SLIIT Journal of Humanities and Sciences
Vol. 5(1), 2024; Page 01-12
DOI: https://doi.org/10.4038/sjhs.v5i1.65

Received: 20 January 2024 *Accepted: 30 May 2024*

Comparing Methods for Detecting Anomalous Values in Automated Laboratory Processes

Madhushanka, H. M. S.[1] & Amaratunga, D.[2]

sahanmadhushanka92@gmail.com,damaratung@yahoo.com
[1]*Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka.*
[2]*Princeton Data Analytics, USA.*

Abstract

Outlier detection is used in many domains. In automated laboratory processes, detecting anomalous values is critical for ensuring the reliability of experimental results. This study compares various outlier detection methods, including traditional statistical approaches like Mahalanobis distance, Median and mean absolute deviation (MAD), as well as modern machine learning techniques such as Isolation Forest, Angle Based Outlier Detection (ABOD), and Local Outlier Factor (LOF). The performances of these methods were evaluated using simulated multivariate data, with different types of outliers and levels of contamination. Comparisons are made using sensitivity, precision, and mainly the F2 score, a weighted metric of sensitivity and precision that gives more weight to precision. The results show that in univariate settings, the Median MAD method works consistently well. For multivariate scenarios, Mahalanobis methods with Minimum Covariance Determinant estimates and Minimum Volume Ellipsoid estimates work well even for high contamination percentages. This study highlights the importance of selecting an appropriate outlier detection method for the situation.

Keywords: Experimental reliability, Machine learning, Multivariate data, Outlier detection, Statistical methods.

Introduction

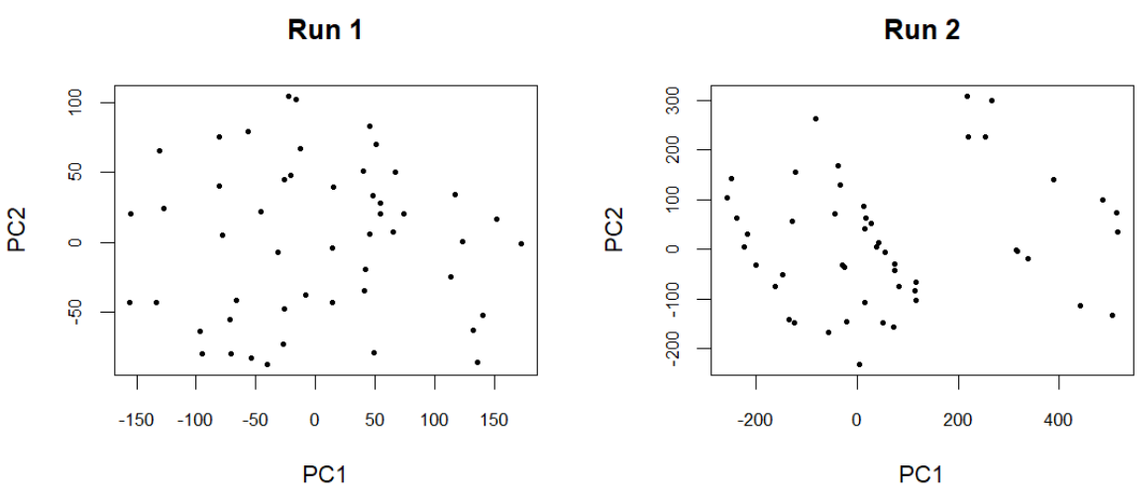
Many laboratory processes now employ various forms of automated systems driven by developments in robotics and other related technologies. To ensure the quality of the experimental results generated by these processes, the progress of such experiments is usually monitored using control samples. Checking the assay results from these control samples and detecting anomalies among them will help identify potential sources of unwanted variation, which can then be dealt with quickly and efficiently to maintain the integrity and reliability of the assay results.

