Chapter 5

HIGH-DIMENSIONAL OUTLIER DETECTION: THE SUBSPACE METHOD

"In view of all that we have said in the foregoing sections, the many obstacles we appear to have surmounted, what casts the pall over our victory celebration? It is the curse of dimensionality, a malediction that has plagued the scientist from the earliest days."– Richard Bellman

1. Introduction

Many real data sets are very high dimensional. In some scenarios, real data sets may contain hundreds or thousands of dimensions. With increasing dimensionality, many of the conventional outlier detection methods do not work very effectively. This is an artifact of the well known *curse of dimensionality*. In high-dimensional space, the data becomes sparse, and the true outliers become masked by the noise effects of multiple dimensions, when analyzed in *full dimensionality*.

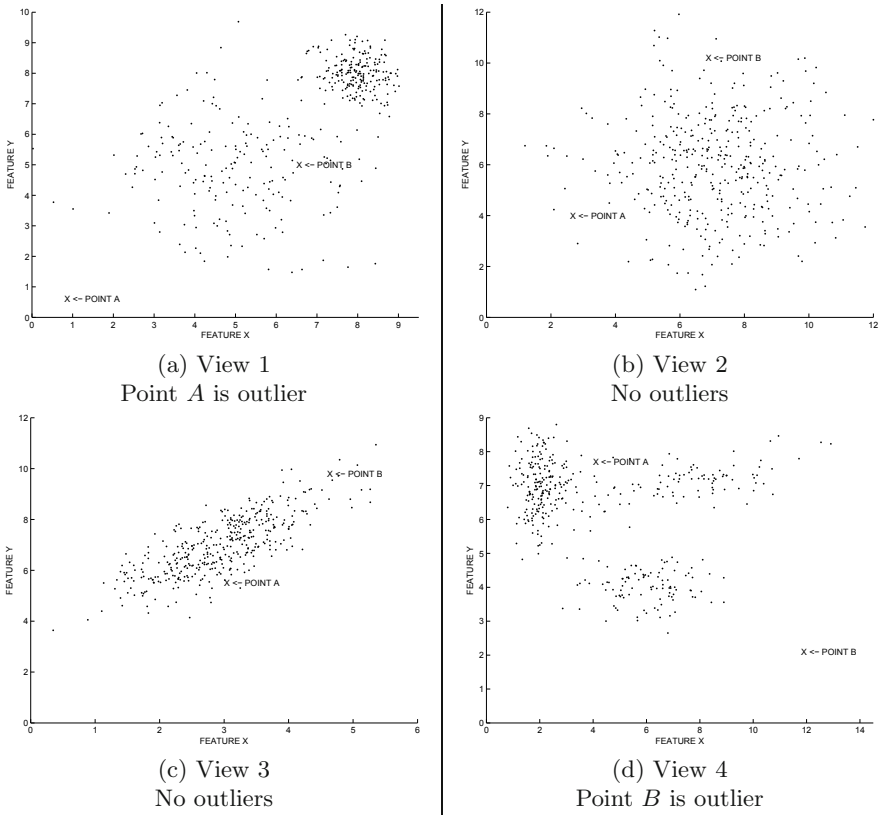
A main cause of the dimensionality curse is the difficulty in defining locality for the high dimensional case. For example, proximity-based methods define locality with the use of distance functions. On the other hand, it has been shown in [65, 215], that all pairs of points are almost equidistant in high-dimensional space. This is referred to as *data sparsity*. Since outliers are defined as data points in sparse regions, this results in a poorly discriminative situation where all data points are situated in an almost equally sparse regions in full dimensionality. The challenges arising from the dimensionality curse are not specific to outlier detection. It is well known that many problems such as clustering and similarity search experience qualitative challenges with increasing

C.C. Aggarwal, *Outlier Analysis*, DOI 10.1007/978-1-4614-6396-2_5,

dimensionality [5, 7, 95, 215]. In fact, it has been suggested that almost any algorithm which is based on the notion of proximity would degrade qualitatively in higher dimensional space, and would therefore need to re-defined in a more meaningful way [8]. The impact of the dimensionality curse on the outlier detection problem was first noted in [4].

