ADAPTIVE CLUSTER-BASED OUTLIER DETECTION

Tilo Strutz

Deutsche Telekom AG, Hochschule für Telekommunikation
Institute of Communications Engineering, Gustav-Freytag-Str. 43-45, 04277 Leipzig, Germany
phone: +49 341 3062 210, email: tilo.strutz@hft-leipzig.de

ABSTRACT

The analysis of data is typically accompanied by concern as to the correctness of recorded data points; some of the points might be contaminated, thereby distorting the result of the analysis. This paper proposes a novel cluster-based and distribution-independent method for outlier detection. Based on Monte Carlo simulations, the new method is tested with different data distributions and compared with the method of standardised residuals (also known as the z-score). It is shown that the cluster-based approach identifies outliers more reliably, even for a normal data distribution, and the advantages are discussed in detail.

1. INTRODUCTION

Outliers, also called as mavericks or contaminant observations, are data points that deviate so much from other points that they seem to be generated by a different mechanism than the ‘good’ observations. When observations are subject to data analysis, at least two scenarios have to be distinguished. Either (i) outliers negatively influence the results of analysis, or (ii) the search for outliers is the main task of data analysis. In data mining, for instance, outlier detection is also regarded as the detection of novelty or anomaly. Many texts propose values in the range of ±3σ, i.e. less than, say, a certain multiple of the standard deviation.

The true value σα is not known in advance. Since the number of observations i = 0, 1, . . . , N − 1 is limited, σα can only be estimated to a certain degree of accuracy by σα = √ ∑ (Δi − ̄X)2 / (N − 1).

The majority of all observations drawn from a normal distribution are less distant from the mean than a certain multiple of its standard deviation. The method of standardised residuals (also called the z-score) utilises this fact to identify contaminants

\[ |\frac{\Delta_i - \overline{X}}{\sigma_{\Delta}}| > \kappa_0. \]

All observations leading to a standardised residual larger than \( \kappa_0 \) are considered to be outliers. Many texts propose values in the range 3 ≤ \( \kappa_0 \) ≤ 4. The value could also be adapted to the number of observations \( N \) according to Chauvenet’s criterion [6]

\[ \kappa_0 = \sqrt{2 \cdot \text{inverf} \left( 1 - \frac{\nu_0}{N} \right)} , \]

where \( \nu_0 \) expresses the average number of observations with \( |\Delta| > \lambda_0 = \kappa_0 \cdot \sigma_{\Delta} \) for a given \( N \). Chauvenet proposed a proportion of \( \nu_0 = 0.5 \). This, however, would imply 0.5 outliers per data set on average. In practice, the outliers would not be evenly distributed over all possible data sets, i.e. less than fifty percent of all data sets would be said to contain outliers, but some would contain more than one. The value \( \nu_0 = 0.5 \) is obviously much too high. In principle, it is up to the implementer to choose another value \( \nu_0 \). In order to tighten the limits for outliers, a lower value should be used.

2. APPROACHES TOWARDS OUTLIER DETECTION

In the abovementioned case of least-squares approximation, \( \Delta_i \) would be centered on zero and typically follow a Gaussian distribution. Values of \( \Delta_i \) close to zero indicate good observations, whereas large absolute values indicate suspicious ones. A threshold \( \lambda_0 \) is required to discriminate between good and bad observations. \( \lambda_0 \) is a hard threshold and its determination is the critical task in outlier detection.

2.1 Standardised residuals

2.1.1 The Method

The criterion of standardised residuals assumes that the values of \( \Delta_i \) are normally distributed with a mean \( \overline{X} \) and a standard deviation of \( \sigma_{\Delta} \)

\[ f(\Delta) = \frac{1}{\sqrt{2\pi} \cdot \sigma_{\Delta}} \cdot \exp \left[ -0.5 \cdot \left( \frac{\Delta - \overline{X}}{\sigma_{\Delta}} \right)^2 \right] . \] (1)

The true value \( \sigma_{\Delta} \) is not known in advance. Since the number of observations \( i = 0, 1, . . . , N - 1 \) is limited, \( \sigma_{\Delta} \) can only be estimated to a certain degree of accuracy by \( \sigma_{\Delta} = \sqrt{\sum (\Delta_i - \overline{X})^2 / (N - 1)} \).

