under Gaussian distribution. $k = \binom{h}{2}$ where $h = [n/2] + 1$.

The Qn estimator is computed by taking the absolute distance between the x_i and x_j, when i differs from j. Of which we take an ordered value corresponding to k. Its characteristics are defined by: an explicit formula, being a location free estimator-therefore suited for asymmetrical distributions, boundedness, a 50 percent breakdown point, while keeping an efficiency level of approx. 82 percent. The influence curve of Qn is smooth, being close to the BLn efficiency level. Additionally, the computational complexity is solved by a O(n log n) time.

1.1.6 The estimator LMSn

$$LMS_n = c'min_i|x_{i+h-1} - x_i|$$

Where c'$= 0.7413$, which achieves consistency at Gaussian distributions and h=[n/2]+1 a half sample.

The method of the LMSn implies minimizing the shortest half sample. The LMSn estimator is a location free estimator, therefore works well in the case of asymmetrical distributions. Some notable characteristics include: being an explicit scale estimator with a 50 percent breakdown point, boundedness, an efficiency level of $36, 74$ percent. Its influence function (hence efficiency) is the same as that of the MAD. Therefore the Sn and Qn are more efficient alternatives. Computationally wise it also only needs explicit scale estimator with a 50 percent breakdown point and a $O(nlogn)$ time.

1.2 Main Findings of our reference paper

In summary, the influence functions that are derived asymptotically give us the information in which the estimators behave with the presence of an outlier. While the MADn efficiency is quite undesirable, the SN and Qn influence curves showed much higher efficiencies. However, the MADn remains more robust than the Sn estimator, with a lower Gross Error sensitivity of 1.167, while Sn has a Gross Errors sensitivity of 1.625. The Qn estimator is the least robust with regards to both estimators with a Gross Error sensitivity of 2.069. Its efficiency in a Gaussian distribution is the highest of the three.

Rousseeuw and Croux, verified their results with finite-sample simulation and the results show that the asymptotic variance provides a good approximation for (not too small) finite samples. In the finite-sample simulations, the Sn is still more efficient than MADn, and that even for small n while the empirical bias curves showed that the explosion bias curve of S is nearly as good as that of the MAD, and for a contamination close to the breakdown point it is even better. For the implosion bias curve the MADn performs a bit better than Sn overall. Noting that the MADn is slightly more robust than Q_n regarding their explosion bias curves, whereas Q_n is more robust than MAD for implosion bias.

Although the following alternative estimators were presented in the case of an underlying Normal distribution, some changes occur when estimators are studied under the Cauchy and Exponential distribution. Hence, the consistency factors were readjusted. In a Cauchy distribution, the results show that the gross-error sensitivity for the Qn is 2.2214 asymptotically Therefore, the absolute asymptotic efficiencies become: e(MADn) = 81 percent, e(Sn) = 95 percent, and e(Qn) = 98 percent. But the Sn performs somewhat better than Qn at small samples. With regards to the exponential distribution, the influence function of the MAD looks very different as it is not a symmetric distribution. The gross-error sensitivity of Sn is smaller than that of the MADn. Other estimators are proposed such as the LMSn which in the case of normality has an influence function (IF) equivalent to the MADn [7], and its efficiency equals that of the MADn as well [1]. Thus meaning that LMSn is less efficient than both Sn or Qn (asymptotically and for finite samples), while their gross-error sensitivities and bias curves are almost as good. In conclusion, the choice between Sn and Qn comes down to a trade-off between efficiency and robustness. Qn is more efficient, while Sn is advantageous for its robustness and easier computation.

The following aim of this report would be to simulate the same study in a finite- sample setting. A comparison will be made between our report and the results of the paper. Furthermore, we will expand on an application of real dataset to test the similarities and limits to our methodology. Lastly some discussion will be made to conclude our main findings.

2.0 Q2.Finite-sample sensitivity curves of all 6 estimators

In this section, a random sample size n=100 from three different distributions is used to compare the sensitivity curves for each estimator.

2.0.1 Gaussian Distribution

The sensitivity curve for each estimator in the case of the Gaussian distribution depicts interesting aspects from figure 2.1 as follows: Firstly, following the paper, the Average Deviation does not show robustness and is unbounded. However, in a finite-sample the Average Deviation appears to be very efficient as its behavior is similar to that of the standard deviation. Secondly, the Qn estimator is bounded and has the second smallest gross error, but it is not significantly different from the gross error for Sn and BL, indicating strong robustness. Moreover, the sensitivity curve also shows a smooth curve, as the paper demonstrates through the influence function, and the sensitivity curve also confirms a high efficiency, illustrating the statement in the paper that mentions it is nearly 88%. Thirdly, the Sn estimator also shows adequate but lower efficiency than the Qn estimator according to the curve's smoothness, and their similarity to SD curve behavior mainly in the center. This also reflects Rousseeuw and Croux's statement, which indicates that Sn and Qn have 58% and 82% of normal efficiency, respectively. Therefore, this plot depicts the main argument in the paper, which is that Sn and Qn, could be better options than the MAD estimator. Besides, the BL estimator acts similarly to the Sn and Qn estimators, so it is better than the MAD according to the sensitivity curve in the Gaussian case, and matching with [6] as well.

Furthermore, the LMSn estimator is the one that depicts the smallest gross error among all estimators, but some efficiency is sacrificed as its behavior in the center of the curve, differs from the standard deviation. Moreover, the Average deviation's sensitivity curve does not illustrate boundedness, thus it is not a robust option.

Finally, although the MAD estimator appears to be robust due to its boundedness, it has a larger gross error than the rest, 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.

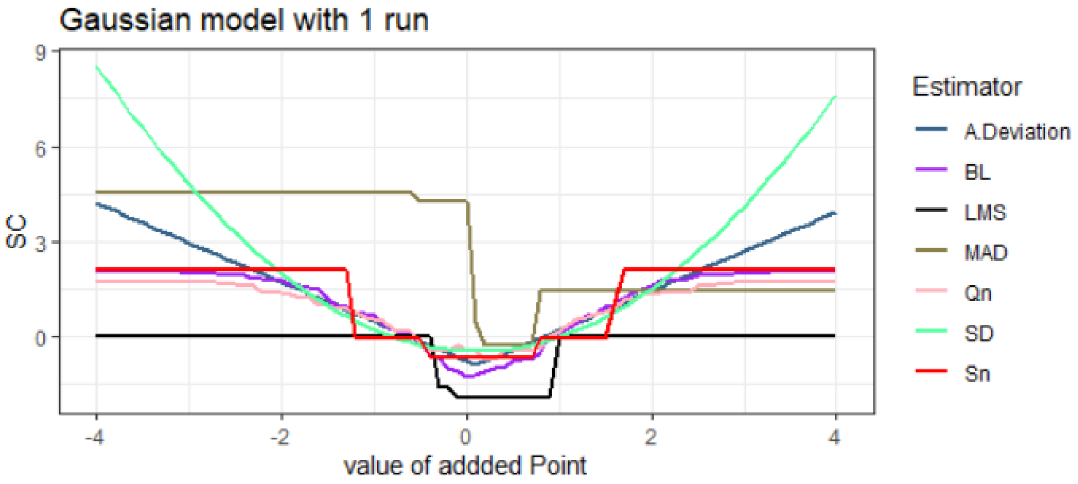

Figure 2.1: Sensitivity curve for the Gaussian Distribution - 1 Estimation

2.0.2 Cauchy Distribution

Contrary to Gaussian distribution, in a Cauchy distribution seen in figure 2.3, the Average Deviation estimator seems to be bounded with the lowest gross error. However, the interpretation of such result should be restrictive as Mean deviation theoretically has the breakdown point of 0 and an unbounded influence function [6]. In this case its bounded sensitivity curve is likely to be attributed to its large variance in the finite sample of 100 taking only one trial (run). In other words, the Average Deviation uses the location estimator-mean in its deviation equation, whose values vary significantly from sample to sample as the expected value of mean for the Cauchy distribution is undefined, and thereby, the sensitivity curve would not be constant throughout a series of runs.

