

Monte Carlo Analysis of Univariate Statistical Outlier Techniques

Mark W. Lukens

This paper examines three techniques for univariate outlier identification: Extreme Studentized Deviate (ESD), the Hampel identifier and the Rousseeuw identifier. The purpose is to determine how these outlier identification techniques perform under varying conditions. An experimental design along with two different Monte Carlo simulations provides insights into the problem. Under certain assumptions it is shown that the ESD identifier performs well with very small data contamination and the robust Hampel and Rousseeuw identifiers perform better with large samples and with multiple outliers.

1. Introduction

In the article, “The Identification of Multiple Outliers”, JASA, September 1993, Laurie Davies and Ursula Gather examine the outlier identification problem. Their approach was to develop an outlier-generating model that allows a small number of observations from a random sample to come from a different distribution than the target distribution. This contamination is then considered as outliers.

The target distribution is assumed to be a normal distribution through the course of the article. The outlier region for $N(\mu, \sigma^2)$ is defined by:

$$\text{out}(\alpha, \mu, \sigma^2) = \{x : |x - \mu| > z_{1-\alpha/2} \sigma\}$$

A number x is called an outlier with respect to $N(\mu, \sigma^2)$ if

$$x \in \text{out}(\alpha, \mu, \sigma^2)$$

The authors then explain and formulate the outlier identification problem and explain three robust outlier identification techniques using the above methodology. The techniques are: Extreme Studentized Deviate – Extreme Deviate Removal (ESD-EDR), the Hampel identifier and the Rousseeuw identifier. A Monte Carlo simulation is used to determine

which of these techniques performs the best. This paper will attempt to confirm the results of, but not replicate, the Monte Carlo experiment using a slightly different experimental design and approach.

2. Outlier Identification

Given the above definition for the outlier region an example graph of this region for the standard normal is shown below.

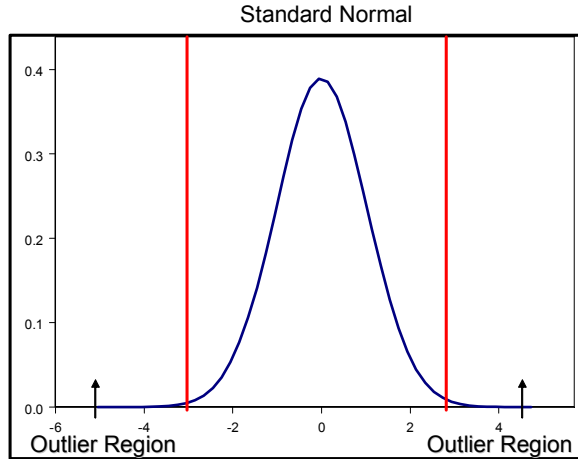


Figure One

An outlier is any data point x that falls into the outlier region.

Now consider a random sample of size N : $\mathbf{X}_N = (X_1, \dots, X_N)$. Outliers are any data points which lie in the outlier region defined as:

$$\text{OR}(\mathbf{X}_N, \alpha_N) = \{x : |x - \hat{\mu}_N| > \hat{\sigma}_N g(N, \alpha_N)\}$$

where $\alpha_N = 1 - (1 - \alpha)^{1/N}$

$\hat{\mu}_N$ is a statistical measure of location and $\hat{\sigma}_N$ is a statistical measure of scale or spread of the random sample \mathbf{X}_N . The function $g(N, \alpha_N)$ relays the distributional information of the statistics in consideration.

The Davies and Gather article called the methods that have this type of outlier regions single step identifiers. Three single step outlier identifiers were examined in this article: Extreme Studentized Deviate (ESD), the Hampel identifier and the Rousseeuw identifier. The latter two are robust with a breakdown point of $1/2$. The ESD identifier is not robust and has a breakdown point of $1/(n+1)$. Therefore, the authors expanded this technique into a multiple step identifier. The technique they tested is the ESD-EDR (extreme deviate removal). This identifier is robust with a breakdown point of $1/2$.

This paper will only consider the single-step outlier identifiers. In this way a comparison between robust (Hampel and Rousseeuw) and non-robust (ESD) methods might occur.

3. Single Step Outlier Identifiers

Lower and upper bounds can be derived for single step identifiers from the above outlier region equation. Another way to state the region is as follows:

$$\text{OR}(\mathbf{X}_N, \alpha_N) = (-\infty, L(\mathbf{X}_N, \alpha_N)) \cap (R(\mathbf{X}_N, \alpha_N), \infty)$$

where $L(\mathbf{X}_N, \alpha_N)$ is the lower bound and $R(\mathbf{X}_N, \alpha_N)$ is the upper bound of the outlier region. These bounds are

useful, especially in a Monte Carlo setting for testing whether a data point is in the outlier region.