The majority of all observations drawn from a normal distribution are less distant from the mean than a certain multiple of its standard deviation. The method of standardised residuals (also called the z-score) utilises this fact to identify contaminants

\[ |\frac{\Delta_i - \overline{X}}{\sigma_{\Delta}}| > \kappa_0 . \] (2)

All observations leading to a standardised residual larger than \( \kappa_0 \) are considered to be outliers. Many texts propose values in the range 3 ≤ \( \kappa_0 \) ≤ 4. The value could also be adapted to the number of observations \( N \) according to Chauvenet’s criterion [6]

\[ \kappa_0 = \sqrt{2 \cdot \text{inverf} \left( 1 - \frac{\nu_0}{N} \right)} , \]

where \( \nu_0 \) expresses the average number of observations with \( |\Delta| > \lambda_0 \) for a given \( N \). Chauvenet proposed a proportion of \( \nu_0 = 0.5 \). This, however, would imply 0.5 outliers per data set on average. In practice, the outliers would not be evenly distributed over all possible data sets, i.e. less than fifty percent of all data sets would be said to contain outliers, but some would contain more than one. The value \( \nu_0 = 0.5 \) is obviously much too high. In principle, it is up to the implementer to choose another value \( \nu_0 \). In order to tighten the limits for outliers, a lower value should be used.

2.1.2 Implications of the normal distribution

Although the normal distribution has its theoretical foundation, most people are not willing to accept that a measurement can deviate from the mean. Therefore, when talking about outliers, it seems appropriate to consider a modification of the statistical model of the observations. This will not be discussed further in this paper.
Choosing the cut-off value $\lambda_0$, based on the estimated standard deviation $\sigma_0$, raises another problem. Typically, it is expected that the outlier criterion will separate the cluster of ‘good’ observations from contaminants. There should be a certain distance between the outer border of the cluster and the outliers. The standardised residual criterion, however, does not offer a separation of this kind by definition.

The next subsection proposes a new approach to outlier detection based on cluster analysis that is independent of the estimated standard deviation and takes into account that outliers should be remote from the bulk of ‘good’ observations.

### 2.2 Cluster criterion

Outlier detection based on the standardised residuals discussed above is dependent on the normal distribution of scores and the estimate of the standard deviation $\sigma_0$, which is rather uncertain where there are small numbers of observations. The basic idea behind the new method is to find a pattern, or strictly speaking a gap, in the distribution of scores that might point to the existence of outliers.

It is presumed that all non-outliers form a one-dimensional cluster in the sense that their corresponding scores are relatively close to each other, while the scores of contaminant observations are more or less remote from this cluster.

The new approach requires no special distribution of scores. The distribution merely has to be one-sided. If the initial distribution of scores has to be sorted in ascending order and numbered by $n$

$$\Delta_0[0] \leq \Delta_0[n] \leq \Delta_0[n+1] \leq \ldots \Delta_0[N-1]$$

and the differences between them are calculated

$$d[n+1] = \Delta_0[n+1] - \Delta_0[n] .$$

It is expected that the score of an outlier will show a significantly higher difference (distance) from its nearest neighbour downwards, i.e. the score will be more distant from the others than scores of measurements drawn from the correct distribution.

What qualifies a distance $d_0$ as a border (a gap) between a one-dimensional cluster of good observations and possible outliers?