As an example, let us consider data from an experimental procedure that was used to test a large number of compounds to identify those that induce specific changes in the structure of cells due to disease. The spatial organization of cellular structure and components was quantified by five variables. Each run of the

experiment involved 50 plates; each plate had several inactive control samples. The median of each variable for those control samples for each plate was calculated, resulting in 50 observations with 5 variables per run. Robust principal component analyses were performed on the control samples from two runs of the experiment (Fig. 1). In Run 1 of the experiment, the data look reasonably well-behaved, whereas in Run 2, there seem to be several anomalous values. As inactive samples should give approximately the same value, anomalous values indicate some error in that plate. Then, steps should be taken to identify the reason for the anomaly and any experimental problems should be fixed; properly identifying anomalies could help to eliminate potential experimental issues early. Thus, given the value of this issue, it is of interest to compare outlier detection methods to assess which ones would be most effective at correctly identifying anomalous values in such situations.

Figure 1.

Scatter plots of the 1st and 2nd principal components of robust principal component analyses.



A few methods have been proposed for detecting outliers in multivariate data. Some of the initially proposed methods, such as those based on the Mahalanobis Distance (Reprint of: Mahalanobis, 1936), are designed to find outliers in parametric models, whereas more recent methods, such as Isolation Forest (Liu et al., 2008) and Angle Based Outlier Detection (ABOD) (Kriegel et al., 2008), are designed for use in machine learning applications in which very large datasets are being mined. While these latter methods are useful in large data situations, they may not be as effective in small sample situations, especially considering that the performance of the Mahalanobis distance for outlier detection can be enhanced by using robust versions of the mean and variance. In this paper, we shall compare the performance of several of these methods in situations similar to typical laboratory processes.

Methods

For the present study, 7 methods have been considered. The methods were chosen because of their frequent use in multivariate outlier detection.

Univariate method using median and MAD

A commonly used method for detecting outliers in univariate data is the median MAD method. The median and Median Absolute Deviation (MAD) are much more resistant to outliers than the mean and standard deviation and therefore can be used to define a range of acceptable values for a variable (Iglewicz & Hoaglin, 1993). Any observation outside of the range ($Median-l \times MAD$, $Median + l \times MAD$) is considered an outlier. Typically, l is taken as 3. To apply this to a multivariate situation, each variable was analyzed separately using median MAD. If a value is flagged as an outlier in at least one variable, then that observation is considered anomalous.

Mahalanobis distance based methods

Mahalanobis distance was introduced by Mahalanobis in 1936. It is a multivariate generalization of the Z-score that measures how many standard deviations away an observation is from the mean. Observations beyond a specific distance could be considered outliers. Measures of center and spread are needed to calculate Mahalanobis distance. In this case, three possibilities were considered:

- Mean and Covariance
- Minimum Covariance Determinant (MCD) estimates of center and spread (Hubert & Debruyne, 2010)
- Minimum Volume Ellipsoid (MVE) estimates of center and spread (Van Aelst & Rousseeuw, 2009)

Here, mean vector and covariance matrix are most affected by outliers, but MCD and MVE estimates of spread and center are resistant to outliers and therefore likely to have better performance when identifying anomalous values.

Angle based outlier detection (ABOD)

Angle-based outlier detection method (Kriegel et al., 2008) is a parameter-free method that is designed to find outliers based on angles between observations. The idea behind this method is that the angles formed between an outlier and other pairs of points tend to be more similar, resulting in less variation in angles. However, normal points are surrounded by many points in different directions, so the angles vary more. For each point, the Angle-based outlier factor (ABOF) is calculated; this can be done considering all points, but this can be computationally expensive, so in this comparison, for each point, 5 random points were considered to calculate ABOF. An ABOF factor close to 0 indicates an outlier.

Local outlier factor (LOF)

Local outlier factor (LOF) is an outlier detection method introduced in 2000 (Breunig et al., 2000). LOF is a measurement of local density deviation compared to its neighbouring points.

Isolation forest (iForest)

Isolation forest was introduced in 2008 (Liu et al., 2008). Most outlier detection methods work by identifying patterns of normal

observation and then flagging observations that deviate from those patterns as outliers. This method works by isolating outliers. The isolation forest method gives a score between 0 and 1. Observations with scores higher than 0.5 are considered outliers.