The Extreme Studentized Deviate (ESD) identifier uses the sample mean and sample standard deviation to form the outlier region. The outlier region is then defined as:

$$\text{OR}(\mathbf{X}_N, \alpha_N) = \{x : |x - \bar{X}_N| > S_N g(N, \alpha_N)\}$$

where \bar{X}_N is the sample mean and S_N is the sample standard deviation.

The lower bound and upper bound for the outlier region:

$$L(\mathbf{X}_N, \alpha_N) = \bar{X}_N - S_N g(N, \alpha_N)$$

$$R(\mathbf{X}_N, \alpha_N) = \bar{X}_N + S_N g(N, \alpha_N)$$

The Hampel identifier uses the sample median and sample median absolute deviation to form the outlier region. The outlier region is defined as follows:

$$\text{OR}(\mathbf{X}_N, \alpha_N) = \{x : |x - \text{med}(\mathbf{X}_N)| > \text{mad}(\mathbf{X}_N) g(N, \alpha_N)\}$$

where $\text{med}(\mathbf{X}_N)$ is the sample median and $\text{mad}(\mathbf{X}_N)$ is the sample median absolute deviation.

The lower bound and upper bound for the outlier region:

$$L(\mathbf{X}_N, \alpha_N) = \text{med}(\mathbf{X}_N) - \text{mad}(\mathbf{X}_N) g(N, \alpha_N)$$

$$R(\mathbf{X}_N, \alpha_N) = \text{med}(\mathbf{X}_N) + \text{mad}(\mathbf{X}_N) g(N, \alpha_N)$$

The Rousseeuw identifier uses the sample midpoint and the length of the shortest half sample of size $[N/2]+1$ to form the outlier region. The outlier region is defined using this equation.

$$OR(\mathbf{X}_N, \alpha_N) = \{ x : |x - \text{mid}(\mathbf{X}_N)| > \min \text{ half sample}(\mathbf{X}_N) g(N, \alpha_N) \}$$

where $\text{mid}(\mathbf{X}_N)$ is the sample mid-point. The lower bound and upper bound for the Rousseeuw outlier region are easily derived similar to the other two methods.

The authors provide values for $g(N, \alpha_N)$ for each of the three identifiers. They derive these values through simulation and are assumed correct. The $g(N, \alpha_N)$ values that are provided are for $N = 20$, $N = 50$ and $N = 100$. These are the sample sizes used in the experimental design.

4. Experimental Design

Two experimental deviations from the article, “The Identification of Multiple Outliers” have been stated. Namely, the ESD-EDR technique will be replaced by the ESD technique (to compare robust verse non-robust) and the sample size will have only three levels not four ($N = 10$ is excluded since no $g(N, \alpha_N)$ are provided). One more deviation is in the form of outlier generation. The authors generated their contamination by placing them such that they were not identified as outliers but had values as large as possible. This number varies depending on the random sample. A different approach taken here is the fixing of the outlier location by creating another factor within the experimental design. This factor has four levels placing the outlier contamination at 3, 4, 5 and 6 for a standard normal. This provides a decent range in which to examine how the outlier distance from the mean affects the outlier techniques

For the different sample sizes N , random data is generated with k outliers. $N - k$ from the standard normal distribution and k from 3, 4, 5 or 6 depending on the level of the experiment. Shown below is an example histogram the outliers are generated at $x = 6$.

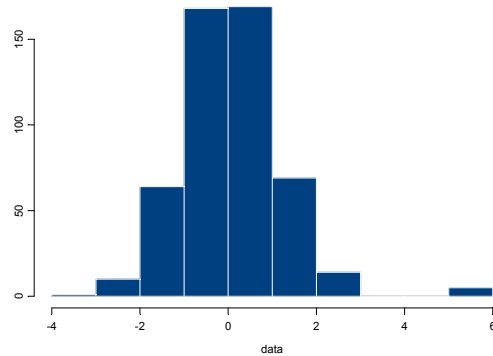


Figure Two

For this type of generated data set (standard normal with contamination) each of the three outlier identification techniques are performed using the statistical software package SPLUS.