Similarly to their behavior in Gaussian distribution, the MADn, Qn, and Sn estimators show their robustness by their bounded sensitivity curve. Although Sn appears to be most robust estimator with its lowest gross error, its influence function indicates differently that it should be less robust than MADn, which is compensated by the increase in Sn's efficiency over MAD.

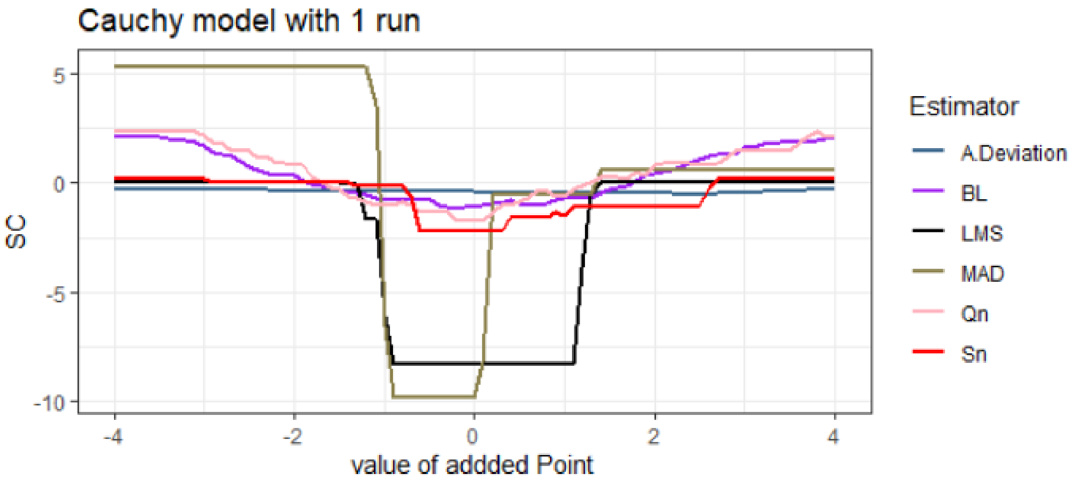

Figure 2.2: Sensitivity curve for the Cauchy Distribution - 1 Estimation

2.0.3 Exponential Distribution

The figure 2.3 presents the sensitivity curves with one run for an Exponential distribution. The sensitivity curves observed of all estimators, except for SD and Mean deviation, seem to be bounded in the graph. MAD has the lowest gross error when the outliers are negative, and LMSn appears to be the most robust when they are positive. Also, BL has a relatively high gross error in contrast to the previous cases, because its bound is the highest in the plot, and it also shows a pronounced peak at the center of the curve which indicates that its efficiency is also lower compared to the results in the Gaussian and Cauchy distributions. Besides, the Qn and Sn curves share a similar shape with Qn having a slightly higher gross error than that of Sn. The efficiency is also lower than it was in the previous cases because all estimators also show sharp peaks at the center of the graph, and Sn is the worst one in this particular case according to the plot.

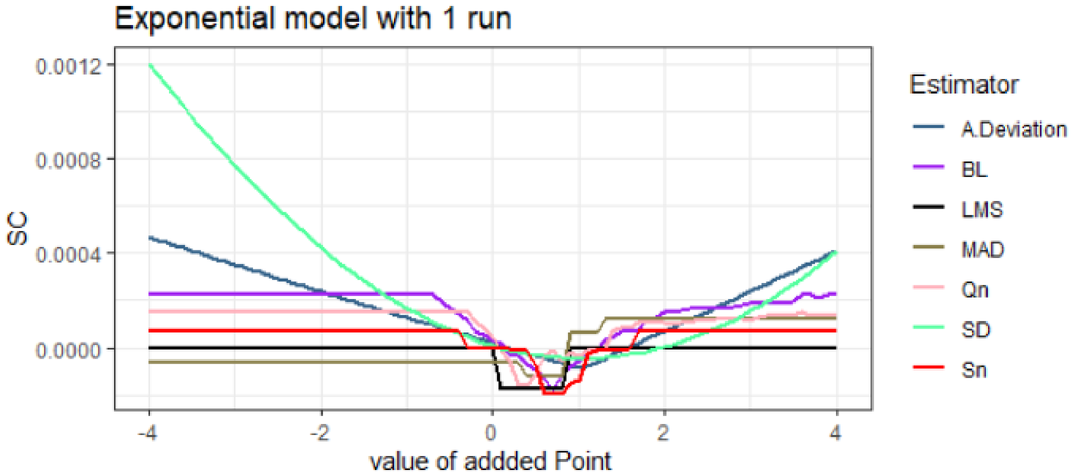

Figure 2.3: Sensitivity curve for the Exponential Distribution - 1 Estimation

3.0 Q3. Obtaining smoother curves using 500 runs

To obtain smoother curves, the derivation of the estimators was repeated 500 computation times (m = 500) and then the curves were calculated by averaging the results.

3.0.1 Gaussian Distribution using 500 runs

Averaging 500 runs significantly improved the smoothness in the curves as it centered the estimation of the estimators to the real value based on the 500 replications. In other words, averaging the results tends to show what the true value of the estimation is, and thereby reflects smoother curves. Also, the plot 3.1 showed some significant changes as follows: The BL estimator, in this case, is more efficient than the Sn estimator, in contrast with the plot for 1 run, as its curvature is closer to the SD curve, and the curve smoothness has a well-defined U-shape. Besides, the MAD estimator depicted a significant change such as the shape of the curve, because it definitively went smoother. Likewise, its efficiency improved drastically, and the gross error related to the boundaries showed a sharp decrease. Furthermore, the Sn decreased in its efficiency but continued to be a highly robust estimator as its boundaries follow those of the LMSn estimator, which seems to be the most robust among them. Following this, the LMSn estimator improved its shape but its efficiency remained constant. Finally, the Qn, Average Deviation, and Standard Deviation present almost the same behavior they had only with one estimation.

Overall, the curves do seem to be smoother, some estimators showed to perform better after averaging an approximately large number of experiments, but some of them just remained constant or slightly decreased the efficiency as is the case of Sn.