Simulation

Comparison criteria

This simulation study shall compare the performance of the various outlier detection methods using the following criteria: sensitivity, precision, and F2 score (Table 1). However, comparisons will be mainly done using the F2 score, which is a weighted combination of sensitivity and precision. A good outlier detection procedure should be able to detect outliers in a dataset with high sensitivity and not incorrectly flag "good observations" as outliers, i.e., have high precision, and thus have a high F2 score.

Table 1.
Confusion matrix.

	Predicted: Outlier	Predicted: Normal
Actual: Outlier	True Positives	False Negatives
Actual: Normal	False Positives	True Negatives

Sensitivity

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}} \quad (1)$$

Sensitivity must be high in this case because it is crucial to find every out-of-control observation. Overlooking any outliers could lead to problems later.

Precision

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives+False Positives}} \quad (2)$$

The percentage of actual outliers out of the observations that were flagged as outliers must be high, as low values mean that detecting non-outliers as outliers is likely. That can raise costs.

F2 score

The F2 score is calculated using the equation (3).

$$\text{F2 Score} = \frac{5*\text{sensitivity}*\text{precision}}{4*\text{precision+sensitvity}} \quad (3)$$

This must also be high, as this is a combination of sensitivity and precision. The F2 score gives more weight to sensitivity, as concern is more on outliers not being flagged as outliers than acceptable values mistakenly being flagged as outliers. Although sensitivity and precision are considered as performance measures, the ultimate comparison was done using the F2 score.

Simulation

All the simulations were done in R Studio. Multivariate normal data with five variables was simulated with the mean vector and the covariance matrix of the observations that were assumed to be non-outliers. At each iteration, 50 or fewer observations were generated using this distribution, and then outliers were added to bring the total number of observations to 50. Those observations were labelled for later calculation of sensitivity and other measures. Two different contamination percentages (10% and 40%) were considered for all cases. These two contamination percentages were considered to compare outlier detection method performance on a few numbers of outliers and a high number of outliers. For one case, 1000 iterations were done, and average

sensitivity, precision, and F2 scores were calculated.

Multivariate normal data was generated using the mean vector (μ) and the covariance matrix of observations (Σ) of the given dataset (dataset 1) that were assumed to be non-outliers. Although the given data does not exactly follow a multivariate normal distribution, to reduce the complexity of simulating a complex multivariate distribution that closely resembles the given data set, this study has approximated it to a multivariate normal distribution with the same mean vector and sample covariance. Outliers were generated in seven ways. When generating outliers few different cases were considered.

Σ− the covariance matrix, Σ_{ij}− the element in i^{th} row and j^{th} column
k− determines the outlyingness of the outliers.

i. **1 variable:** $\Sigma_{11} = k \times \Sigma_{11}$
ii. **2 variables:** $\Sigma_{ii} = k \times \Sigma_{ii}$ for i =1,2
iii. **3 variables:** $\Sigma_{ii} = k \times \Sigma_{ii}$ for i=1,2,3
iv. **4 variables:** $\Sigma_{ii} = k \times \Sigma_{ii}$ for i=1,2,3,4
v. **5 variables:** $\Sigma_{ii} = k \times \Sigma_{ii}$ for i=1,2,3,4,5
vi. **Shift outliers in the direction of highest variance:** These outliers were generated by $N(k_1,\Sigma)$ where $k_1 = \mu + kv_1$. v_1 is the eigenvector of Σ with the highest eigenvalue.
vii. **Shift outliers in the direction of lowest variance:** These outliers were generated by $N(k_5,\Sigma)$ where $k_5 = \mu + kv_5$. v_5 is the eigenvector of Σ with the lowest eigenvalue.

These types of outliers were generated to cover the main types of outliers. The k values were chosen to clearly separate out the anomalies from the normal observation cluster. Line charts were also drawn to compare performance across different k values.

Results

The performance of the outlier methods is shown in the tables 2-8. The highest F2 score for each row is highlighted in blue, and every F2 score that is higher than 0.7 is highlighted in green.

Table 2.

One variable.

