

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. In many applications, a set of training values is required to define ‘normality’. Security applications are examples, in which atypical behaviour by people or technical systems has to be detected. In applications with a small number of observations, however, the detection method should be able to identify outliers without prior training.

With respect to scenario (i), lively discussion can be found in the past literature on outliers, as to whether to reject suspicious values or always to keep all observations. Beckman and Cook give an overview of the history of attempts to find outliers in data sets [1]. Unfortunately, no established ‘standard’ technique has evolved to date. A comprehensive review of the tests developed for outlier detection can be found in Barnett and Lewis [2]. Outlier tests (also known as discordance tests) are often tailored to the statistical model generating the observations and presume some knowledge of the number of putative outliers. Many of them can only cope with a single outlier. With the focus on machine learning and data mining, approaches to outlier detection have been surveyed in [3] and [4].

This paper proposes a novel method of cluster-based outlier detection via scores Δ_i ($i = 0, 1, \dots, N-1$). These scores could be, for example, the deviates $y_i - \hat{y}_i$ arising from least-squares approximation of N measured data points y_i with a model function $\hat{y}_i = f(\mathbf{x}_i|\mathbf{a})$, whereas \mathbf{a} is the vector of model parameters and \mathbf{x}_i is the vector of conditions. The score Δ_i also could be the number of observations, which are within a certain radius around the observation y_i , enabling the processing of multi-variate data [5]. The new method does not depend on a certain distribution of scores, nor does it require prior training. It implicitly adapts itself and applies a threshold based on distance measures separating putative outliers from the bulk of good observations. Its advantages over the method of standardised residuals are demonstrated via Monte Carlo simulations.

2. APPROACHES TOWARDS OUTLIER DETECTION

In the abovementioned case of least-squares approximation, Δ_i would be centered on zero and typically follow a Gaussian distribu-

tion. Values of Δ_i close to zero indicate good observations, whereas large absolute values indicate suspicious ones. A threshold λ_O is required to discriminate between good and bad observations. λ_O 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 Δ_i are normally distributed with a mean $\bar{\Delta}$ and a standard deviation of σ_Δ

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

The true value σ_Δ is not known in advance. Since the number of observations $i = 0, 1, \dots, N-1$ is limited, σ_Δ can only be estimated to a certain degree of accuracy by $\hat{\sigma}_\Delta = \sqrt{\sum_i (\Delta_i - \bar{\Delta})^2 / (N-1)}$.

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

$$\left| \frac{\Delta_i - \bar{\Delta}}{\hat{\sigma}_\Delta} \right| > \kappa_O. \quad (2)$$

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

$$\kappa_O = \sqrt{2} \cdot \left[\text{inverf} \left(1 - \frac{\nu_0}{N} \right) \right], \quad (3)$$

where ν_0 expresses the average number of observations with $|\Delta_i| > \lambda_O = \kappa_O \cdot \hat{\sigma}_\Delta$ 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 ν_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 deliver a result that arbitrarily deviates from the correct value. This is, however, exactly what the range of definition $-\infty \leq \Delta \leq +\infty$ of the normal distribution is telling us. In practical cases of limited numbers N of values Δ_i everybody may reject a value that is further than, say, a certain multiple of the standard deviation σ_Δ away 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.

Table 1: Threshold κ_1 depending on the number of observations.

N	8	11	16	23	32	45	64	91	128	181	256	362	512	724	1024	1448	2048	2896
κ_1	7.3	7.7	10.1	11.8	14.1	16.7	20.3	25.2	31.5	39.6	51.3	66.6	86.4	112	150	198	261	351

Choosing the cut-off value λ_O based on the estimated standard deviation $\hat{\sigma}_\Delta$ 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 $\hat{\sigma}_\Delta$, 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 is Gaussian and centred on zero, for example, this requirement can be simply achieved by mapping all negative values into the range of positive values. Putative contaminants are identified by comparing the distances between the scores Δ_i .¹

The decision regarding the existence of more than one single cluster of scores is made using the following algorithm. First, the scores have to be sorted in ascending order and numbered by n

$$\Delta_s[0] \leq \dots \leq \Delta_s[n] \leq \Delta_s[n+1] \leq \dots \Delta_s[N-1]$$

and the differences between them are calculated

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

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_b 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:
 $d_b \geq \kappa_1 \cdot d_{\text{glob}}$ (global criterion).
2. It should be substantially larger than its predecessors:
 $d_b \geq \kappa_2 \cdot d_{\text{loc}}$ (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

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

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

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

¹For simplicity, the same symbol Δ is also used for the mapped values.

