Data science for tax administration

Pijnenburg, M.G.F.

4

Singular Outliers: Finding Common Observations with an Uncommon Feature

In this chapter we focus on the research question:

> **Research Question** *Can we develop a new anomaly detection algorithm tailored to find the outliers that are of interest to a tax administration?*

Firstly, we looked carefully at the types of outliers that are of interest for tax administrations. This leads us to define the concept of *singular outliers*. Singular outliers are multivariate outliers that differ from conventional outliers by the fact that the anomalous values occur for only one feature (or a relatively small number of features). Subsequently, we developed a new algorithm (Singular Outlier Detection Algorithm – SODA) for detecting these outliers. The SODA algorithm is applied successfully to tax data. In order to obtain reproducible findings, the algorithm is applied as well to five public, real-world data sets and the outliers found by it are qualitatively and quantitatively compared to outliers found by three conventional outlier detection algorithms, showing the different nature of singular outliers. The chapter is based on the following article:

- M. Pijnenburg and W. Kowalczyk. Singular outliers: Finding common observations with an uncommon feature. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 492–503. Springer, 2018

4.1 Singular Outliers

Currently, many outlier detection algorithms exist [25]. However, when trying to find tax fraud, we found that hardly any of these is able to detect a particular type of outliers, which we will call *singular outliers*. In this chapter we will describe these singular outliers and present a newly developed algorithm (Singular Outlier Detection Algorithm – SODA) that can find these anomalies.

Roughly speaking, singular outliers are observations that show an anomalous value for *one* feature (or a relatively small set of features), while displaying common behavior on all other features. This feature with the anomalous value will be called the *discriminating feature*. The discriminating feature may be different for each outlier and finding the discriminating feature is part of the learning process. Singular outliers are typically overlooked by current outlier detection algorithms that prefer observations with anomalous values on as many features as possible.

To explain the concept of singular outliers more clearly, suppose we have a set of points $\mathbf{X}$ in a 5-dimensional vector space (with the standard metric) and a particular point $\mathbf{x} \in \mathbf{X}$, surrounded by its 20 nearest neighbors. Denote the mean vector of the 20 neighbors with $\mathbf{m}$, i.e., $\mathbf{m} = 1/20 \sum_1^{20} \mathbf{n}_i$. This allows calculating the absolute distances between $\mathbf{x}$ and $\mathbf{m}$ for each dimension. When visualized in a needle plot, the result may look as one of subplots of Figure 4.1. The subplot in the left shows large deviation from $\mathbf{m}$ in all dimensions. This observation can be labeled a conventional outlier. However, the right subplot shows an observation with small deviations to $\mathbf{m}$ on all but one dimension (dimension 4). This is what we will call a singular outlier.

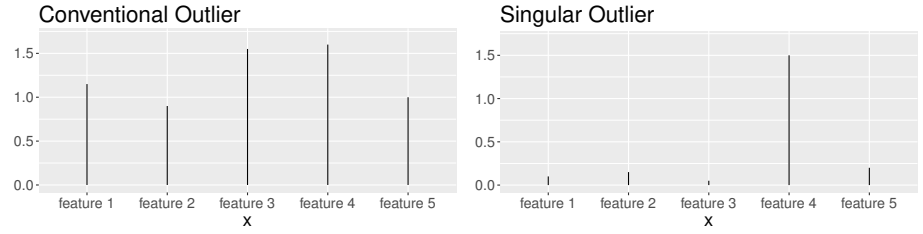

Figure 4.1: Needle plots of the absolute differences $|x_i - m_i|$ ($i \in 1, \ldots, 5$) of a point $\mathbf{x}$ and the center $\mathbf{m}$ of its 20 nearest neighbors.

Singular outliers can provide interesting insights in the field of fraud detection or data quality. We will illustrate this by our experiences at the Netherlands Tax and Customs Authority (NTCA), where we were asked to contribute to the selection of erroneous VAT tax returns for field audits. It turned out that tax returns containing many unusual values, are often less risky than tax returns containing only one (or two) anomalous fields. This apparent contradiction can be explained partly by efforts

of taxpayers to conceal tax evasion. But also by the existence of unconventional businesses with non-standard business models, that produce uncommon values on many fields of a tax declaration, without any increased risk of tax evasion. Moreover, it was noted at the NTCA that taxpayers sometimes unintentionally change two adjacent fields, leading to an example of singular outliers in the field of data quality. The examples in section 4.4 point to other application areas.