Method	5 Outliers			20 Outliers		
	Sensitivity	Precision	F2	Sensitivity	Precision	F2
Mahalanobis	0.337	0.651	0.382	0.102	0.783	0.130
Mahalanobis MCD	0.618	0.328	0.520	0.456	0.779	0.494
Mahalanobis MVE	0.587	0.408	0.533	0.357	0.769	0.396
ABOD	0.615	0.375	0.539	0.458	0.800	0.499
LOF	0.690	0.293	0.536	0.425	0.629	0.451
IForest	0.537	0.283	0.453	0.313	0.617	0.346
Median MAD	0.687	0.281	0.525	0.565	0.676	0.580

Table 3.

Two variables.

Method	5 Outliers			20 Outliers		
	Sensitivity	Precision	F2	Sensitivity	Precision	F2
Mahalanobis	0.601	0.841	0.630	0.181	0.928	0.215
Mahalanobis MCD	0.863	0.434	0.706	0.732	0.937	0.763
Mahalanobis MVE	0.851	0.520	0.742	0.655	0.935	0.693
ABOD	0.616	0.375	0.539	0.458	0.800	0.499
LOF	0.694	0.295	0.538	0.427	0.634	0.453
IForest	0.759	0.460	0.666	0.411	0.818	0.455
Median MAD	0.876	0.352	0.664	0.761	0.778	0.761

Table 4.

Three variables.

Method	5 Outliers			20 Outliers		
	Sensitivity	Precision	F2	Sensitivity	Precision	F2
Mahalanobis	0.779	0.921	0.798	0.271	0.980	0.316
Mahalanobis MCD	0.956	0.467	0.778	0.887	0.964	0.900
Mahalanobis MVE	0.95	0.554	0.818	0.848	0.967	0.868
ABOD	0.731	0.447	0.639	0.433	0.868	0.479
LOF	0.903	0.350	0.676	0.752	0.734	0.745
IForest	0.886	0.625	0.812	0.507	0.953	0.557
Median MAD	0.949	0.389	0.725	0.866	0.845	0.860

Table 5.

Four variables.

Method	5 Outliers			20 Outliers		
	Sensitivity	Precision	F2	Sensitivity	Precision	F2
Mahalanobis	0.878	0.957	0.890	0.357	0.996	0.408
Mahalanobis MCD	0.983	0.477	0.798	0.956	0.969	0.958
Mahalanobis MVE	0.979	0.559	0.839	0.938	0.973	0.944
ABOD	0.744	0.454	0.650	0.432	0.872	0.478
LOF	0.959	0.362	0.712	0.826	0.740	0.804
IForest	0.949	0.777	0.903	0.587	0.998	0.638
Median MAD	0.979	0.419	0.760	0.927	0.903	0.921

Table 6.

Five variables.

Method	5 Outliers			20 Outliers		
	Sensitivity	Precision	F2	Sensitivity	Precision	F2
Mahalanobis	0.938	0.981	0.944	0.450	1.000	0.504
Mahalanobis MCD	0.994	0.478	0.805	0.983	0.968	0.979
Mahalanobis MVE	0.993	0.565	0.850	0.976	0.973	0.975
ABOD	0.752	0.458	0.657	0.432	0.874	0.478
LOF	0.979	0.366	0.724	0.853	0.742	0.825
IForest	0.978	0.865	0.949	0.658	1.000	0.705
Median MAD	0.993	0.449	0.786	0.959	0.972	0.961

As expected, the Median MAD method performs well for the one variable outlier situation with high F2 scores. As the number of outlier variables increases, the number of methods that give a score of more than 0.7 increases. For 2,3,4, and 5 outlier variables situations with 20 outliers, the Mahalanobis MCD method has the best F2 scores, but methods like Mahalanobis MVE and Median MAD come close in terms of F2 scores. For 2 and 3 variables scenarios with 5 outliers Mahalanobis MVE method has the highest F2 scores. For 4 and 5 variables scenarios with 5 outliers, the Isolation Forest method has the highest F2 scores. As the number of outliers increases, the F2 scores of methods like Mahalanobis MCD increase, but for some methods like Mahalanobis and ABOD, the F2 scores decrease. F2 scores of other methods sometimes increase, but sometimes decrease, depending on the scenario.