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

n	$\Delta_s[n]$	$d[n]$	$d_{\text{glob}}[n]$	$\frac{d[n]}{d_{\text{glob}}[n]}$	$d_{\text{loc}}[n]$	$\frac{d[n]}{d_{\text{loc}}[n]}$
0	1.70	0.00	0.000	0.000	0.000	0.000
1	2.00	0.30	0.000	0.000	0.000	0.000
2	2.50	0.50	0.300	1.667	0.300	1.667
3	3.10	0.60	0.402	1.492	0.464	1.294
4	3.20	0.10	0.472	0.212	0.578	0.173
5	3.70	0.50	0.373	1.341	0.196	2.553
6	4.60	0.90	0.400	2.249	0.430	2.095
7	5.10	0.50	0.500	1.000	0.816	0.613
8	10.50	5.40	0.505	10.692	0.572	9.446
9	10.70	0.20	1.335	0.150	4.451	0.045
10	18.30	7.60	1.213	6.263	1.139	6.673
11	18.40	0.10	2.231	0.045	6.235	0.016
			$\kappa_1: 8.18$		$\kappa_2: 2.00$	

The local criterion utilises the same averaging process, but with weights falling off more rapidly in order to express closeness

$$d_{\text{loc}}[n] = \frac{1}{C_{2,n}} \cdot \sum_{j=1}^{n-1} d[n-j] \cdot \exp \left[-\frac{1}{2} \cdot \left(\frac{j}{N/12} \right)^2 \right]. \quad (6)$$

The denominators $N/2$ and $N/12$ have been determined empirically.

Please note that both, $d_{\text{glob}}[n]$ and $d_{\text{loc}}[n]$ are different for each distance $d[n]$ reflecting the adaptive character of the approach.

In summary, the presence of one or more outliers is indicated if there is a score $\Delta_s[n]$ showing a distance $d[n]$ from the next score down $\Delta_s[n-1]$, which has the two properties

$$d[n] \geq \kappa_1 \cdot d_{\text{glob}}[n] \quad \text{and} \quad d[n] \geq \kappa_2 \cdot d_{\text{loc}}[n]. \quad (7)$$

The value of the corresponding score $\Delta_s[n]$ is taken as the cut-off value λ_O .

Experiments with different kinds of data sets have shown that κ_2 can be set to a fixed value of 2, while κ_1 should be dependent on the number of observations N . Suitable values for κ_1 have been derived from computer simulations, **Tab. 1**. If N is a number between these points, the corresponding κ_1 should be interpolated.

Using the approach described, even multiple outliers can be eliminated at once, because the score Δ_b corresponding to the distance d_b fulfilling Eqs. (7) only marks the border between the two clusters of good observations and contaminants. It is evident that all observations having higher absolute scores also belong to the cluster of outliers.

The entire procedure of outlier detection based on distances is explained in following example.

Example:

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

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

Please note that the relation $q[10] = 6.263$ is below the threshold κ_1 , although the corresponding distance $d[10] = 7.6$ is higher than $d[8]$. This is caused by the effect of accommodation.

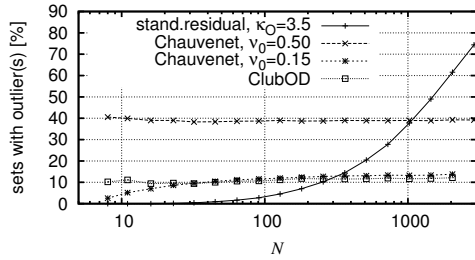


Figure 1: Normally distributed observations: percentage of data sets with at least one observation classified as outlier.