3.0.2 Cauchy Distribution using 500 runs

For the Cauchy distribution as well as the normal distribution the MAD estimator presented a dramatic change, as its curve is significantly smoother, its boundaries are symmetric and based on the other estimators, its efficiency and gross error appeared to have improved. Although the Sn estimator's curve is way smoother, its efficiency seems to have decreased and its gross error to have increased. Besides, the LMSn

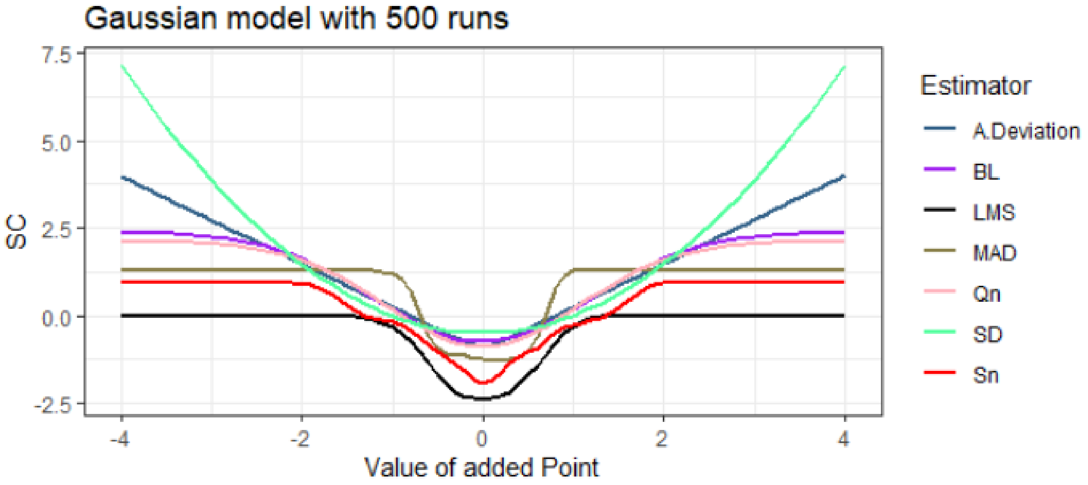

Figure 3.1: Sensitivity curve for the Gaussian Distribution - 500 Estimations

estimator remains constant but its curve smoothness improved as it is the case for the Average Deviation. Finally, the Qn and BLn estimators show a particular behavior as they do not seem to be bounded and its gross error increased significantly. Therefore, these estimators presented a lower performance after averaging 500 results.

Hence, in this case, the curves are also smoother, the MAD estimator also improved its performance, Sn is also less efficient, and Qn and BLn changed their boundaries and robustness in comparison to one run.

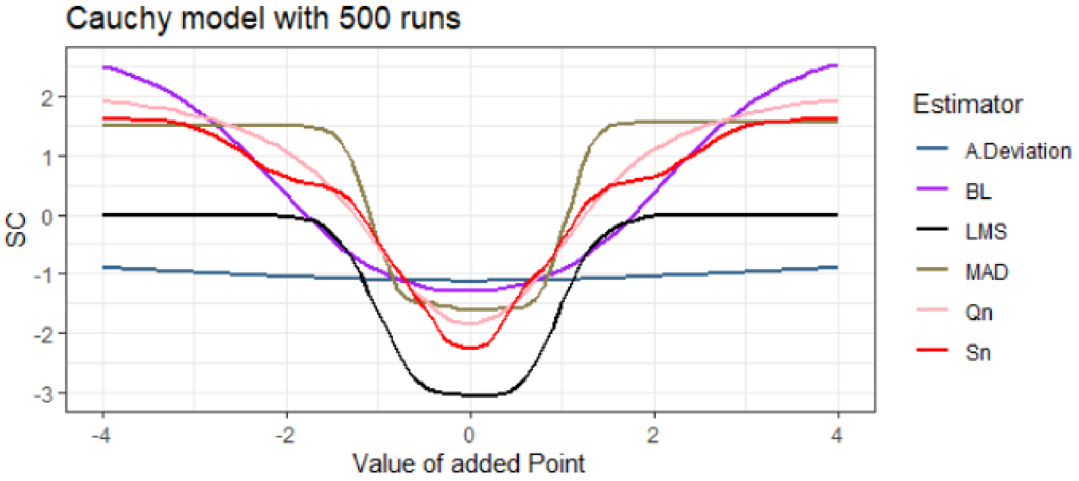

Figure 3.2: Sensitivity curve for the Cauchy Distribution - 500 Estimations

3.0.3 Exponential Distribution using 500 runs

Taking the average of 500 trails in figure 3.3, we expect to see a smooth approximation of the robust estimator's influence functions in their corresponding sensitivity curves. This generally holds true in terms of the the estimators' robustness and efficiency, despite few inconsistencies. When identifying the most robust estimator, we found that although MAD has the lowest gross error for negative outliers, which is also seen on its influence function when compared to Sn and Qn [6],for positive outliers the LMSn estimator overtakes MAD as the most robust. However, in the paper, LMSn is suggested to have the same influence function as of MAD. Indeed, as pointed by the paper, in asymmetric distributions such as the exponential distributions, LMSn might have an advantage over MAD due to the later' use on location estimator. We have also observed that the robust BL estimator has shown slightly higher efficiency but the lowest robustness when compared to other robust estimators.

From the graph the trade-off between efficiency and robustness is evident; the less robust estimators the higher the efficiency such as the case of LMSn which tends to show strong robustness but lower efficiency. For the estimators with high efficiency and thus low variance, the value of the estimates should approach approximately to the true value of the scale parameter in the graph when the data is uncontaminated(corresponding to the central region around 0). whilst when the data is uncontaminated, the efficiency of the robust estimators is always lower than that of their non-robust counterparts. This is, however,compensated by their bounded values of gross error when data is contaminated[4].

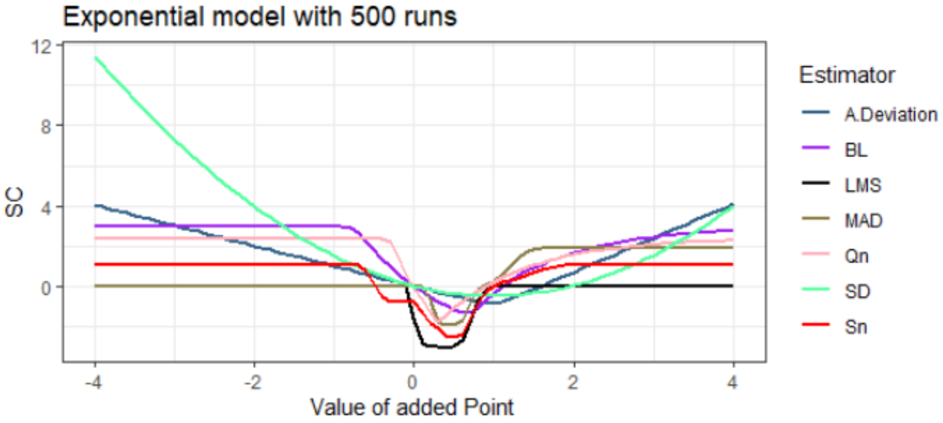

Figure 3.3: Sensitivity curve for the Exponential distribution - 500 Estimations

4.0 Q4. Simulations on Gaussian and Cauchy distributed data

4.0.1 Gaussian Data

In order to create table 4.1 we have taken averages and average standardized variations over 5000 trials of each of the six estimators with increasing sample size. The average standardized variance was calculated as mentioned in eq. 2.10 in [6],