Table 7.

Shift outliers in the direction of the highest variance.

Method	5 Outliers			20 Outliers		
	Sensitivity	Precision	F2	Sensitivity	Precision	F2
Mahalanobis	0.612	0.781	0.641	0.037	0.496	0.083
Mahalanobis MCD	1.000	0.48	0.809	1.000	0.964	0.992
Mahalanobis MVE	1.000	0.566	0.855	0.729	0.786	0.755
ABOD	0.739	0.437	0.640	0.238	0.557	0.267
LOF	0.000	0.000	NaN	0.233	0.410	0.256
IForest	1.000	0.574	0.867	0.546	0.623	0.558
Median MAD	1.000	0.491	0.816	1.000	1.000	1.000

The Isolation Forest MAD method has the highest F2 score. Mahalanobis MCD, Mahalanobis MVE, Isolation Forest MAD, and Median MAD outlier methods have F2 scores higher than 0.7. The LOF method seems to perform the worst with 0 sensitivity and 0 PPV. For 20 outlier cases, the Median MAD method has the highest F2 score of 1.

The only other methods to have F2 scores higher than 0.7 are Mahalanobis MCD and Mahalanobis MVE. All other methods seem to perform worse with low F2 scores.

Table 8.

Shift outliers in the direction of lowest variance.

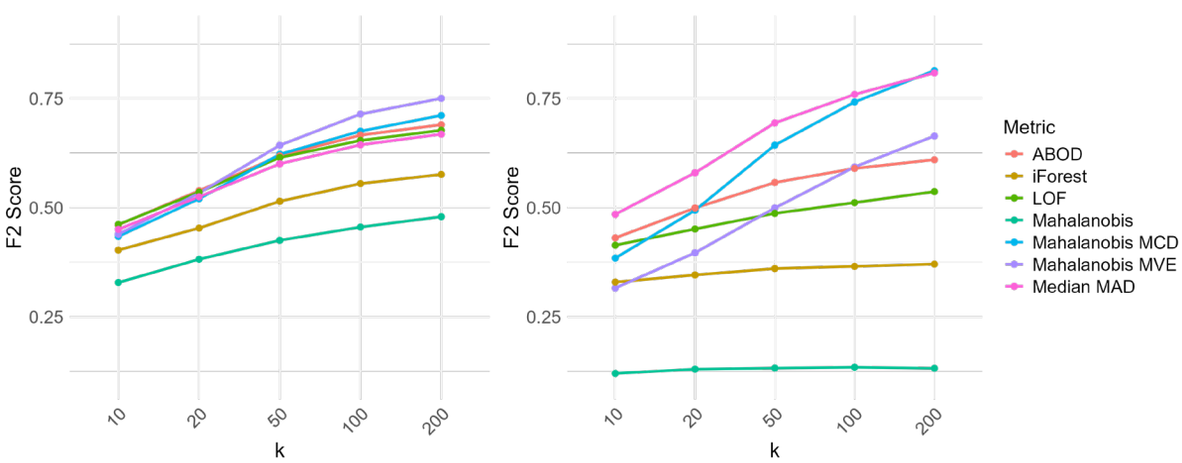
Method	5 Outliers			20 Outliers		
	Sensitivity	Precision	F2	Sensitivity	Precision	F2
Mahalanobis	0.611	0.776	0.635	0.037	0.518	0.081
Mahalanobis MCD	1.000	0.48	0.810	1.000	0.964	0.992
Mahalanobis MVE	1.000	0.566	0.855	1.000	0.965	0.992
ABOD	0.647	0.486	0.6	0.084	0.735	0.124
LOF	0.000	0.000	NaN	0.233	0.41	0.256
IForest	1.000	0.546	0.853	0.509	0.613	0.525
Median MAD	1.000	0.427	0.776	1.000	0.881	0.973

For 5 outliers case, the Isolation Forest has the highest F2 score. Other methods that have a score of more than 0.7 are Mahalanobis MCD, Mahalanobis MVE, Isolation Forest, and MAD outlier. Here, also LOF performs worst with 0 sensitivity and 0 PPV. For 20 outliers case, both Mahalanobis MCD and Mahalanobis MVE have the highest F2 scores. The only other method to work well with a high score is the Median MAD method.