Singular outliers usually differ from those found by conventional outlier detection algorithms. Let us consider a real life example that will be more elaborate treated in section 4.4. In the example two algorithms, the newly developed SODA-algorithm and the popular LOF algorithm, [22] are used to find outliers in a data set containing the marks on 5 courses of 88 students[74]. Figure 4.2 shows two identical parallel coordinates plots (for more information on parallel coordinates plots see e.g., [23]) of this data set, but each plot highlights different outliers. The left plot highlights three singular outliers. We see that the singular outliers are students that have common marks on at least three subjects, but one or two exceptional marks. In contrast, the right subplot shows the same parallel coordinates plot but with three *conventional* outliers highlighted. These students have exceptional marks on all subjects (in this case exceptionally high). The singular outliers are interesting, since valuable insight might be gained in exploring the cause of the one (or two) exceptional mark(s) of the students.

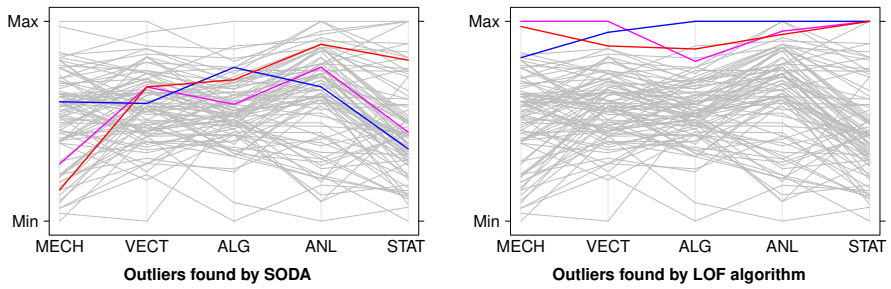

Figure 4.2: Two identical parallel coordinates plot of the Marks data set. The left plot highlights three observations that are singular outliers (detected by the SODA algorithm). The right plot highlights three conventional outliers (detected by the LOF algorithm). We see that the singular outliers have one or two features with exceptional values, whereas the conventional outliers show exceptional values on all features.

We will now define formally a singular outlier.

Definition 1 *An observation is called a* singular outlier *if there exits one feature (or a relatively small number of features), called the* discriminating feature(s) *such that the*

observation is an outlier when the discriminating feature is taken into account, but no outlier when the features are restricted to the non-discriminating features.

The purpose of this chapter is to point to the class of singular outliers and present an algorithm for finding them. The chapter is organized as follows. Section 4.2 gives a short overview of classes of outlier detection algorithms. Subsequently Section 4.3 describes the SODA algorithm for finding singular outliers and explains the experimental setup to test the algorithm on five public data sets. Section 4.4 contains the results of the experiments. Finally, Section 4.5 contains the conclusions of the chapter and a discussion of the results.

4.2 Related Work

Chandola, Banerjee, and Kumar [25] present an overview of commonly used outlier detection techniques. They distinguish five classes of unsupervised outlier detection algorithms: nearest neighbor-based algorithms (including density based approaches), clustering-based algorithms, statistical algorithms, information theoretic algorithms, and spectral algorithms. Goldstein and Uchida [43] present a similar categorization.

Nearest neighbor and clustering-based outlier detection are the most used categories in practice, according to Goldstein and Uchida. Of these two categories, nearest-neighbor based algorithms perform better in most cases [43]. Moreover it is useful to split the class of nearest neighbor algorithms in two subclasses: *local* algorithms and *global* algorithms. In the remaining categories outlier detection algorithms, the statistical algorithm HBOS (Histogram-based Outlier Score) performs remarkably well in experiments [43].

We will briefly describe three algorithms that we will compare with SODA: Local Outlier Factor, k^{th}-Nearest Neighbor and HBOS.

The Local Outlier Factor (LOF) is introduced by Breunig et al. [22]. The basic idea is to compare the density of an observation $\mathbf{x}$ with the density of its k closest neighbors $N^k(\mathbf{x})$. If we ignore a small subtlety that may arise if the k^{th}-neighbor of a point is not unique, the LOF score can be calculated simply by,

$$LOF_k(\mathbf{x}) = \frac{1}{k} \sum_{o \in N^k(\mathbf{x})} \frac{d_{Eucl}(\mathbf{x}, \mathbf{x}_{(k)})}{d_{Eucl}(\mathbf{o}, \mathbf{o}_{(k)})}, \tag{4.1}$$

where $d_{Eucl}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ is the Euclidean distance.