$$\text{STDVAR }(T_n) = n * \text{var}_m(T_n)/(\text{ave}_m(T_n))^2. \tag{4.1}$$

where denominator given in equation 4.1 is given in order to provide a more natural measurement of the variability of the estimator for a scale parameter by rescaling according to the estimated scale parameter of the underlying distribution. By repeating estimations with an increasing number of samples(n) we can glean the asymptotic relative efficiencies of these estimators by taking the ratios of their inverse standardized variances at large (∞) values of n. In each of the estimators we have employed consistency factors as illustrated in [6]. While there is slight variation in these results from those in the reference paper, we can confirm that, in addition to having higher precision, both S_n and Q_n converge to the true value more

	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

quickly than MAD_n. We can also confirm some results from [6] on the relative trad-offs of S_n and Q_n, While we see that Q_n is more efficient than either S_n or MAD_n, we also can note its small sample size bias. While we do not directly confirm results concerning the relative efficiency of the mean average deviation, with Gaussian efficiency of approximately 88% [8], we do note that it shows the highest precision. This is likely because we are making scale estimates using uncorrupted data drawn from a standard Gaussian distribution.

We can also compare S_n, Q_n, MAD_n and $meanAD_n$ to other traditionally robust estimators from other sources. Notably BL_n and LMS_n. On uncorrupted Gaussian data, BL_n only second most efficient to the mean average deviation$_n$ and is shown to be unbiased with the correct consistency factor for Gaussian data $(c = \frac{\sqrt{2}}{\Phi(0.75)} \approx 1.0483)$. It is interesting to note that the Bickell-Lehmenn estimator, with a breakdown value of 29% and an efficiency of about 86% [6], was used as a form of inspiration for the more robust Q_n. We can see that it has a similar, although slightly smaller efficiency than the mean average deviation, but a much higher efficiency than the more robust Q_n, highlighting a trade-off between robustness and efficiency. We can also investigate the relative properties of the LMS_n estimator, which is second least efficient on standard Gaussian data only to MAD_n. While it has a 50% breakdown value [6], we can remark that it shows a noticeable small sample size bias similar in magnitude to that of Q_n but of opposite direction.

4.0.2 Cauchy Data

We can repeat the above experiment and perform a similar analysis on data generated by a standard Cauchy distribution as well, in order to examine the estimator's robustness against departures from normality. Again we have run simulations with increasing numbers of data points in order to assess the small sample size performance of each of these six estimators and hopefully glean some numerical insight into their asymptotic behavior. We have also used consistency factors mentioned in [6] for S_n, MAD_n, and Q_n in addition to own derived factors which works for both BL_n and LMS_n of $c = \frac{1}{Q_{0,2}(0.75)} = \frac{1}{2}$ where $Q_{0,2}$ is the quantile function of a Cauchy distribution with zero location parameter and scale parameter of 2. No factor was found for the mean AD as the Cauchy distribution has no defined mean or variance [5]. Our data confirms results highlighted in [6]: the continued small sample size bias in Q_n, that Q_n has relatively

	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

higher efficiency than S_n and MAD_n, and that both Q_n and S_n are more efficient than MAD_n.

We are now able to compare results on Q_n, MAD_n, and S_n to a broader range of estimators, namely LMS_n, BL_n and the mean average deviation. First and foremost, the mean average deviation entirely breaks down under this non-normal data. Critically, it relies on values of the mean in order to produce estimates of scale, however the mean of a Cauchy distribution is not defined [5] and this estimator cannot produce good results. Furthermore, as illustrated in [5], as more samples of data are taken from a standard Cauchy distribution, more samples are taken from the tails of the distribution, inflating the value of the variance of the sample. Since the mean average deviation is calculated using the mean of a Cauchy distribution, it is straightforward to see that it should not converge in a meaningful way, which can be seen in table 4.2.

We can also see the change in behavior of BL_n and LMS_n. While BL_n remains fisher consistent, it loses quite a bit of efficiency in this case, especially at small sample sizes. Notably BL_n is the least efficient of the convergent estimators. LMS_n, while more efficient than the MAD_n estimator at large sample sizes also maintains a small bias at lower sample sizes.

By comparing the results from table 4.1 with those of table 4.2, we can see overall that the more robust estimators (S_n, Q_n, MAD_nandLMS_n), each with a 50% breakdown, are now *more* efficient than several other candidates under departures from normality. We also see how changes in distribution can severely hamper estimation using the mean, and the overall loss of efficiency when the underlying data generating distribution is non-normal.

5.0 Q5. Studying the effects of asymmetric corrupt data

5.0.1 Incremental Gaussian Corruption

We generated figure 5.1 using ϵ proportion of corrupted data drawn from a $\mathcal{N}(a, 1)$ distribution, with a set to 10. This has the interpretation of either removing an ϵ proportion of good samples and replacing them with improper ones from another distribution or shifting an ϵ proportion of the data to the right. As more and more corrupted data is added, the estimator can be moved further away from the true scale

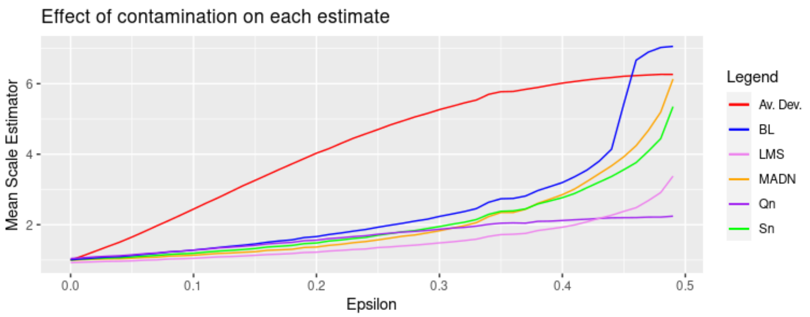

Figure 5.1: As Gaussian contamination percentage increases we start to see breakdown of the estimators

value of 1. Since the corruption added is $\mathcal{N}(a, 1)$, if the corruption were allowed to continue, the data would all have been shifted to the right and the estimators would reconverge on a value of 1 centered at an arbitrary point, a.