Different distance

Figure 2.

One variable 5 outliers (left) 20 outliers (right).



As the distance increases F2 score generally increases for every method (Fig. 2). With 5 outliers for small k values like 10 and 20, LOF and ABOD methods seem to have high F2 scores but for higher distances, Mahalanobis MVE outperforms other methods.

Mahalanobis MCD method also comes in 2nd place in the F2 score. In the case of 20 outliers, the median MAD method has high F2 scores for all k distances. Mahalanobis MCD method also gets a comparable F2 score, but only for high K values.

Figure 3.

Two variables 5 outliers (left) 20 outliers (right).

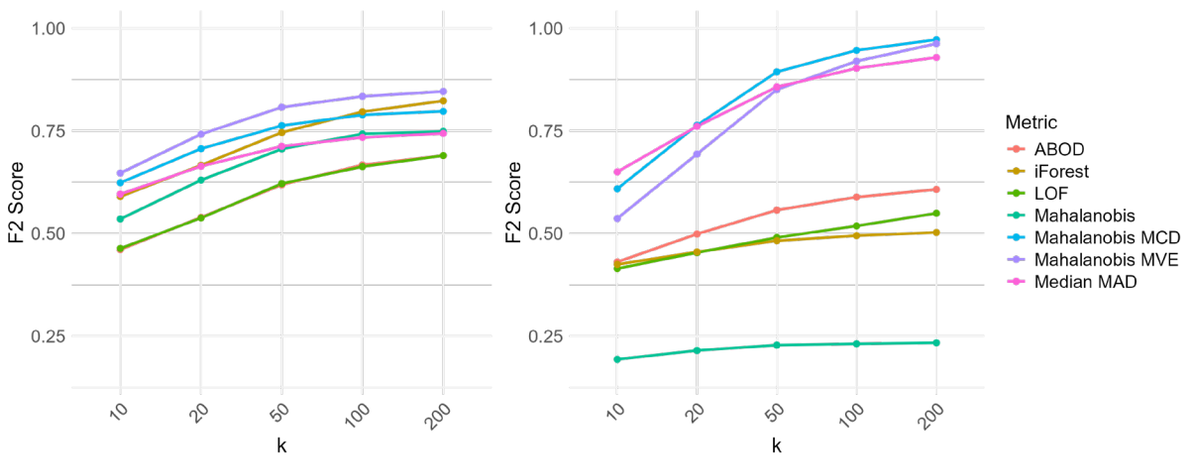
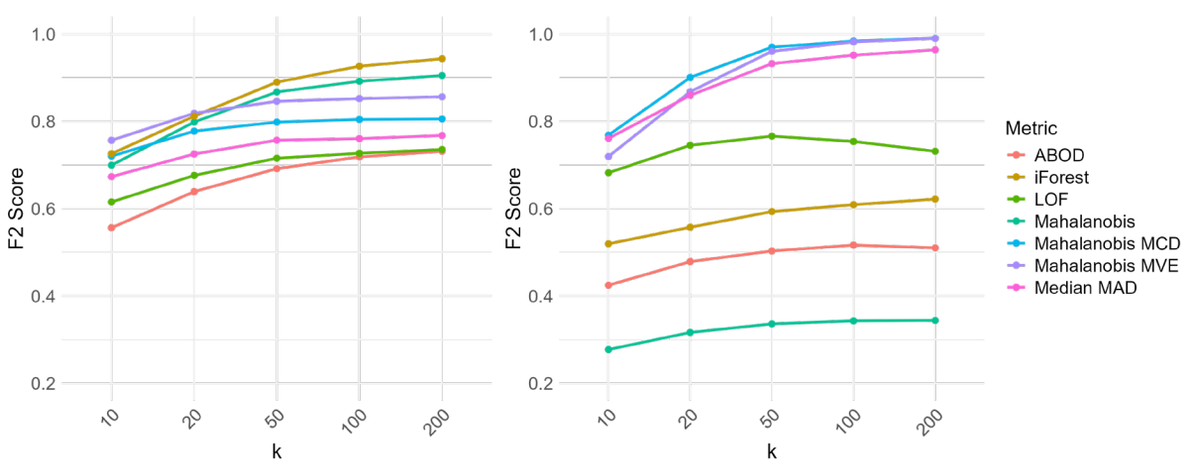