1. It must be distinctly larger than a typical distance:

$$d_0 \geq \kappa_1 \cdot d_{glob} \text{ (global criterion).}$$

2. It should be substantially larger than its predecessors:

$$d_0 \geq \kappa_2 \cdot d_{loc} \text{ (local criterion).}$$

We define the typical distance for a certain score as the weighted average of distances belonging to scores which are smaller than the score corresponding to the distance $d[n]$ under investigation

$$d_{glob}[n] = \frac{1}{C_1,n} \sum_{j=1}^{n-1} d[n-j] \cdot w_j \text{ with } C_{1,n} = \sum_{j=1}^{n-1} w_j . \tag{4}$$

This avoids the influence of other potential outliers. The weighting becomes smaller with increasing $j$

$$w_j = \exp \left[ -\frac{1}{2} \left( \frac{j}{N/2} \right)^2 \right] . \tag{5}$$

### Example

Table 2 shows the sorted scores $\{ \Delta_0 \}$ (first column) and also the distances $d[n]$, averaged distances $d_{glob}[n]$ and $d_{loc}[n]$, as well as the relations $q[n] = d[n] / d_{glob}[n]$ and $r[n] = d[n] / d_{loc}[n]$.

The distances range from 0.10 to 7.60. Only one of them, $d[8]$, fulfills the global criterion $d[n] / d_{glob}[n] \geq \kappa_1$. Since the local criterion $d[8] / d_{loc}[8] \geq \kappa_2$ is also satisfied, the detection is successful.

Please note that the relation $q[10] = 6.263$ is below the threshold $\kappa_1$, although the corresponding distance $d[10] = 7.6$ is higher than $d[8]$. This is caused by the effect of accommodation.

### Table 1: Threshold $\kappa_1$ depending on the number of observations.

<table>
<thead>
<tr>
<th>$N$</th>
<th>8</th>
<th>11</th>
<th>16</th>
<th>23</th>
<th>32</th>
<th>45</th>
<th>64</th>
<th>91</th>
<th>128</th>
<th>181</th>
<th>256</th>
<th>362</th>
<th>459</th>
<th>666</th>
<th>1286</th>
<th>1502</th>
<th>1848</th>
<th>2048</th>
<th>2896</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa_1$</td>
<td>7.3</td>
<td>7.7</td>
<td>10.1</td>
<td>11.8</td>
<td>14.1</td>
<td>16.7</td>
<td>20.3</td>
<td>25.2</td>
<td>31.5</td>
<td>39.6</td>
<td>51.3</td>
<td>66.6</td>
<td>86.4</td>
<td>122</td>
<td>152</td>
<td>198</td>
<td>261</td>
<td>351</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Example of intermediate values of cluster-based outlier detection, see text for details.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\Delta_0[n]$</th>
<th>$d[n]$</th>
<th>$d_{glob}[n]$</th>
<th>$\frac{d[n]}{d_{glob}[n]}$</th>
<th>$\frac{d[n]}{d_{loc}[n]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.70</td>
<td>0.00</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>2.00</td>
<td>0.30</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>2.50</td>
<td>0.30</td>
<td>0.300</td>
<td>1.667</td>
<td>1.667</td>
</tr>
<tr>
<td>3</td>
<td>3.10</td>
<td>0.40</td>
<td>0.402</td>
<td>1.492</td>
<td>0.464</td>
</tr>
<tr>
<td>4</td>
<td>3.20</td>
<td>0.10</td>
<td>0.472</td>
<td>0.212</td>
<td>0.578</td>
</tr>
<tr>
<td>5</td>
<td>3.70</td>
<td>0.50</td>
<td>0.373</td>
<td>1.341</td>
<td>0.196</td>
</tr>
<tr>
<td>6</td>
<td>4.60</td>
<td>0.90</td>
<td>0.400</td>
<td>2.249</td>
<td>0.430</td>
</tr>
<tr>
<td>7</td>
<td>5.10</td>
<td>0.50</td>
<td>0.500</td>
<td>1.000</td>
<td>0.816</td>
</tr>
<tr>
<td>8</td>
<td>10.50</td>
<td>5.40</td>
<td>0.505</td>
<td>10.692</td>
<td>0.572</td>
</tr>
<tr>
<td>9</td>
<td>10.70</td>
<td>0.20</td>
<td>0.373</td>
<td>1.341</td>
<td>0.196</td>
</tr>
<tr>
<td>10</td>
<td>18.30</td>
<td>7.60</td>
<td>1.213</td>
<td>6.263</td>
<td>1.139</td>
</tr>
<tr>
<td>11</td>
<td>18.40</td>
<td>0.10</td>
<td>2.231</td>
<td>0.045</td>
<td>6.235</td>
</tr>
</tbody>
</table>