The k^{th} nearest neighbors outlier detection algorithm is straightforward: the distance to the k^{th} neighbor is used as an anomaly score [96]. By taking the distance to the k^{th} neighbor instead of the average distance to the k^{th} nearest neighbors, the algorithm is better able to detect a small cluster of outliers.

The Histogram Based Outlier Score [43] starts with making a histogram for each feature in the data set. Then, for each observation, the height of the bins it resides are multiplied. This will result in a positive number. Subsequently the negative of this number is taken as an outlier score. In the experiments, the frequently used Sturges' formula is applied to compute the number of bins.

4.3 Singular Outlier Detection Algorithm

4.3.1 The Algorithm

We propose an algorithm called SODA (Singular Outlier Detection Algorithm) to find singular outliers. The algorithm involves several steps which are specified in Algorithm 1. The input of the algorithm is a data set with n observations and p numeric features as well as a parameter k that specifies the number of nearest neighbors. The output of the algorithm is an outlier score for each observation; large scores represent outliers. Additionally, the discriminating feature for each observation is given as output. The latter is a convenient starting point when manual inspection of outliers is required.

In the first step, for each observation $\mathbf{x_i}$, $i = 1, \ldots, n$, its k neighbors N_i^k are found using the Manhattan metric (also called the 'distance based on the L_1 norm'). Subsequently, the center of these neighbors, $\mathbf{m_i}$, is determined as the trimmed mean, i.e., the mean calculated after removing the largest value and the smallest value along each dimension. These extreme values are ignored to limit their impact on $\mathbf{m_i}$.

As a next step, the Euclidean distance $d_{Eucl}(\mathbf{x_i}, \mathbf{m_i})$ as well as the Manhattan distance $d_{Manh}(\mathbf{x_i}, \mathbf{m_i})$ are calculated. We will call the ratio of these two distances LEMR (*Local Euclidean Manhattan Ratio*), i.e.:

$$\text{LEMR}(\mathbf{x_i}) = \frac{d_{Eucl}(\mathbf{x_i}, \mathbf{m_i})}{d_{Manh}(\mathbf{x_i}, \mathbf{m_i})}. \tag{4.2}$$

This ratio is the 'singular outlier score' for observation $\mathbf{x_i}$. It is an indicator of the 'spread' of the values of the vector $\mathbf{v_i} = \|\mathbf{x_i} - \mathbf{m_i}\|$. The left subplot of Figure 4.3 displays a 2-dimensional example with the Euclidean unit sphere (i.e., the circle) as well as the Manhattan unit sphere (the square: all points with a Manhattan distance 1 to the origin). It is clear from the figure that points on the coordinate axes (i.e., singular outliers) are in both spheres, so the ratio of the Euclidean to the Manhattan distance is 1. On the other hand, points that are on the diagonals (the opposite of singular outliers) have the largest difference in Euclidean and Manhattan distance to the origin and consequently have a low LEMR value. The right subplot of Figure 4.3 shows the LEMR value in this 2-dimensional example.

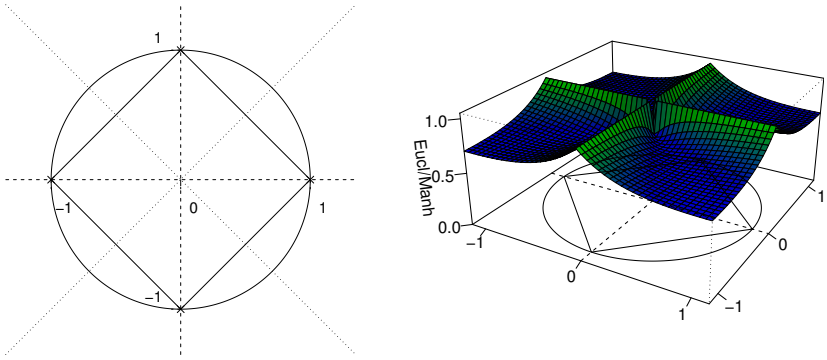

Figure 4.3: Left: the Euclidean unit sphere and the Manhattan unit sphere in 2 dimensions. Points on coordinate axes (singular outliers) have a maximal value of the Local Euclidean Manhattan Ratio (LEMR = 1). Points on the diagonals (i.e., conventional outliers) have a minimal LEMR value (LEMR $=1/\sqrt{2}$), see subplot right. In more dimensions the difference between the Euclidean and the Manhattan distance is more pronounced.