Figure 3 indicates that in two-variable cases with 5 outliers, the Mahalanobis MVE method has the highest F2 scores across all distances. For k 10, 20, and 50, Mahalanobis MCD performs well with the 2nd best F2 score, but for k values 100 and 200 Isolation Forest method has the 2nd highest F2 score. In the case of 20 outliers, the median MAD method has the highest F2 score for k=10, but for k=20, Mahalanobis MCD has an equivalent F2 score. For high k values, the best-performing method seems to perform better than the best-performing methods in the 5 outliers' case.

Figure 4.

Three variables 5 outliers (left) 20 outliers (right).



In three variable cases with 5 outliers, the Mahalanobis MVE method achieves the highest F2 scores for k values of 10 and 20, while for k values of 50, 100, and 200, the isolation forest has the highest scores (Fig.4). In the case of 20 outliers, the Mahalanobis MCD method has high F2 scores for all k values. Mahalanobis MVE also has a close F2 score for high k values. The median MAD method has a score that is close to the highest F2 score for k=10 and the 3rd highest scores for other k values.

Figure 5.

Four variables 5 outliers (left) 20 outliers (right).

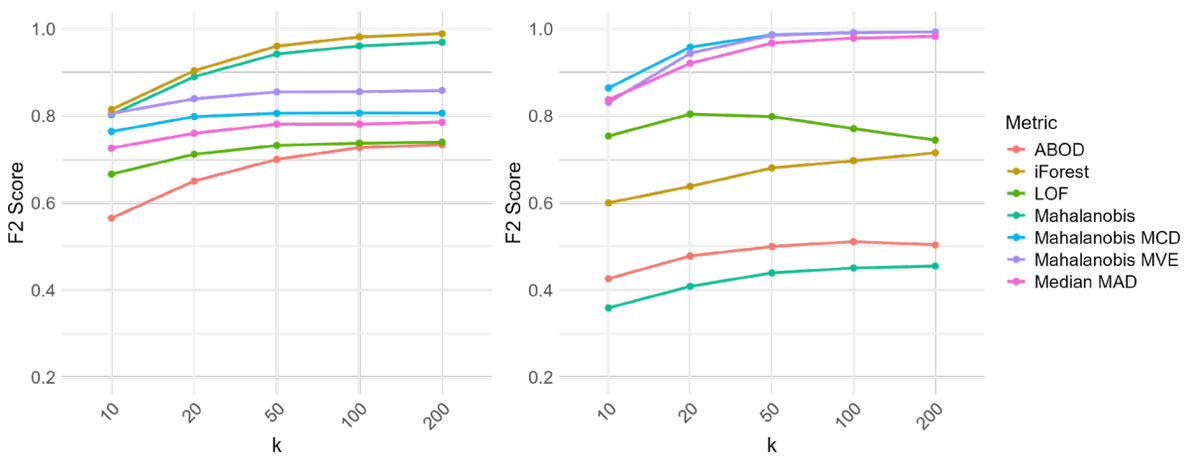
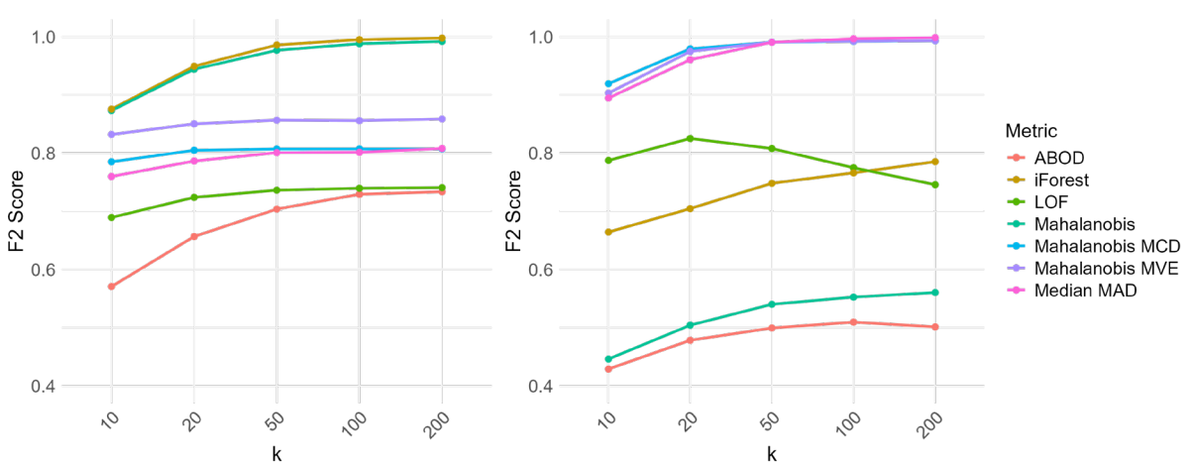