From the graph we can make a qualitative comparison of the robustness of the various estimators. It is desirable that a more robust estimator would be resistant to the introduction of corrupted data, so the estimators whose value changes most rapidly are least robust. We can see that the mean average deviation is the least robust as evidenced by the rapid increase in value of the red line in figure 5.1. Then we see that the BL_n estimator is fairly robust up to a point, but not as robust as LMS_n, Q_n, S_n, and MAD_n. This behavior corresponds well to the breakdown value of 29% versus the 50% breakdown value of the other estimators. It is also worth noting that after its breakdown value of 29% the BL_n estimator actually is more affected by the data than the almost entirely un-robust mean average deviation as evidenced by the rapid increase and high value of the mean scale estimator at $\approx 46\%$ corruption. LMS_n, MAD_n, Q_n and S_n all behave similarly, showing how strongly related they are. While here we see that MAD_n is actually least robust in the context of unbalanced Gaussian corruption, however it performs similarly to S_n. Here Q_n and LMS_n perform with the least change in value, and it is worth noticing that they 'switch' near 43% where the LMSn estimator is pulled more violently by the higher amounts of corruption.

5.0.2 The effect of the magnitude of deviation

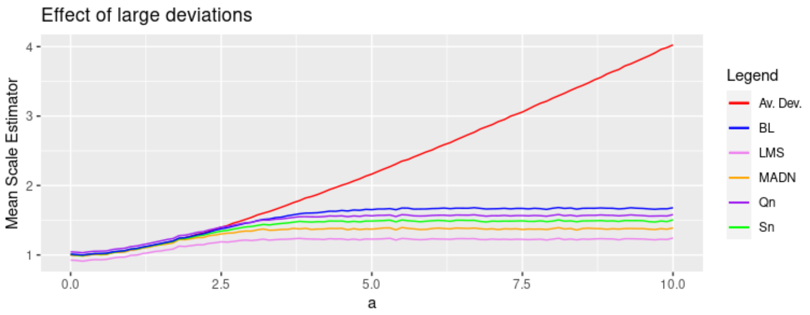

Figure 5.2: All estimators besides the average deviation stay bounded when a moderate amount of contamination is added

In figure 5.2, for a moderate amount of corruption we can see how the estimators change in value. Particularly, since $\epsilon = 20\% \leq 29\%$ (the breakdown value of the BL_n estimator) 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. Since in figure 5.2 we are comparing the value of the scale estimates with a *percentage* of corrupt data as opposed to a single changing data point we see slightly different behavior of the estimator's bounded values, although qualitatively it remains a similar picture.

6.0 Q6. Studying the effects of symmetric corrupt data

The results of question 5 were replicated, changing the conditions upon which using the ϵ proportion of corrupted data is spread evenly over $N(a,1)$ and $N(-a,1)$ distribution.

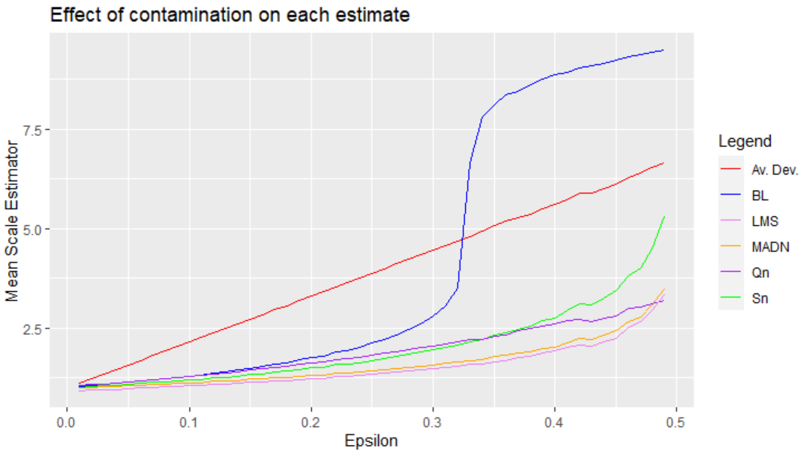

Figure 6.1

In figure 6.1 we see the effect of the varying contaminating (spread evenly) proportion of the data on each estimator. As mentioned previously, a breakdown point is seen when the estimator become highly nonlinear.

The earliest departure of the mean scale by a positive proportion of corrupted data is the average deviation, which is the same as in the case of non symmetric corrupt data (Q5). This testifies for the lack of robustness. Regarding the BLn estimator, it increases abruptly at approx. 32 percent of contamination, which surpasses the average deviation. That happens much earlier than in Q5, that was at approx. 45 percent. It would mean that the BLn is less robust earlier passed an increasing proportions of outliers when the contamination is symmetric. Looking at the LMSn, the MADn and the Sn, we can observe again similar behaviors. Their breakdown values are verified to 50 percent, the increase seems less pronounced than in Q5 but follows closely the MADn. The LMSn and MADn seem closer together in terms of behavior than in a non symmetric case. While the Sn seems to have the same shape and curvature, as MADn and LMSn, however, the Sn estimator is shifted to upper side, where in Q5 it followed more closely MADn. Lastly, the Qn remains seemingly the most stable for most of the increase in σ, with some slight fluctuations starting from approx. 32.5 percent. Approaching 50 percent of outliers, it is the lowest among other estimators, which testifies to its robustness. 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. The Sn and Average deviation are somewhat in between. We can conclude that with an symmetrical distribution of corrupt data, behaviors in robustness of the estimators can vary like the MADn. While at the limit of 50 percent corrupt data, the Qn remained the most robust in both cases, the BLn remains the least robust.

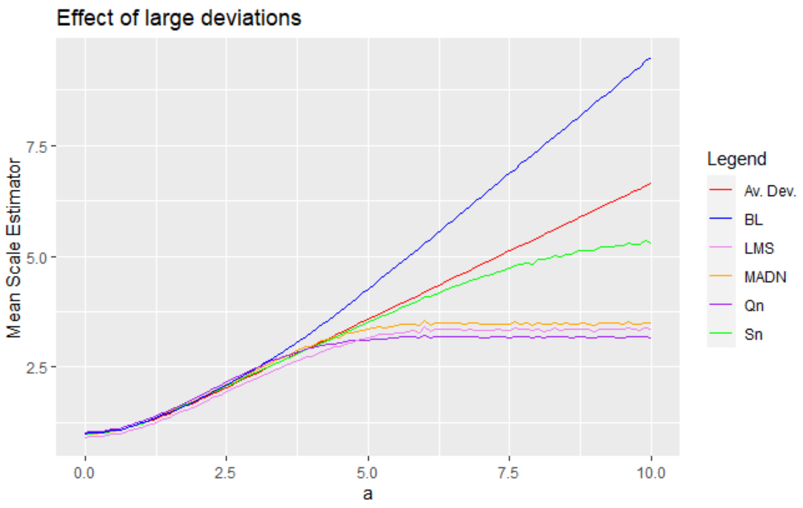

Figure 6.2: large symmetric deviations cause estimates to diverge

In 6.2 the moderate amount of corruption (a=10) is visualized over those same estimators. Here only the MADn, LMSn, and Qn are bounded. The three other estimators BLn, Average Deviation, and Sn remain divergent. The estimators that are less sensitive to shift only do so around $a = 5$ on-wards, while it was seems to happen earlier on in Q5. 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.

7.0 Q7. Real dataset, Pima Indians

In this section, a data set based on patients that may or may not present diabetes is analyzed, and the goal is to compare the performance of the aforementioned different estimators. The aim is to analyze if the robust estimators behave differently, ihow the results depend on the data set, and if the estimators' behavior is consistent with the results presented in questions 5 and 6, and the introductory paper [6].