The value of LEMR cannot exceed 1 since the Euclidean length of a vector $\|\mathbf{v}\|$ is always smaller or equal to the L_1-norm $\|\mathbf{v}\|_1 = \sum |v_j|$. The LERM value of 1 occurs when all but one components of $\mathbf{v}$ equal 0. The value of LEMR is a minimum in case the elements of $\mathbf{v}$ are all of equal length (and unequal 0). The minimal value depends on the number of features p and is given by $1/\sqrt{p}$.

The calculation of the nearest neighbors in the first step of the algorithm is performed using the Manhattan distance in contrast to the more frequently used Euclidean distance. The reason is that the discriminating feature of a singular outlier ideally deviates strongly from other observations. This deviation may however be so large that it can have too big influence on the determination of the nearest neighbors. In the worst case, this would lead to neighbors that only share a similar value for the discriminating feature. To diminish the effect, we prefer the Manhattan distance that is less influenced by extreme values in one feature.

The algorithm has one parameter k. A (too) small value of k will lead to observations $\mathbf{x}$ with very few neighbors. Consequently, it might result in outliers whose non-discriminating features have a risk of being not so common. A (too) large value of k usually has less severe consequences. However, in theory it can lead to bad results if the data points are gathered in many small clusters and k exceeds the cluster size. For instance this may happen if the data contains binary features. In our experiments we used $k = n/5$.

The time complexity of the singular outlier detection algorithm is $\mathcal{O}(n^2pk)$.

Algorithm 1: singular Outlier Detection Algorithm (SODA)

Input : A data set $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots\}$ with p (numeric) features,
k the number of nearest neighbors

Output: 1) outlier score. The observations with the largest scores represent the singular outliers,
2) discriminating feature for each $\mathbf{x}_i$. This indicates the feature that contains the anomalous value.

1 Find the k nearest neighbors N_i^k for each observation $\mathbf{x_i}$ in $\mathbf{X}$ based on the Manhattan distance d_{Manh}.
2 **for** *each observation (row)* $\mathbf{x}_i$ *in* $\mathbf{X}$ **do**
3 Compute the vector of trimmed means $\mathbf{m}_i$ of the neighbors N_i^k:
$\mathbf{m_i} = \underset{\mathbf{x}_j \in N_i^k}{\text{mean}_{trimmed}}\{\mathbf{x}_j\}$
4 Compute
5 $d_{Eucl}(\mathbf{x}_i, \mathbf{m}_i) = \sqrt{\sum_{s=1}^{p}(x^s - m^s)^2}$
6 $d_{Manh}(\mathbf{x}_i, \mathbf{m}_i) = \sum_{s=1}^{p}|x^s - m^s|$
7 outlier score$(\mathbf{x}_i) = LEMR(\mathbf{x}_i) = \frac{d_{Eucl}(\mathbf{x}_i, \mathbf{m}_i)}{d_{Manh}(\mathbf{x}_i, \mathbf{m}_i)}$
8 discriminating feature$(\mathbf{x}_i) = \underset{s \in \{1,\ldots,p\}}{\text{argmax}}\,|x^s - m^s|$
9 **end**

4.3.2 Measurement of Characteristics of Singular Outliers

To determine whether an outlier can be called *singular*, two characteristics must hold. These two characteristics follow from definition 1 and are listed below.

No outlier with respect to non-discriminating features When the discriminating feature is removed, the observation stops to be an outlier.

Outlier with the discriminating feature The observation must become an outlier when the discriminating feature(s) is taken into account.

To quantify these two qualitative characteristics, we will formulate two measures. These measures will be applied in section 4.4, Table 4.2 to see whether the outliers found by SODA can be called singular outliers.

The first characteristic is measured for an outlier by removing the discriminating feature from the data set and then calculate an outlier score. This outlier score must be low for the characteristic to be valid. In principle any outlier score can be used. In this chapter the LOF score is chosen, since it is a well established outlier score that performs well on a broad variety of data sets [43]. A LOF score around 1 will be accepted as a sign that the first characteristic is applicable.