Figure 5 indicates that in four variable cases with 5 outliers, the Isolation Forest method has the highest F2 scores for all k values. 2nd highest scores are for the Mahalanobis method. For 20 outliers, the Mahalanobis MCD method has the highest scores for k values 10 and 20, and for high k values, Mahalanobis MCD and Mahalanobis MVE have the highest F2 scores.

Figure 6.

Five variables 5 outliers (left) 20 outliers (right).



For the 5 variable cases with 5 outliers, the Isolation Forest method has the highest scores, and the Mahalanobis method has 2nd highest scores for all k values. For 20 outliers, the Mahalanobis MCD, Mahalanobis MVE, and Median MAD have the highest scores (Fig. 7). Those highest scores are close to 1.

Conclusions

Detecting anomalies in laboratory processes is critical for maintaining the integrity and reliability of experimental results. Here, the goal was to find out which anomaly detection methods work well and how the performance of those methods varies for different numbers of outliers, different distances, and different outlier types. Various traditional multivariate outlier detection methods and modern machine learning methods were considered in this study.

The median MAD method consistently performs well in univariate scenarios, offering high F2 scores for lower and higher numbers of outliers, indicating that it is reasonable to use if the focus is on a single variable. For multivariate cases, the Mahalanobis MCD and Mahalanobis MCD methods tend to perform well, particularly in situations with many outliers. In a few circumstances, such as when there are a few shift outliers, the Isolation Forest method performs well, but, in general, none of the machine learning methods, designed as they are for large sample data mining situations, performed particularly well compared to the robust Mahalanobis methods in the situation studied. Data in quality control situations such as the one described above are such that (a) there are a moderate number of samples and a moderate number of variables and (b) a majority (if not all) of the samples are behaving in a consistent pattern reflecting a system that is operating properly for the most part – any observations that are not are considered anomalous and could possibly indicate a problem with the experimental procedure. In such cases, this study has shown that the Mahalanobis MVE method and the Mahala Nobis MCD method seem to work well, only lagging behind in some situations.

References

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *SIGMOD 2000 - Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104. https://doi.org/10.1145/342009.335388.

Hubert, M. & Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 36–43. https://doi.org/10.1002/wics.61.

Iglewicz, B. & Hoaglin, D. C. (1993). *Volume 16: How to Detect and Handle Outliers*. ASQ Quality Press. https://books.google.lk/books?id=FuuiEAAAQBAJ.

Kriegel, H. P., Schubert, M., & Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 444–452. https://doi.org/10.1145/1401890.1401946.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 413–422. https://doi.org/10.1109/ICDM.2008.17.

Reprint of: Mahalanobis, P.C. (1936) On the Generalised Distance in Statistics. (2018). *Sankhya A*, *80*(1), 1–7. https://doi.org/10.1007/s13171-019-00164-5.

Van Aelst, S. & Rousseeuw, P. (2009). Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 71–82. https://doi.org/10.1002/wics.19.