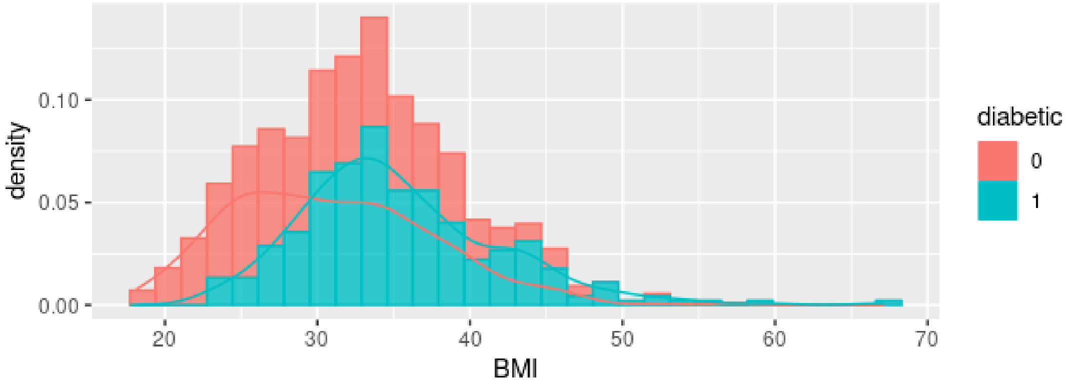

Figure 7.1: Two different populations are contained in the variable BMI in the Pima indian diabetes dataset. Zero values, which correspond to missingness have been removed

The Pima Indians diabetes dataset, originally produced by the National Institute of Diabetes and Digestive and Kidney Diseases, contains measurements taken of a selection of individuals: females at least 21 years old of Pima Indian heritage. All diabetics in this database have type II diabetes. While several diagnostics were taken, one which is known to correlate with type II diabetes is BMI, making it a good choice for study of outlying behavior. As a point of note, this dataset contains several zeros, which we will not consider outliers, as the zero values correspond to missingness. The variable BMI, plotted in figure 7.1, shows two populations present those with and those without type II diabetes. Since this histogram is similar to a 'contaminated normal' example as seen previously, it makes a good ideal object for study.

In order to assess the performance of the six scale estimators we first must verify that we are using the correct consistency factors. Using a similar technique as that employed in question 4, we can first standardize our data and then calculate the average of a subsample of size 640 with 1000 repetitions in order to estimate the correct scale factor. The values that we obtained were 1.267, 1.494, 1.201, 1.069, 2.256, and 0.770 for the average deviation, MAD, S_n, BL, Q_n, and LMS_n respectively. Since these values are extremely close to the consistency factors for Gaussian distributed data, coupled with a visual inspection of the distribution in figure 7.1, *we continued the analysis using the defined consistency factors for Gaussian distributed data*.

We can then use the estimated scale parameters to check for and remove outliers using the following rule:

$$z_i = \frac{x_i - \mu(X)}{\sigma(X)} \geq 3, \tag{7.1}$$

which we have borrowed from [3]. In the above equation, μ is an estimator of location, which we have chosen to be the median, and σ is a scale estimate, which we vary in order to assess the performance of each individual scale estimator. It is desirable since a robust estimator of scale would be able to 'unmask' outliers that were previously hidden. Using the above rule using each estimate of scale in turn as the denominator in eq. 7.1 we can classify a small number of outliers, both graphically as seen in figure 7.2, and numerically.

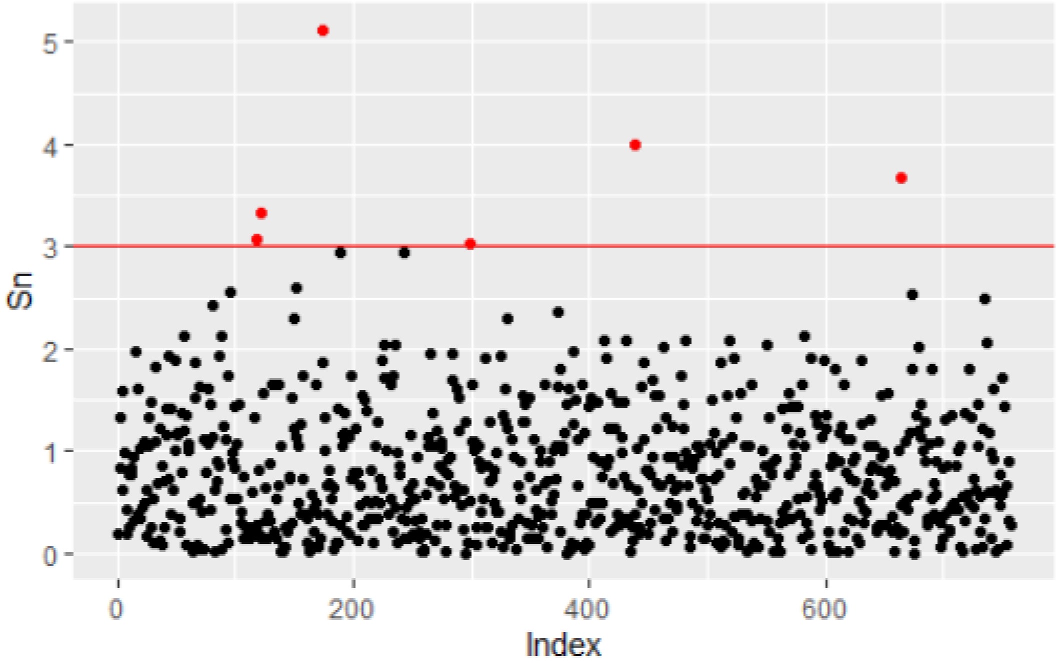

Figure 7.2: Sensitivity curve for the Exponential distribution - 500 Estimations

	mean AD	MAD_n	S_n	BL_n	Q_n	LMS_n	SD_n
Number of Outliers Detected	6	6	6	6	5	6	5

Table 7.1: All scale estimators select observations 117, 122, 173, 299, 439, and 665 except Q_n and S_n which leave observation 299 masked

Following this, and according to table 7.1 nearly all estimators produced the same six outliers but Qn and Sd because they detected only five. Indeed, five out of six outliers correspond to diabetic patients, classified as one (1) by the variable Outcome, but sample 665 corresponded to a patient without diabetes marked as zero (0). Besides, the variable BMI represented the body mass index (weight in kg/(height in m)^2) which ranged between 18 and 58 and using the z-scores method mentioned in eq. 7.1, those values between 52.9 and 58 were selected as outliers. The average BMI corresponds to 32.45, and there were 96 values higher than 40, and 6 values between 49 and 52. Therefore, it is clear that the distance between outliers and regular observations is not significantly far or separated. This indicates, that the robust estimators are not expected to perform notoriously differently than the SD and the MeanAV as the extreme cases are no extreme compared to the regular data.