Treatments: ESD identifier, Hampel identifier and Rousseeuw identifier

The experimental design contains the following factors and levels:

Factor: Sample Size (N)
Levels: 20, 50 and 100

Factor: Number of Contaminants (k)
Levels: 1, 2, 3, 5, 7, 10, 15, 20, 25, 30 and 49. Not to exceed $[N/2]$.

Factor: Outlier Location
Levels: 3, 4, 5 and 6

The following data table displays how the design looks and how the experiment was recorded.

Outlier =
N =

k	ESD	ROU	HAMP
1			
2			
3			
5			
7			
10			
15			
20			
25			
30			
49			

Table One

To keep the performance metric similar to the original experiment, I used the upper bound $R(\mathbf{X}_N, \alpha_N)$ for each of the three treatments. This represents the largest possible non-identified outlier. The smaller this upper bound the better the outlier method.

5. Experimental Results

SPLUS was used to code each of the outlier identification techniques and to analyze performance. Each treatment has its own separate code (Appendix A). Changes in the design points occur by changing the initialization of the variables at the beginning of each code and then rerunning the code.

Each design point had 2000 runs. The output of the code was the mean of the 2000 calculated upper bounds: $R(\mathbf{X}_N, \alpha_N)$. The variance of this mean was quite small for all but the Rousseeuw identifier. The resultant upper bounds are shown in Appendix B. One can easily compare the mean

upper bound to the actual outlier value to see if it was captured by the method. The values highlighted are considered the best based on this metric.

Using the upper bound as the measure of performance gave the following results:

ESD:

1. Performed the best for small sample size (N = 20) with close in outliers (3 and 4).
2. Did the best for all design points with very small contamination (k=1).

ROUSSEEUW:

Generally did better when the sample size was large (N=100) and the contamination was greater than 5%.

HAMPEL:

Performed the best only 3 times out of 290 total design points. This only occurred when the outlier was far away from the mean (x = 6).

Overall, it appears that the robust methods do not perform well when the sample size is small and the outliers are close in. Correspondence with one of the authors, Ursula Gather, implied this was the why they used the largest non-identified outlier as a metric. Close in outliers are not easily picked up by the robust techniques. The ESD method generally did not perform well in larger samples with contamination greater than 5%.

These conclusions coincide with the results of the simulation in the article by Davies and Gather. Their findings state that the ESD-EDR method is superior when there are a very small number of outliers. They also state that the Hampel and Rousseeuw

identifiers work well in worst-case situations (large number of outliers).

Using the metric from the original article has its faults, however. Calculating the lowest bounds does not imply that the outlier technique actually identified the outliers in the random sample. After a few minor revisions of the code in Appendix A, a different experiment was derived to determine the percentage of time the technique correctly identified outliers in the sample. The same experimental design was used while changing the performance metric to percent outlier detected (out of 2000). This yielded different results (Appendix C).

The three outlier identification techniques only correctly identified outliers in 103 out of 290 design points (often in very low percentages). When the outlier was closest to the mean, the robust techniques failed almost completely and the ESD technique only detected outliers a small percentage of the time when $k=1$. Here is a summary of the results.

ESD:

1. Worked best when contamination is less than 5%.
2. Performance was superior when contamination small and with outliers closer in to the mean.

ROUSSEEUW:

1. This method performed the best only 3 times out of the 290 total design points.
2. Appears to do better than the Hampel for smaller contamination.
3. Did not perform well for small sample ($N=20$).

HAMPEL:

1. Generally did the best for large number of outliers (greater than 5%)
2. Better performance as the outlier distance from the mean increases.

The second experiment, using outlier detected as the metric, might seem more reasonable than the metric from the first experiment. Based on these results it appears that the non-robust ESD is superior for contamination less than 5% and the Hampel identifier is generally better for data sets with larger contamination.

6. Departures from Normality

The main underlying assumption so far has been that the target distribution is standard normal. This section of the paper examines what occurs to the above outlier identification techniques when the target distribution is not normal. Two departures from normality are examined. The first departure examines what happens when the target distribution is uniform (no tails). The second departure from normality uses a double exponential as a target distribution (heavy tails). The original paper by Davies and Gather only provided values $g(N, \alpha_N)$ for normal data. To find these values the authors used simulations to derive the distribution of:

$$\frac{x_i - \hat{\mu}_N}{\hat{\sigma}_N} \text{ at } \alpha_N$$

where x_i is from the target distribution $\hat{\mu}_N$ is the estimate of location and $\hat{\sigma}_N$ is the estimate of scale.