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 λ_O , i.e. all observations with scores $\Delta_i \geq \lambda_O = \Delta_s[8] = 10.5$ are marked as outliers. \square

3. ANALYSIS AND COMPARISON

3.1 Normally distributed data

It has been investigated whether the removal of observations according to the $\pm\kappa_O \cdot \hat{\sigma}_\Delta$ 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, 10^5 sets of observations y_i , drawn from a normal distribution ($\sigma_y = 1; \bar{y} = 0$), 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 $\kappa_O = 3.5$ 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 κ_O , 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 $\nu_0 = 0.5$ 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 κ_1 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 \hat{y} of the entire set with the mean value \hat{y}' of the reduced set, i.e. the set after removal of putative outliers. Theoretically, \hat{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

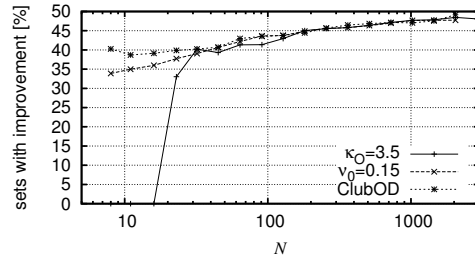


Figure 2: Normally distributed observations: percentage of data sets with better estimates of y after the removal of observations classified as outliers.

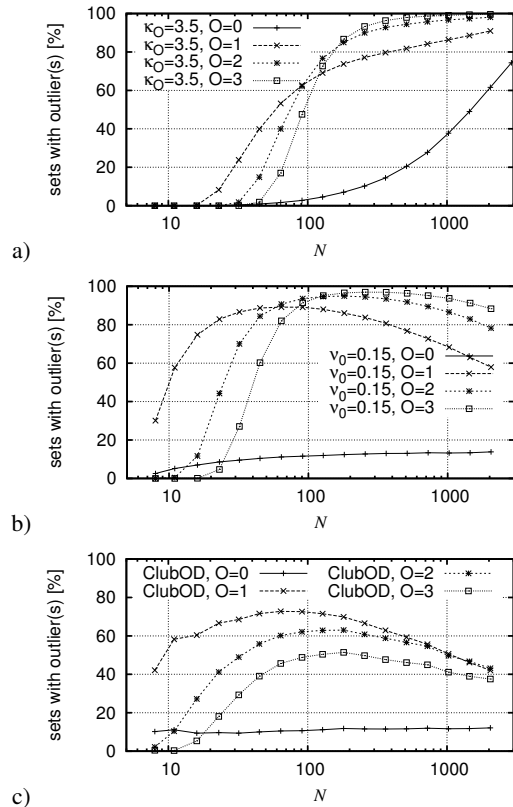


Figure 3: Normally distributed data: percentage of data sets with at least one observation classified as outlier after insertion of one, two or three contaminants; a) $\kappa_O = 3.5$, b) Chauvenet's criterion $\nu_0 = 0.15$, c) cluster criterion.

the different values of κ_O 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 \hat{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 ($\sigma_y = 1, \bar{y} = 4.0$). 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%

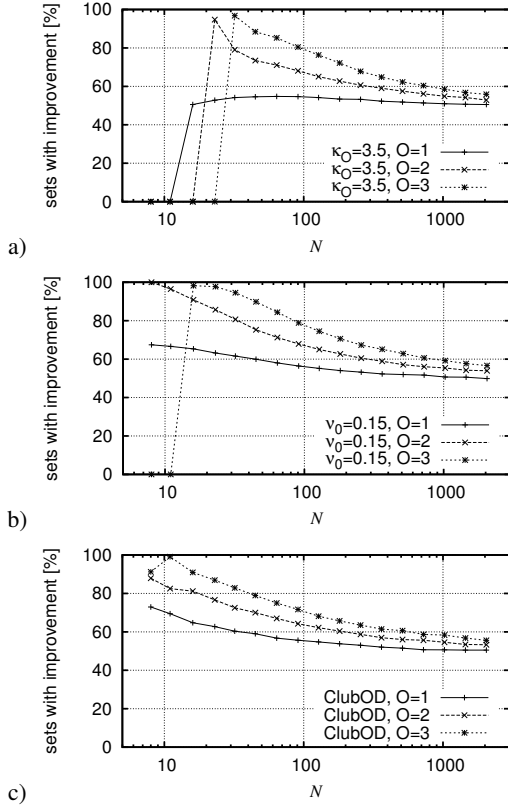


Figure 4: Normally distributed data: percentage of data sets with better estimates of y after the removal of observations classified as outliers. Data sets with insertion of one, two or three contaminants; a) $\kappa_O = 3.5$, b) Chauvenet's criterion $v_0 = 0.15$, c) cluster criterion.