After removing the above outlying values we are now in a position to compare the performance of our set of estimators on the clean and unclean data:

	mean average deviation	$MADN$	S_n	BL	Q_n	LMS	SD
mean AD	6.645356	6.6717	6.79782	6.70912	6.65742	6.59757	6.565042
MAD	6.645356	6.6717	6.79782	6.70912	6.65742	6.59757	6.565042
Sn	6.645356	6.6717	6.79782	6.70912	6.65742	6.59757	6.565042
BL	6.645356	6.6717	6.79782	6.70912	6.65742	6.59757	6.565042
Qn	6.671014	6.74583	6.79782	6.70912	6.65742	6.59757	6.603713
LMS	6.645356	6.6717	6.79782	6.70912	6.65742	6.59757	6.565042
SD	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

Comparing $MADN$, S_n, BL, LMS, and Q_n to the mean average deviation and SD, we actually see little difference in the change in scale estimate with the largest difference being a 5% change in the standard deviation. The reason for this might be associated with the contamination level in the data set, which can be compared with those results from figure 5.1. The number of outliers detected (6) corresponded to 0.8% of the data set, and this indicates that the contamination was minimal. As shown in question 5, when the contamination levels are too small the estimators tend to attain the same or quite a close value, and at this point 5.1 demonstrated that all estimates approximately reached the same value. Moreover, it can be mentioned that the outlying values are all close to the central distribution. Taking inspiration from question 5, particularly 5.2. We can view adding diabetic persons to the database as shifting a healthy patient's BMI to the right, changing the scale estimate. However, for small changes, all estimators act approximately the same and the boundedness of the more robust estimators does not come into effect.

8.0 Q8. Discussion

8.0.1 Discussing estimators

Comparison of Estimators

	Scale Estimators (Gaussian)					
Criteria	$MeanAD_n$	MAD_n	S_n	BL_n	Q_n	LMS_n
Boundedness	Unbounded	Bounded	Bounded	Bounded	Bounded	Bounded
Efficiency	86	37	58.21	86	82	37
Gross Error (rank)	1	4	5	2	3	6
	Scale Estimators (Cauchy)					
Criteria	$MeanAD_n$	MAD_n	S_n	BL_n	Q_n	LMS_n
Boundedness	Unbounded	Bounded	Bounded	Bounded	Bounded	Bounded
Gross Error rank	6	4	3	1	2	5
Efficiency	0	81	95	74	98	81
	Scale Estimators (Exponential)					
Criteria	$MeanAD_n$	MAD_n	S_n	BL_n	Q_n	LMS_n
Boundedness	Unbounded	Bounded	Bounded	Bounded	Bounded	Bounded
Rel. Efficiency w/ MAD_n	1.429	1	1.1214	1.452	1.453	1.054
Gross Error rank	1	4	5	2	3	6

Table 8.1: Estimators comparison in a Exponential distribution; Relative efficiencies were computed in the exponential case as there is no fisher information for the exponential distribution [6]

In general, we see a very similar pattern of the behaviours for all our robust estimators across three distributions, despite minor differences in their efficiency. Consistent with the paper,MAD_n's gross error is the lowest amongst Q_n and S_n in Cauchy and Exponential distribution. In the Gaussian case, although we do see its gross error higher than that of S_n, this can be justified by its non-trivial variance; Because of the variances of these estimators, we may observe the results different from their influence functions given.

We have also observed that BL_n's sensitivity curve has the similar shape of Q_n. The justification is given that they both take the statistics of the pairwise inter-point distances, although Q_n uses the kth order statistics[6].

It is surprising that,despite its breakdown value of 0, the Mean-ad estimator appears to have the bounded lowest gross error in Cauchy distribution. The fact that we are not able to find its consistency factor in Cauchy is potentially playing a role here. Mean-ad, by the definition, is not a Fisher consistent scale estimator for Cauchy distributed data. In fact, when we plot it separately, not only the curve itself is unbounded, but also its position shows the significant deviation from the rest of the estimators

It is also worth noting their relative computational efficiencies. $MeanAD_n$ takes the mean, subtracts values, and takes the mean again, all of which are linear in n so it is one of the faster options for

$MeanAD_n$	MAD_n	S_n	BL_n	Q_n	LMS_n
O(n)	O($n \log n$)	O($n \log n$)	O(n^2)	O($n \log n$)	O($n \log n$)

Table 8.2: Computational time requirements for each estimator. Space requirements can depend on sorting algorithm.

estimating scale. Several of the other estimators take at least $n \log n$ time to compute, making them tenable options for larger datasets. BL_n however, with a computational complexity of n^2 is unfeasible for a large dataset.

On the Trade-off between Robustness and Efficiency

In several cases we see a noted trade-off between robustness and efficiency for these estimators of scale. This can be seen plainly by the high standardized variances of the more robust MAD_n, S_n, BL_n, Q_n and LMS_n on Gaussian data in 4.1. Of these, BL_n is the most efficient, coupled with the lowest breakdown value at 29% as opposed to 50%. While the breakdown value is rarely model dependent [2], The relation to the efficiency is clearly model dependent, however, as the standardized variances qualitatively change when estimating the scale of Cauchy distributed data.

The above behavior can be explained by examining the sensitivity curves seen in part 2. Since, for large n, the sensitivity curve can be seen as being approximately equal to the influence function (more on that below), a bounded sensitivity curve corresponds to a bounded influence function, which is model dependent. An influence function with a low gross error sensitivity is, by definition, one with a low least upper bound on the magnitude of its influence function, a desirable quantity. There is, however, a conflict between a low gross error sensitivity and a low asymptotic variance [2], meaning that estimators with a low gross error sensitivity will generally have higher variances and lower efficiencies. In general, a high breakdown value cannot be reached while also minimizing the gross error sensitivity [2]. We see this in the bounded behavior of the sensitivity curves for the more robust estimators in figures 3.1, 3.2, and **??** of question 3, the estimators with bounded influence functions have lower efficiencies in part four than the estimators with unbounded influence functions. They are also more robust.

8.0.2 Discussing Methodological link to paper

Link between IF and sensitivity curves

Rousseeuw and Croux presented their results asymptotically by visualizing and computing Influence functions of the various estimators across three different distributions in order to see the effect of a small number of outliers on their behaviour. Coupled with the bias curves that would enable to view the estimators with an increasing proportion of outliers in dataset. However, our report bases itself on finite-sample simulations. It is therefore relevant to further explore the link between the two, and how for the notions of Influence curves are translated into sensitivity curves. As a reminder we have the following expressions:

- Sensitivity Curve:

$$S(x, T_n, X_{n-1}) = (T(X_{n-1}, x) - T(X_{n-1}))n \tag{8.1}$$

- Influence Function:

$$IF(x,T,F) = \lim_{\epsilon \to 0} \frac{T(F_{\epsilon,x}) - T(F)}{\epsilon} = \frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=0} T(F_{\epsilon,x}) \tag{8.2}$$

As seen above the sensitivity curve evaluates the difference of the the estimator has with and without the presence of an outliers, function of the sample size n. While the Influence function on an asymptotic difference of the estimators with and without the presence of a proportion of contaminated data ϵ. Those estimators now depend on an underlying distribution F and are corrected for the proportion ϵ (numerator). In more details, the limit is calculated from from the right up to the point x_o equal to ϵ. Intuitively, if we added an added a new observation x_o to the sample x1, . . . , xn the fraction of contamination is 1(n + 1), and so we define the standardized sensitivity curve (SC) as:

$$SC_n(x_0) = (n+1)((T_{n+1}(x1, ..., xn, x0)) - T_n(x1, ..., xn))) \tag{8.3}$$

With the assumption of x_i being iid., then for a large n, then $SC_n(x_0)$ approximately equals $IF(x_0, F)$. A possible limit worth mentioning is that while the convergence from SC to IF depends on n, which in itself depends on x_o, then there are cases of no uniformity. This is the furthered in [4].