The second characteristic is measured for an outlier by computing an outlier score with and without the discriminating feature. Subsequently the ratio of the outlier scores is taken as a measure. Again, in principle any outlier score can be taken, but in this chapter we choose the LOF score. Hence, the measure of the second characteristic becomes $\frac{\text{LOF}_{all}(\mathbf{x})}{\text{LOF}_{w.o.discr}(\mathbf{x})}$. A large ratio signifies that the addition of the discriminating feature has substantially increased the outlier score (LOF) and is taken as a sign that the second characteristic is applicable.

In section 4.4 the outliers found by SODA will be compared with outliers of the three conventional outlier algorithms mentioned in section 4.2. In order to apply the two measurements above, we have to determine the discriminating feature for outliers found by these algorithms as well. For the algorithms based on nearest neighbors (LOF and k^{th}NN) a natural choice is to take:

$$\text{discriminating feature}(\mathbf{x}) = \underset{s\in\{1,\ldots,p\}}{\text{argmax}} |x^s - x^s_{(k)}|, \tag{4.3}$$

where $x^s_{(k)}$ is the s^{th} feature of the k-nearest neighbor of $\mathbf{x}$. For the HBOS algorithm we take the feature with the lowest bin height as the most discriminating feature.

4.4 Comparison

This section compares the outliers found by SODA with outliers detected by three conventional outlier detection algorithms (mentioned in section 4.2) to five real world data sets, that will be described shortly. The comparison consists of parallel coordinates plots and the two measurements described in section 4.3.2.

The selected data sets are listed in Table 4.1, along with some key properties. One of these properties is the value k that is used in the SODA algorithm and the two conventional outlier detection algorithms based on nearest neighbors (LOF and k^{th}NN). In the experiments, k is set to (approximately) one-fifth of the number of observations, see section 4.3.1.

data set name	# rows	# columns	k
Marks	88	5	20
Istanbul Stock Exchange	536	9	100
Wholesale Customers	440	6	80
Polish Bankruptcy	5,907	5	1,000
Algae	306	8	60

Table 4.1: Data sets used in the singular outlier experiments with key indicators.

Measurement 1 - no outlier without discriminating feature: LOF_{-d}					
	data sets				
	Marks	Istanbul	Wholesale	Polish	Algae
algorithm					
SODA	1.02	0.99	1.16	1.01	1.01
LOF	1.45	2.57	2.98	147.7	3.84
k^{th}NN	1.42	2.56	3.69	147.7	4.20
HBOS	1.39	2.57	4.02	122.0	3.28
Measurement 2 - outlier with discriminating feature: $\frac{\text{LOF}_{all}}{\text{LOF}_{-d}}$					
	data sets				
	Marks	Istanbul	Wholesale	Polish	Algae
algorithm					
SODA	1.09	1.28	2.31	5.75	1.99
LOF	1.04	1.08	2.11	1.46	1.73
k^{th}NN	1.02	1.04	1.60	1.46	1.27
HBOS	1.06	1.06	1.06	1.13	1.08

Table 4.2: Measurements of the two characteristics of singular outliers (see section 4.3.2) for the outliers selected by SODA, LOF, k^{th}NN and HBOS on five data sets. Each algorithm is allowed to pick 3 outliers and the values reported are averages over these three outliers. The first measurement computes an outlier score (LOF) without the discriminating feature. The second measurement computes the ratio of the LOF score with all features and the LOF score without the discriminating feature. Singular outliers are characterized by a value close to 1 for measurement 1 and a large value for measurement 2. We see that the outliers selected by SODA can be termed 'singular outliers' and differ on these characteristics from the outliers found by conventional outlier detection algorithms.

4.4.1 Marks Data

The first data set, *Marks*, comes from Mardia, Kent and Bibby [74] and consists of the examination marks of 88 students in the five subjects mechanics, vectors, algebra, analysis and statistics. We used these unstandardized marks.

The parallel coordinates plot of this data set is displayed in the introduction, see Figure 4.2. Clearly, the SODA algorithm picks up students that perform around average for most subjects, except for one subject. In contrast, the LOF algorithm selects students that perform exceptionally well on all subjects. The measurements of Table 4.2 confirm that the SODA outliers can be called singular outliers: the average value for the SODA outliers on measurement 1 equals 1.02, which is close to 1 and much