In order to further explain the causes of the ineffectiveness of full dimensional outlier analysis algorithms, a motivating example will be presented. In Figure 5.1, four different 2-dimensional views of a hypothetical data set have been illustrated. Each of these views corresponds to a disjoint set of dimensions. It is evident that point A is exposed as an outlier in the first view of the data set, whereas point B is exposed as an outlier in the fourth view of the data set. However, neither of the data points A and B are exposed as outliers in the second and third views of the data set. These views are therefore *noisy* from the perspective of measuring the outlierness of A and B. In this case, three of the four views are quite non-informative and noisy for exposing any *particular* outlier A or B. In such cases, the outliers are lost in the random distributions within these views, when the distance measurements are performed in *full* dimensionality. This situation is often naturally magnified with increasing dimensionality. For data sets of very high dimensionality, it is possible that only a very small fraction of the views may be informative for the outlier analysis process.

What does the aforementioned pictorial illustration tell us about the issue of locally relevant dimensions? The physical interpretation of this situation is quite intuitive in practical scenarios. An object may have several measured quantities, and significantly abnormal behavior of this object may be reflected only in a small subset of these quantities. For example, in an airplane mechanical fault detection scenario, the results of thousands of different airframe tests on the same plane may mostly be normal, with some noisy variations, which are not significant. On the other hand, some deviations in a small subset of tests may be significant enough to be indicative of anomalous behavior. When the data from the tests are represented in full dimensionality, the anomalous data points will not appear significant in virtually all views of the data, except for a very small fraction of the dimensions. Therefore, aggregate proximity measures are unlikely to expose the outliers, since the noisy variations of the vast number of normal tests will mask the outliers. Furthermore, when different objects (instances of different airframes) are tested, then different tests (subsets of dimensions) may be relevant to finding the outliers, which emphasizes the *local* nature of the relevance.



(a) View 1
Point A is outlier

(b) View 2
No outliers

(c) View 3
No outliers

(d) View 4
Point B is outlier

Figure 5.1. The outlier behavior may be lost in a majority of randomly chosen subspaces in the high dimensional case.

What does this mean for full-dimensional analysis in such scenarios? When full-dimensional distances are used in order to measure deviations, the dilution effects of the vast number of "normally noisy" dimensions will make the detection of outliers difficult. In most cases, this will show up as concentration effects in the distances, from the noise in the other dimensions. This may make the computations more erroneous. Furthermore, the additive effects of the noise present in the large number of different dimensions will interfere with the detection of actual deviations. Simply speaking, *outliers are lost in low-dimensional subspaces, when full-dimensional analysis is used, because of the masking and dilution effects of the noise in full dimensional computations* [4].

Similar effects are also experienced for other distance-based methods such as clustering and similarity search. For these problems, it has been shown [5, 7, 215] that by examining the behavior of the data in subspaces, it is possible to design more meaningful clusters which are specific to the particular subspace in question. This broad observation is generally true of the outlier detection problem as well. Since the outliers may only be discovered in low dimensional subspaces of the data, it makes sense to explore the lower dimensional subspaces for deviations of interest. Such an approach filters out the additive noise effects of the large number of dimensions, and results in more robust outliers.

Such a problem is very challenging to address effectively. This is because the number of possible projections of high dimensional data is exponentially related to the dimensionality of the data. The problem of outlier detection is like finding a needle in a haystack, *even when we know* the relevant dimensions of interest. Being forced to determine the relevant subsets of dimensions *in addition to this challenge* is equivalent to suggesting that even the haystack of interest is hidden in an exponential number of possible haystacks. An important observation is that subspace analysis in the context of the outlier detection problem is generally more difficult than in the case for problems such as clustering, which are based on aggregate behavior. This is because outliers, by definition, are rare, and therefore statistical aggregates on individual dimensions in a given locality often provide *very weak* hints for the subspace exploration process as compared to aggregation-based methods such as clustering. When such weak hints result in the omission of relevant dimensions, the effects can be much more drastic than the inclusion of irrelevant dimensions, especially in the interesting cases when the number of locally relevant dimensions is a small fraction of the full data dimensionality. A common mistake is to assume that the complementarity relationship between clustering and outlier analysis can be extended to the problem of local subspace selection. In particular, blind adaptations of dimension

selection methods from earlier subspace clustering methods, which are unaware of the nuances of subspace analysis principles across different problems, may sometimes miss important outliers. In this context, it is also crucial to recognize the difficulty in identifying relevant subspaces for outlier analysis, and use robust methods which combine the results from different subspaces.