8.0.3 Conclusions

Robust Estimation is Not Always Preferable

As we have seen in question 7, and throughout the exercises, there is a trade-off apparent in robust estimation in terms of efficiency. In the case of certain estimators, notably Q_n, there is also more potential for a bias which can be introduced by the more robust procedure. A key take-away is that if there is good reason to believe that the underlying distribution is normal with few, or minor, outliers it is actually more preferential to use a less robust estimator as one can get more powerful estimates with less data than with their robust alternatives.

Robust estimation can also be beneficial under model mis-specification

As has been seen in questions 3, 4, and 5, when the model is misspecified or corrupted the bounded nature of these univariate scale estimators limits the effects of large deviations from the assumed underlying model. This can be invaluable in cases where we cannot be sure of the distribution of our data. One example of this is in stock returns, which are assumed to be log-normal in the Black-Scholes-Merton model. They are, however, famously heavy tailed and estimating their properties using the assumption of log-normally distributed data can lead to gross errors.

Robustness comes at a cost to computation time

We see in 8.2 that the more robust estimation procedures come at a cost in terms of computation complexity. While having a high breakdown value comes at the cost of efficiency [2], it also comes at the cost of an increase in computation time. We can see that the robust estimators of scale MAD_n, S_n, Q_n, BL_n, and LMS_n require one to sort the dataset, which often takes $n \log n$ time and can require auxiliary space. This step is a necessary part of identifying 'idealized' points with which to generate the scale estimate, however identifying and selecting these points comes with a corresponding increase in computation complexity which must be taken into account.

.1 Derivations of Scale Factors

.1.1 BL_n

In order to derive the consistency factor for the BL_n estimator, we follow similar steps as to that of the Q_n estimator. We first define

$$BL(F) = cH_F^{-1}(1/2) \quad \text{with} \quad H_F = \mathcal{L}(|X-Y|)$$

This is different than $BL(F_n)$ but in the asymptotic limit they are equivalent.

We first find $\mathcal{L}(|X-Y|)$.
In the case of a cauchy distribution, the variable $Z = X - Y$ is distributed as a cauchy random variable with sale factor 2. This can be seen using an identity of the characteristic function with $X_1, ..., X_n$ i.i.d.:

$$\phi_{a_1X_1+...+a_nX_n}(t) = \phi_{X_1}(a_1t)\phi_{X_2}(a_2t)...\phi_{X_n}(a_nt)$$

Plugging in $Z = X - Y$ we find that

$$\phi_{X-Y} = \phi_X(t)\phi_Y(-t)$$

The characteristic function for a cauchy distributed random variable is as follows:

$$\phi_X(t) = ex_0it - \gamma|t| \tag{4}$$

With location parameter x_0 and scale parameter γ. The characteristic function for the difference in cauchy random variables with zero location parameter and scale parameter of 1 then becomes:

$$\phi_{X-Y} = e^{-2|t|}$$

i.e. a Cauchy distributed random variable with location parameter 0 and scale parameter 2.
To find the distribution of the absolute value of this cauchy distributed random variable, we first remember that the absolute value is not differentiable at the origin and so standard formulas for the transformation of variables do not apply. We must instead go directly to the CDF.

$$F(|Z| \leq z) = F(Z \leq z) - F(Z \leq -z)$$

We can then again exploit the symmetry of the Cauchy distribution: $F(Z \leq -z) = 1 - F(Z \leq z)$ and the above equation becomes:

$$F(|Z| \leq z) = 2F(Z \leq z) - 1$$

Since we take the *median* of these values, it corresponds to the point at which $F(|Z| \leq z) = 1/2$, setting the above equation and solving for $F(z)$ we see:

$$\begin{aligned} F(Z \leq z) &= 3/4 \\ z &= F_Z^{-1}(3/4) \end{aligned}$$

As a sanity check, if we allow Z to be a normally distributed random variable (again with variance of 2) we see that $z = 0.9538726$, with the correct inverse and scale factor $c = 1.048358$. With Z a Cauchy distributed random variable with scale parameter 2 this leads to a value of z = 2 and so the inverse leads to a scale factor of 1/2.

LMS_n

In the case of the LMS_n estimator, we are trying to find the 'shorth' length, the minimum length which contains half of the data. It can be defined for arbitrary coverage $\alpha \in (0, 1)$ as:

$$S_\alpha(x) = \min\{|I|\colon I = [a, b], x \in I, P(I) \geq \alpha\}.$$

In the case of a symmetric unimodal distribution, the interval $[a, b]$ must be centered about the location parameter μ as $[\mu - c, \mu + c]$ for some $c \geq 0$.

$$\begin{aligned}
\inf_c P(x \leq \mu + c) - P(x \leq \mu - c) &\geq 1/2 \\
\inf_c P(\frac{x-\mu}{\theta} \leq \frac{c}{\theta}) - P(\frac{x-\mu}{\theta} \leq \frac{-c}{\theta}) &\geq 1/2 \\
\inf_c F_{\mu,\theta}(\frac{c}{\theta}) - F_{\mu,\theta}(\frac{-c}{\theta}) &\geq 1/2 \\
\inf_c 2 * F_{\mu,\theta}(\frac{c}{\theta}) - 1 &\geq 1/2 \\
\inf_c c &\geq \theta F^{-1}_{\mu,\theta}(3/4) \\
c &= \theta F^{-1}_{\mu,\theta}(3/4)
\end{aligned}$$

Clearly the value of c is the third quartile of the data!, So the lower value $-c$ is the first quartile. The LMS_n estimator is the collection of half samples of data ordered by index. This means that the asymptotic consistency factor should be the same as that of the IQR. In the case of the normal distribution, this becomes:

$$\frac{1}{c} = 2\Phi^{-1}(0.75)$$

and in the case of a cauchy distributed R.V. this should be:

$$\frac{1}{c} = 2\gamma \tag{5}$$

with γ being the value of the scale parameter of the standard cauchy distribution, $\gamma = 1$

Bibliography

[1] R Grubel et al. “The length of the shorth”. In: *The Annals of Statistics* 16.2 (1988), pp. 619–628.

[2] F Hampel et al. “Robust Statistics: The Approach Based On Influence Functions”. In: *Wiley-Interscience* 502 (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: *Journal of the American Statistical Association* 88 (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 Neerlandica* 42.2 (1988), pp. 103–116.

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