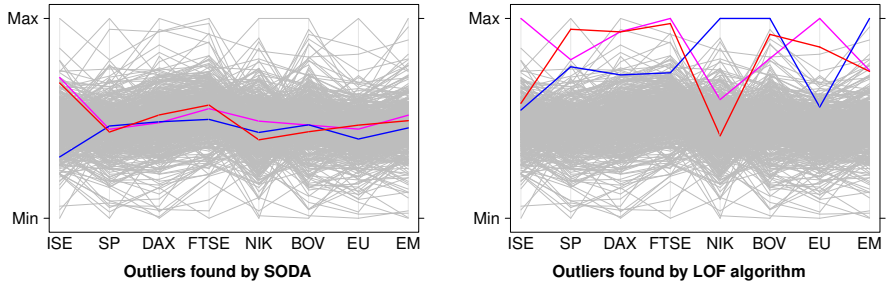

Figure 4.4: Parallel coordinates plot for the Istanbul Stock Exchange data set with the three largest outliers highlighted.

lower compared to the values of other algorithms. The average value on the second measurement (1.09) is not very large, but at least larger than the values for the outliers found by the other algorithms.

4.4.2 Istanbul Stock Exchange Data

The Istanbul Stock Exchange data set [3] displays the daily increase in eight major stock exchange indices for the period January 5 2009 to February 22, 2011. The data are already standardized.

The parallel coordinates plots of Figure 4.4 show that the SODA algorithm has picked three days where the average stock return is common (0.3 %), but the Istanbul stock exchange performed deviantly. In contrast, the LOF algorithm has selected three days on which most stock indices got exceptional good returns (average return of these days was 4.1 %, average of all days is less than 0.1 %). Table 4.2 confirms the differences of the outliers found by SODA and the three conventional algorithms.

4.4.3 Wholesale Customers Data

The Wholesale Customers data set [1] contains the annual spending on six product categories for 440 customers of a wholesaler in Portugal. These not standardized spending amounts were input for the algorithms. It is clear from the parallel coordinates plots in Figure 4.5 that the SODA algorithm selects customers with average sales on all product categories, but one. The LOF algorithm selects customers with exceptional high purchases on at least three product categories.

The statement that SODA picks out singular outliers is confirmed by looking at the measurements of Table 4.2. However, some experts might be inclined to call the red and / or pink line of the LOF algorithm singular outliers as well. See Section 4.5 for

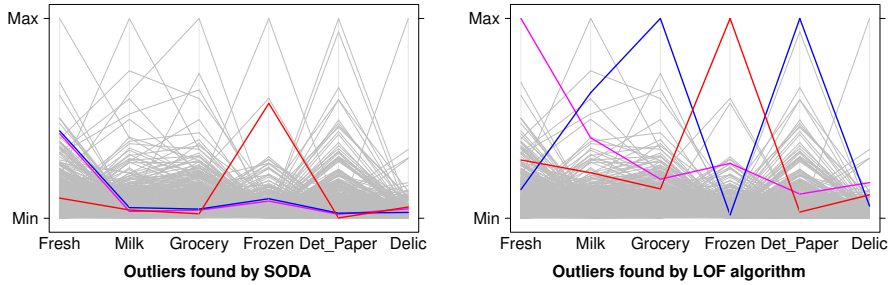

Figure 4.5: Parallel coordinates plot for the Wholesale Customers data set with the three largest outliers highlighted.

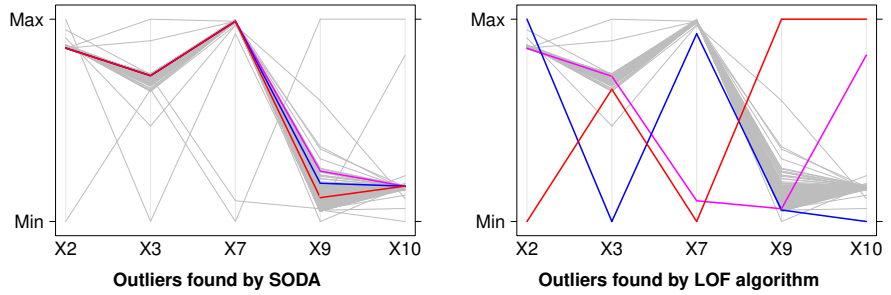

Figure 4.6: Parallel coordinates plot for the Polish Bankruptcy data set with the three largest outliers highlighted.

a discussion of this aspect.

4.4.4 Polish Bankruptcy Data

The Polish Bankruptcy data contains financial information of Polish companies and was originally used to model the probability of a bankruptcy [117]. We took the 'five year set' and standardized the ratios. We selected five financial ratios that correspond to major economic indicators of the company: X2 (total liabilities / total assets), X3 (working capital / total assets), X7 (Earnings Before Interest and Taxes / total assets), X9 (sales / total assets), and X10 (equity / total assets). After removing observations with missing values, 5,907 observations remain.