An effective outlier detection method would need to search the data points and dimensions in *an integrated way*, so as to reveal the most relevant outliers. This is because different subsets of dimensions may be relevant to different outliers, as is evident from the example in Figure 5.1. The integration of point and subspace exploration leads to a further expansion in the number of possibilities which need to be examined for outlier analysis. This chapter will focus on subspace exploration methods, which attempt to find the relevant outliers by sifting through different subsets of dimensions in the data in an ordered way. This is accomplished simultaneously with a data-specific evaluation process, so that relevant data points are reported as outliers without having to explore all the subspaces in an exhaustive way. The idea is to determine the relevant subsets of dimensions in which the most important outliers are revealed as quickly as possible. This model is referred to as *projected outlier detection* [4]. Correspondingly, this chapter will present a number of algorithms, which achieve this goal.

Several classes of methods are commonly used in order to discover the relevant subspaces:

- **Rarity-based:** These methods attempt to discover the subspaces based on rarity of the underlying distribution. The major challenge here is computational, since the number of rare subspaces is far larger than the number of dense subspaces in high dimensionality.

- **Unbiased:** In these methods, the subspaces are sampled in an unbiased way, and scores are combined across different subspaces.

- **Aggregation-based:** In these methods, aggregate statistics such as cluster statistics, variance statistics, or non-uniformity statistics of local or global subsets of the data are used in order to determine the relevance of subspaces. Note that the difference from rarity-based statistics, is that instead of trying to determine the *number of data points* in a pre-specified local subspace, these methods typically analyze the statistical distributions of pre-specified local or global reference sets of points. Since such methods use statistics over local or global *subsets* of the data, it provides some *hints* for relevant subspaces for exploration. However, since such hints

are weak, and are not guaranteed to be the correct ones, multiple subspace sampling is crucial.

This chapter is organized as follows. Evolutionary algorithms for outlier detection are discussed in section 2. These algorithms are based on a grid-based approach for defining outliers. Distance-based methods for subspace outlier detection are studied in section 3. Methods for using and combining multiple subspaces in order to determine relevant outliers are discussed in section 4. The problem of determining outliers in generalized subspaces is discussed in section 5. The limitations of subspace analysis are discussed in section 6. The conclusions and summary are presented in section 7.

2. Projected Outliers with Grids

A first approach to projected outlier detection was presented in [4]. Projected outliers are determined by finding *localized regions of the data in low dimensional space*, which have abnormally low density. Thus, the first step is to identify and mine those *localized* patterns which contain data points, but have abnormally low density. Thus, the goal is to determine interesting anomalies, rather than the noise in the data. Once such localized regions have been identified, then the outliers are defined as those records which have such patterns present in them. An interesting observation is that such lower dimensional projections can be determined even in data sets with missing attribute values. This is quite useful for many real applications, in which feature extraction is a difficult process and full feature descriptions often do not exist. For example, in the airframe fault detection scenario introduced earlier in this chapter, it is possible that only a subset of tests may have been applied, and therefore the values in only a subset of the dimensions may be available for outlier analysis.

2.1 Defining Abnormal Lower Dimensional Projections

In order to find such abnormal lower dimensional projections, it is important to provide a proper statistical definition of an abnormal lower dimensional projection. An abnormal lower dimensional projection is one in which the density of the data is exceptionally lower than average. In this context, the methods for extreme value analysis introduced in Chapter 2 are useful.

A grid-based approach is used in order to determine projections of interest. The first step is to perform a grid discretization of the data. Each attribute of the data is divided into ϕ ranges. These ranges are

created on an equi-depth basis. Thus, each range contains a fraction $f = 1/\phi$ of the records. The reason for using equi-depth ranges as opposed to equi-width ranges is that different localities of the data have different densities. Therefore, such an approach partially adjusts for the local variations in data density during the initial phase. These ranges form the units of locality which are used in order to define low dimensional projections which have unreasonably sparse regions.

Consider a k-dimensional cube which is created by picking grid ranges from k different dimensions. The expected fraction of the records in that region is equal to f^k, if the attributes were statistically independent. Of course, the data is far from statistically independent and therefore the actual distribution of points in a cube would differ significantly from average behavior. Many of the local regions may contain very few data points, if any. It is precisely these abnormally sparse regions, which are useful for the purpose of outlier detection.