$\kappa_1 = 8.18$ and $\kappa_2 = 2.00$.
As one large distance has already been seen for predecessors, the new occurrence of a similar distance is no longer an indication of contaminant observations, but only of a sparse distribution. This is an important feature of the proposed method.

In order to exclude all observations not belonging to the cluster of good points, the value of the score corresponding to the critical distance \( d[8] \) is taken as threshold \( \kappa \), i.e. all observations with scores \( \Delta_i \geq \kappa \Delta_i = \kappa [8] = 10.5 \) are marked as outliers.

\[ \Delta_i \geq \kappa \Delta_i = \kappa [8] = 10.5 \]

3. ANALYSIS AND COMPARISON

3.1 Normally distributed data

It has been investigated whether the removal of observations according to the \( \pm \kappa \Delta_i \) rule of the standardised residuals method (Eq. 2) is in fact critical, and whether the proposed cluster-based approach leads to more reliable results. For different numbers \( N \) of data points, \( 10^3 \) sets of observations \( y_i \), drawn from a normal distribution \( ( \sigma = 1; \bar{y} = 0 ) \), have been generated individually. The function to be used to parameterise the data is simply the identity \( \Delta_i = |y_i| \).

3.1.1 Data sets without outliers

The method of standardised residuals has been tested in three modes: with a constant value of \( \kappa_0 = 3.5 \) and with two different adaptive values according to Eq. (3), respectively. The cluster-based outlier detection (ClubOD) has been applied as described in subsection 2.2.

The average percentage of data sets having at least one data point declared as being an outlier has been recorded (Fig. 1). When using a constant value of \( \kappa_0 \), the chance of classifying data points as outlier naturally increases with increasing \( N \), since the tails of the Gaussian bell become more and more filled. The curve corresponding to \( \kappa_0 = 0.5 \) does not seem to converge to the value of 50%. The reason for this lies in counting only the number of sets with outliers without considering the number of outliers per set. If one counts sets with two outliers twice, sets with three outliers three times, and so on, the result will in fact converge towards 50%. According to the chosen thresholds \( \kappa \) for the cluster-based method (ClubOD), each set shows on average 0.15 outliers leading to about 10\% – 12\% sets containing potential outliers.

The results of Figure 1 reveal one major problem. Even though all values have been drawn from the same distribution, some of them have been classified as outliers by definition. The question is, does the removal of these falsely classified observations harm the data analysis, i.e. the estimation of the true value of \( y \)? In order to answer, the simulation mentioned above has also compared the mean value \( \bar{y} \) of the entire set with the mean value \( \bar{y}_i \) of the reduced set, i.e. the set after removal of putative outliers. Theoretically, \( \bar{y} \) should be equal to zero, due to the parameters of the normal distribution used. It has been found that the removal improves the estimate of the mean value of \( y \) in less than fifty percent of all sets containing one or more observations classified as being contaminant (Fig. 2).

There are no significant differences between all three cases, despite the different values of \( \kappa_0 \) or the different methods.

It follows that the removal of putative outliers actually yields poorer results if sets of normally distributed data are free of contaminants, whereby ClubOD shows for \( N < 30 \) less degradation than the method of standardised residuals.