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 σ_Δ 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 σ_Δ . 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 Δ_i . Two examples are discussed here.

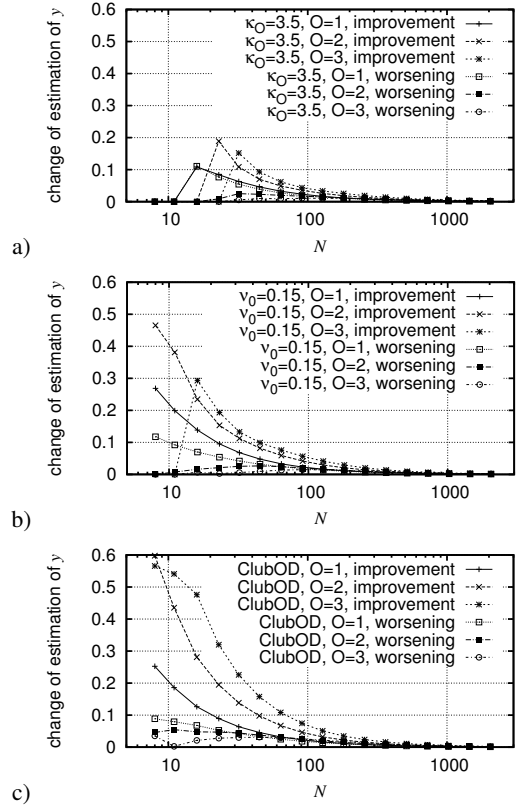


Figure 5: Normally distributed data: quantitative change of estimated \hat{y} after the removal of observations classified as outliers. Data sets with insertion of one, two or three contaminants; a) $\kappa_O = 3.5$, b) Chauvenet's criterion $v_0 = 0.15$, c) cluster criterion.

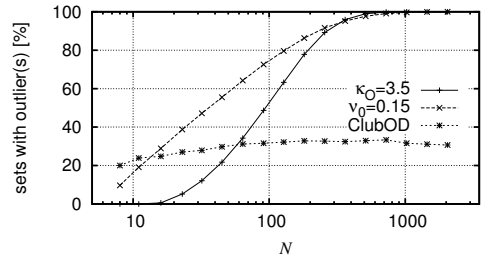


Figure 6: Laplace distribution: percentage of data sets with at least one observation classified as an outlier.

3.2.1 Laplace distribution

The Laplace distribution is a two-sided exponential distribution

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

which was investigated with $\mu = 0$ and $b = 1$. It turns out that the standard deviation σ_Δ is not suitable anymore as basis for discrimination of good observations and outliers. In fact, it causes the $\kappa_O \cdot \hat{\sigma}_\Delta$ 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

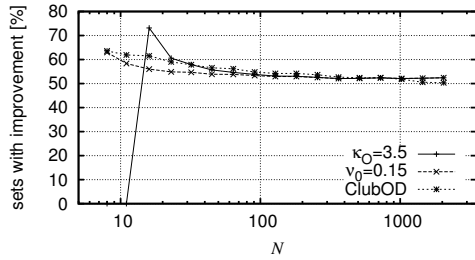


Figure 7: Laplace distribution: percentage of data sets with better estimates of y after the removal of observations classified as outliers.

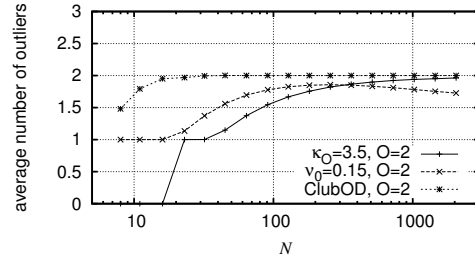


Figure 9: Uniformly distributed observations: average number of detected outliers per data set, with two outliers inserted on purpose.