Comparing the outliers found by SODA and LOF in the parallel coordinates plots of Figure 4.6, we see that the SODA outliers show more common behavior than the LOF outliers. It is not clear from the figure whether the SODA outliers have at least one anomalous value. The latter is confirmed by Table 4.2. The first measurement shows

an average LOF value of 1.01 for the SODA outliers, indicating that the observations are not considered outliers without the discriminating feature. Simultaneously, measurement 2 shows that the average LOF score increases by a factor of over 5 when the discriminating feature is added. However, if one is interested in two-feature singular outliers, one may find the blue and pink line of the LOF plot interesting. See Section 4.5 for a discussion on this aspect.

4.4.5 Algae Data

The data set *Algae*, available via the UCI repository [65], comes from a water quality study where samples were taken from sites on different European rivers of a period of approximately one year. We used the (standardized) concentrations of 8 chemical substances from these samples as features. We exclude the additional measures on different algae populations. After removing rows with missing data, 306 observations and 8 features remain.

The parallel coordinates plots of Figure 4.7 show that the SODA outliers have one chemical substance that has an unusual value. One of these outliers is picked up by the LOF algorithm as well. The other two LOF outliers show unusual values for three chemical characteristics. Table 4.2 confirms the on average different nature of the outliers selected by SODA and the three conventional algorithms. If one is interested in three-feature outliers, the blue and pink lines of the LOF plot are interesting observations as well. See Section 4.5 for a discussion on this aspect.

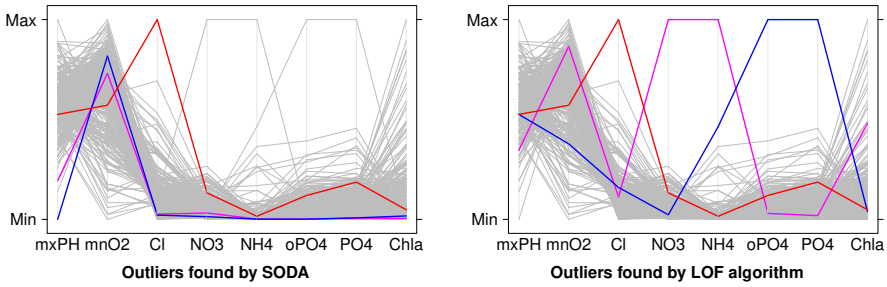

Figure 4.7: Parallel coordinates plot for the Algae Data with the three largest outliers highlighted.

4.5 Conclusion and Discussion

In this chapter, we introduced the concept of a *singular outlier* and pointed to its usefulness in the contexts of fraud detection, data quality and other areas. We introduced

an algorithm (SODA) to detect these outliers, based on the Local Euclidean Manhattan Ratio (LEMR). The algorithm has been applied to five publicly available data sets. The parallel coordinates plots, as well as the the results shown in Table 4.2, confirm that the SODA algorithm is suited for finding singular outliers.

Although Table 4.2 points out that the SODA algorithm finds observations that better match the definition of singular outliers when one restricts oneself to *one* discriminating feature, the parallel plots of the latter three data sets show that there might be interesting outliers that have two discriminating features. These observations may not get the highest outlier scores for SODA, despite the fact that the deviations in the discriminating features may be large. In such a situation, the SODA algorithm may be adapted to focus more on two-feature singular outliers or to take the absolute value of the deviations into account. This might be a fruitful aspect for further research. Part of such research can be to adjust the Local Euclidean Manhattan Ratio. The Euclidean and the Manhattan distances are based on the L_p norm. Other ratios can be constructed by taking other values for p instead of 2 and 1.

Another interesting path for further research might be to view the detection of singular outliers as an optimization problem where simultaneously attention is given to maximize dissimilarity (discriminating feature) as well as maximizing similarity (non-discriminating features).

Finally, singular outliers might be found with an algorithm with a lower time complexity than SODA. The time complexity of the SODA algorithm is dominated by finding the nearest neighbors of each observation. The nearest neighbors component gives a local flavor to SODA that might be beneficial for data sets that possess local structures (like separate clusters). However, for data sets that do not display this property, comparison of an observation $\mathbf{x}$ to the global center $\mathbf{m}_{global}$ might be used. This reduces computational time considerably.