However, as the removal does not always negatively influence the estimate of \( \bar{y} \), there is a chance that this relation will change in favour of removal where the presence of outliers can be expected.

3.1.2 Data sets containing outliers

In order to investigate the effects of real outliers, the simulations have been run again with one, two or three of the original observations substituted by some values drawn from a normal distribution with other parameters \( ( \sigma = 1, \bar{y} = 4.0) \). Fig. 3 shows the results in comparison to the outlier-free case. Naturally, the percentage of data sets with putative contaminants has increased. It is not 100\%
because the inserted outliers may have values similar to the other observations, and are not detected in these cases. Furthermore, the chance of detecting outliers with the method of standardised residuals decreases for small $N$ with an increasing number of inserted contaminants, because the estimate of $\sigma_y$ is strongly influenced by the outliers and it becomes less likely that the remaining ‘good’ observations will form a distinct unit.

The proposed cluster-based approach proves advantageous when detecting outliers in small data sets, because it is not dependent on the estimation of $\sigma_y$. In larger data sets, observations are more frequently located in the tails of the Gaussian distribution, closing the gap between the bulk of good observations and outliers. Consequently, fewer outliers are detected on average. This behaviour is also beneficial, as we have to ask ourselves whether the outliers included deliberately can still be regarded as contaminant if similar values are also common for true data points.

Fig. 4 clearly shows that, as soon as outliers are present, the removal of outliers is statistically advantageous, especially for small data sets. It should also be noted that the amount of improvement of the estimated value $\hat{y}$ is on average higher than its degradation (Fig. 5). The changes are given as absolute values. With increasing $N$, the influence of outliers on the estimation of $y$ decreases, and so do the changes.

### 3.2 Non-Gaussian distribution

Albeit originally developed for normally distributed scores, the new method also works well for other distributions. As a matter of course, the method of standardised residuals will fail in these cases. The distribution of scores is dependent on the function converting the observations $y_i$ into scores $\Delta_i$. Two examples are discussed here.

#### 3.2.1 Laplace distribution

The Laplace distribution is a two-sided exponential distribution

$$f(\Delta) = \frac{1}{2 \cdot b} \exp \left( - \frac{|\Delta - \mu|}{b} \right),$$

which was investigated with $\mu = 0$ and $b = 1$. It turns out that the standard deviation $\sigma_y$ is not suitable anymore as basis for discrimination of good observations and outliers. In fact, it causes the $K_0$ criterion to reject far too many observations with increasing $N$. The cluster-based criterion, however, only shows a somewhat increased tendency to declare observations as contaminant (Fig. 6, in comparison with Fig. 3, $O=0$). On average, about 1.5 samples from the end of the tail are declared as contaminant. So it is resistant to long tails in the distribution of scores.

Most interestingly, the removal of putative outliers improves the estimated value on average. The proposed cluster-based method has
The inherent principle of the novel method is generally compatible to any distribution of scores, as soon as the scores of outliers are more distant to others than the scores of true data points making it a very versatile method. Application-specific properties must not be taken into account, since these can be incorporated into the parameterisation of data points to scores. The method is also suitable for online applications, where each newly occurring observation has to be tested. Removing an old observation as soon as a new one is included would make the approach adaptable to varying statistics.

REFERENCES

[5] Knorr, E.M; Ng, R.T.: Algorithms for Mining Distance-Based Outliers in Large Datasets. Proc. of the 24rd Int. Conf. on Very Large Data Bases, 1998, 392–403

Figure 7: Laplace distribution: percentage of data sets with better estimates of y after the removal of observations classified as outliers.

Figure 8: Uniformly distributed observations: a) percentage of data sets with at least one observation classified as an outlier; b) percentage of data sets with better estimates of y after the removal of observations classified as outliers.

Figure 9: Uniformly distributed observations: average number of detected outliers per data set, with two outliers inserted on purpose.