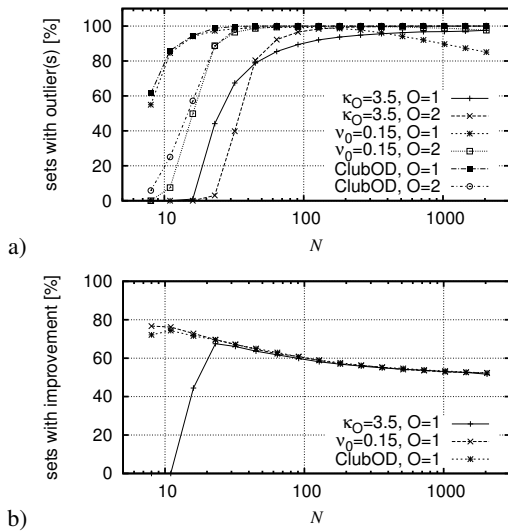


Figure 8: Uniformly distributed observations: a) percentage of data sets with at least one observation classified as an outlier; b) percentage of data sets with better estimates of y after the removal of observations classified as outliers.

here the highest percentage of improved data sets (Fig. 7).

3.2.2 Uniformly distributed data

When applying both outlier detection schemes to uniformly distributed and outlier-free data ($-1 \leq y_i \leq +1$), almost no outliers are detected and the few cases of wrong classification mostly decrease the quality of the estimated \hat{y} . Using a fixed threshold $\kappa_O = 3.5$, none of the observations is classified as an outlier.

After the insertion of outliers, uniformly distributed in the range of $2 \dots 4$, the proportion of sets classified as containing at least one outlier significantly increases (Fig. 8 a). Only in cases of small data sets, do the inserted outliers influence the determination of σ_Δ (standardised residuals) or the computation of $d_{g\text{lob}}$ (cluster-based approach) so strongly that the outliers are likely to become part of the cluster of ‘good’ observations. In terms of improvement after removal of putative contaminants, there is no significant difference between the approaches. That is why Fig. 8 b only shows the results based on a single inserted outlier.

It might be of interest that if the data contains two outliers, the proposed approach typically finds both, whereas the method of standardised residuals often detects only one (Fig. 9).

4. DISCUSSION AND SUMMARY

The Monte-Carlo simulations presented in this paper underline numerically that the removal of observations decreases the estimation accuracy if normal distribution is assumed and the data set contains

no outliers. This result can be generalised for any unbounded distribution. If any arbitrarily large value is an element of the distribution, none of the observations may be regarded as an outlier, regardless of the criterion that is used. The situation changes as soon the presence of outliers can be assumed. In the majority of cases, the removal of potential outliers improves the data analysis, and the improvements are higher than the effect, by which the estimation becomes worse.

The simulations have also revealed that the method of standardised residuals sometimes places the threshold λ_O in between similar observations, which contradicts the intuitive decision as to the definition of outliers. In contrast, the proposed method considers the distances between scores, and only removes those points that are in fact remote from the cluster of true observations. In addition, the new cluster-based method is able to adapt itself to the possibly sparse distribution of scores.

The benefits from the cluster-based approach become especially apparent if the scores are not normally distributed. In the case of Laplace distribution, the method of standardised residuals declares far too many observations as contaminant, whereas the proposed method is only slightly affected. In case of uniformly distributed data, only the proposed approach is able to detect true outliers with sufficient reliability, because it finds multiple outliers, while the method of standardised residual often only finds one out of two inserted outliers, for example.

The statistical relevance of results has been obtained by large scale tests based on simulated data. Presenting results of a particular data set from a concrete application would not prove or disprove the effectiveness of the proposed approach.

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 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

- [1] Beckman, R.J.; Cook, R.D., “Outliers”, *Technometrics*, Vol.25, No.2, May 1983, 119–149
- [2] Barnett, V.; Lewis, T., *Outliers in Statistical Data*, 3rd edition, WILEY, 1994, ISBN 0-471-93094-6
- [3] Markou, M.; Singh, S., “Novelty Detection: A Review. Part I + II”, *Signal Processing*, 2003, Vol. 83, 2481–2521
- [4] Hodge, V.J; Austin, J., “A survey of Outlier Detection Methodologies”, *Artificial Intelligence Review*, 2004, 22, 85–126
- [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
- [6] Chauvenet, W., “Method of Least Squares”, Appendix to *Manual of Spherical and Practical Astronomy*, Vol.2, 4th edition, J.B. Lippincott & Co., Philadelphia, 1871, 469–566