Changing the target distribution to uniform or to double exponential changes the values for $g(N, \alpha_N)$. Therefore, new simulations are needed. A uniform with a minimum of negative three and a maximum of three were chosen to keep similar outlier values (three to six). A double exponential with a parameter of one was chosen to keep a symmetric distribution centered on zero. Using SPLUS code similar to that found in Appendix D the necessary values were derived. They are shown in the table below.

$g(N, \alpha_N)$ values Uniform (-3 , 3)

ESD	N=20	N=50	N=100
$g(N, \alpha_N)$	1.73	1.73	1.74
HAMP	N=20	N=50	N=100
$g(N, \alpha_N)$	2.01	2.01	2.02
ROU	N=20	N=50	N=100
$g(N, \alpha_N)$	1.04	1.04	1.05

$g(N, \alpha_N)$ values Double Exponential

ESD	N=20	N=50	N=100
$g(N, \alpha_N)$	3.71	4.35	4.75
HAMP	N=20	N=50	N=100
$g(N, \alpha_N)$	7.54	8.65	9.29
ROU	N=20	N=50	N=100
$g(N, \alpha_N)$	3.64	3.81	4.20

Table Two

With these values calculated a new experimental result tables are derived through Monte Carlo simulation for both uniform and double exponential distributions.

6. Departures from Normality: Experimental Results

Overall the three outlier identification techniques appeared to work better

when replacing the normal by a uniform (Appendix E). This is likely due to the lack of tails which might create masking. In the largest data set, when the outlier was far enough out all three techniques captured up to 15% contamination. This was the best performance by far. In general these outlier techniques seem to improve when the tails of the distribution are removed or reduced. Below are the summarized results for the Monte Carlo experiment using an uniform target distribution.

ESD:

1. Worked better with larger contamination than the normal (up to 15%).
2. No longer the best method for close in contamination.

ROUSSEEUW:

1. This method performed much better. Clearly the best for close in contamination under a uniform target distribution.
2. Work well for very large contamination (up to 49%).
3. The high percentage values at the right boundary of the distribution, however, might indicate a lack of efficiency. There may be potential to tag valid data as outliers.

HAMPEL:

1. Generally outperformed by the other outlier identification techniques when using uniform data.
2. Better performance as the outlier distance increases with small data sets.

Next lets examine the replacement of the normal distribution with a double exponential (Appendix F). It is clear from the experimental results that the non-robust ESD usually fails under

these conditions. The heavy tails of the double exponential create masking when using the sample mean and sample standard deviation. The Hampel outlier identification technique clearly works the best when the outlier distance is large. Below are the summarized results for the Monte Carlo experiment using a double exponential target distribution.

ESD:

1. Was never the best.
2. Did achieve 93% identification with a large sample and one outlier at a large distance.

ROUSSEEUW:

1. The best for close in contamination with a large sample size.
2. Generally outperformed by the Hampel technique.

HAMPEL:

1. Clearly the best for the double exponential (heavy tails).
2. Outperformed the other three techniques when the outlier distance was medium to large.

7. Conclusions

The results from the above Monte Carlo experiments reinforces and extends the results from the Davies and Gather article in JASA. No one statistical outlier technique dominated for every situation. The Extreme Studentized Deviate (ESD) technique performed well under the assumptions of normality and small contamination. The Rousseeuw outlier identification method performed the best when the underlying distribution had no tails. The Hampel outlier identification technique perform the best when the

target probability distribution had heavy tails or under normality when large contamination was present.

Development of a multi-step technique incorporating all three methods might be considered. However, the experiments in this paper used percent outlier detected as a measure of performance. These experiments did not consider efficiency. It might be possible that the outlier identification technique not only identifies outliers but captures a large part of target distributional data as well. This seemed to be the case with the Rousseeuw identifier under uniform assumptions. This might be similar to the Minimum Volume Ellipsoid (MVE) estimator for multivariate data. This multivariate extension of the Rousseeuw identifier is known for its lack of efficiency (Wilcox, p. 204). Measuring the efficiency of these three outlier methods is the logical next step to the development of a multi-step univariate outlier identifier.

8. References

- Barnett, V. and Lewis, T. (1994), *Outliers in Statistical Data* (3rd ed), New York: Wiley.
- Davies, Laurie and Gather, Ursula (1993), "The Identification of Multiple Outliers," *Journal of the American Statistical Association*, 88, 782-792.
- Wilcox, Rand R., (1997), *Introduction to Robust Estimation and Hypothesis Testing*, San Diego: Academic Press.