It is assumed that the total number of points in the database is denoted by N. Under the afore-mentioned independence assumption, the presence or absence of any point in a k-dimensional cube is a bernoulli random variable with probability f^k. Then, the expected fraction and standard deviation of the points in a a k-dimensional cube is given by $N \cdot f^k$ and $\sqrt{N \cdot f^k \cdot (1 - f^k)}$. Furthermore, if the number of data points N is large, then the central limit theorem can be used to *approximate* the number of points in a cube by a normal distribution. Let $n(\mathcal{D})$ be the number of points in a k-dimensional cube $\mathcal{D}$. The sparsity coefficient $S(\mathcal{D})$ of the data set $\mathcal{D}$ can be computed as follows:

$$S(\mathcal{D}) = \frac{n(D) - N \cdot f^k}{\sqrt{N \cdot f^k \cdot (1 - f^k)}}$$

Only sparsity coefficients which are negative are indicative of local projected regions, in which the presence of the points is significantly lower than expected. Since $n(\mathcal{D})$ is assumed to fit a normal distribution, the normal distribution tables can be used to quantify the probabilistic level of significance of its deviation. Of course, while the independence assumption is almost never completely true, it provides a good heuristic for determining the level of abnormality of the underlying data points in practice.

2.2 Evolutionary Algorithms for Outlier Detection

It is evident from the discussion in the introduction, that an exhaustive search of all the subspaces in the data for outliers is unlikely to be

fruitful, because of high computational complexity. Therefore, an ordered search method is required, which prunes off most of the subspaces automatically during the exploration process. Since the search space is noisy and unstructured in this case, this is a natural candidate for the use of evolutionary algorithms.

The nature of this problem is such that there are no upward or downward-closed properties on the grid-based subspaces satisfying the sparsity condition.[1] Unlike problems such as frequent pattern mining [28] where one is looking for large aggregate patterns, the problem of finding subsets of dimensions which are sparsely populated has the flavor of finding a needle in haystack. Furthermore, it may often be the case that even though particular regions may be well populated on certain sets of dimensions, they may be very sparsely populated when such dimensions are combined together. For example, in a given data set, there may be a large number of individuals clustered at the age of 20 (low local variance), and a modest number of individuals with varying levels of diabetes (modest local variance). However, *very rare* individuals would satisfy both criteria, because the disease does not affect young individuals. From the perspective of outlier detection, a 20-year old with diabetes is a very interesting record. However, the interestingness of the pattern is not even hinted at by its lower dimensional projections, or the relative variances in these individual projections. Therefore, the best projections are often created by an unknown combination of dimensions, whose lower dimensional projections may contain very few hints for proper subspace exploration. One solution is to change the measure in order to force better closure or pruning properties; however this can worsen the quality of the solution substantially by forcing the choice of the measure to be driven by algorithmic considerations. In general, it is not possible to predict the behavior of the data when two sets of dimensions are combined. Therefore, a natural option is to develop search methods which can identify such hidden combinations of dimensions. In order to search the exponentially increasing space of possible projections, the work in [4] borrows ideas from a class of evolutionary search methods in order to reduce the size of the search space.

Evolutionary Algorithms [223] are methods which imitate the process of organic evolution [125] in order to solve parameter optimization problems. In evolutionary methods, every solution to an optimization problem can be disguised as an individual in an evolutionary system. The

[1] An upward closed pattern is one in which all supersets of the pattern are also valid patterns. A downward closed set of patterns is one in which all subsets of the pattern are also members of the set.

measure of fitness of this "individual" is equal to the objective function value of the corresponding solution, and the other species which this individual has to compete with are a group of other solutions to the problems. Appropriate operations are defined in order to imitate the recombination and mutation processes as well, and the simulation is complete. Each feasible solution is encoded in the form of a string and is the chromosome representation of the solution. The process of conversion of feasible solutions of the problem into strings which the algorithm can use is referred to as its *encoding*. The measure of fitness of a string is evaluated by the *fitness function*. This is equivalent to the objective function value of the solution. The better the objective function value, the better the fitness value. As the process of evolution progresses, all the individuals in the population typically improve in fitness and also become more similar to each other. Dejong [134] defined convergence of a particular position in the string, as the the stage at which 95% of the population had the same value for that gene. The population is said to have converged when all positions in the string representation have converged.

The relevant localized subspace patterns can be easily represented as strings. Let us assume that the grid range for the ith dimension is denoted by m_i. Then, the value of m_i can take on any of the values 1 through ϕ, or it can take on the value $*$, which denotes a "don't care". Thus, there are a total of $\phi + 1$ values that the dimension m_i can take on. Thus, consider a 4-dimensional problem with $\phi = 10$. Then, one possible example of a solution to the problem is given by *3*9. In this case, the ranges for the second and fourth dimension are identified, whereas the first and third are left as "don't cares". The evolutionary algorithm uses the dimensionality of the projection k as an input parameter. Therefore, for a d-dimensional data set, the string of length d will contain k specified position and $(d - k)$ "don't care" positions. The fitness for the corresponding solution may be computed using the sparsity coefficient discussed earlier. The evolutionary search technique starts with a population of p random solutions and iteratively used the processes of selection, crossover and mutation in order to perform a combination of hill climbing, solution recombination and random search over the space of possible projections. The process is continued until the population converges to a global optimum according to the *Dejong convergence criterion*[134]. At each stage of the algorithm, the m best projection solutions (most negative sparsity coefficients) are kept track of. At the end of the algorithm, these solutions are reported as the best projections in the data. The following operators are defined for selection, crossover and mutation:

- **Selection:** The copies of a solution are replicated by ordering them by rank and biasing them in the population in the favor of higher ranked solutions. This is referred to as *rank selection*.

- **Crossover:** The crossover technique is key to the success of the algorithm, since it implicitly defines the subspace exploration process. One solution is to use a uniform two-point crossover in order to create the recombinant children strings. The two-point crossover mechanism works by determining a point in the string at random called the crossover point, and exchanging the segments to the right of this point. However, such a blind recombination process may create poor solutions too often. Therefore, an optimized crossover mechanism is defined. In this case, it is guaranteed that both children solutions correspond to a k-dimensional projection as the parents, and the children typically have high fitness values. This is achieved by examining a subset of the different possibilities for recombination and picking the best among them.

- **Mutation:** In this case, random positions in the string are flipped with a predefined mutation probability. Care must be taken to ensure that the dimensionality of the projection does not change after the flipping process.

At termination, the algorithm is followed by a postprocessing phase. In the postprocessing phase, all data points containing the abnormal projections are reported by the algorithm as the outliers. The approach also provides the relevant projections which provide the *causality* (or intensional knowledge) for the outlier behavior of a data point. Thus, this approach also has a high degree of interpretability in terms of providing the reasoning for *why* a data point should be considered an outlier.

3. Distance-based Subspace Outlier Detection

In these methods, distance-based models are used in lower dimensional subspaces of the data in order to determine the relevant outliers. There are two major variations to the common task.

- In one class of models, the outliers are determined by exploring relevant subspaces.

- In another class of methods, the relevant outlying subspaces for a given data point are determined. This is more useful for providing *intensional knowledge*, for illustrating *why* a specific data point is an outlier.

The second class of methods shares similarities with the approach used in [262] for finding intensional knowledge from distance-based outliers. Both classes of methods will be discussed in subsequent sections.

3.1 Subspace Outlier Degree

A distance-based method for finding outliers in lower dimensional projections of the data is proposed in [273]. In this approach, instead of trying to find local subspaces of abnormally low density over the whole data, a local analysis is provided specific to each data point. For each data point $\overline{X}$, a set of reference points $S(\overline{X})$ are determined, which represent the proximity of the current data point being examined.

Once this reference set $S(\overline{X})$ has been determined, the relevant subspace for $S(\overline{X})$ is determined as the set $Q(\overline{X})$ of dimensions in which the variance is small. The specific threshold is picked as a user-specified fraction of the average dimension-specific variance of the data points in $S(\overline{X})$. Thus, this approach analyzes the statistics of individual dimensions independently of one another during the crucial step of subspace selection, though this may sometimes not be helpful for picking the best subspace projections. The approach of analyzing the distance behavior of individual dimensions for picking the subspace set $Q(\overline{X})$ is a rather naive generalization derived from subspace clustering methods. Unlike data clustering, the effectiveness of subspace outlier methods is almost entirely dependent upon the identification of dimensions containing rare points rather than dimensions with specific kinds of aggregate statistics. In outlier analysis, aggregate data measures such as the dimension-specific variance tell us very little about the subspace behavior of the rare points, and which choices of subspaces are likely to be most relevant for identification of these very unusual points. In some cases such as the example of the young diabetes patient discussed earlier, the unusual behavior is manifested in combinations of dimensions rather than the variances of the individual dimensions. If the absolute variance of a particular dimension such as the diabetes level is not deemed to be sufficiently low, it will not selected in the projection.

In the interesting cases, where the number of relevant dimensions is limited, the negative effects of removing a single relevant dimension can be even more drastic than keeping many irrelevant dimensions. The particularly problematic factor here is that if a mistake is made in subspace selection, there is virtually no chance of recovering from the mistake, when a single subspace is picked for analysis. As we will discuss later, other more insightful techniques in [256, 337] mitigate these impacts by using multiple subspaces for outlier analysis.

The euclidian distance of $\overline{X}$ is computed to the mean of the reference set $S(\overline{X})$ *in the subspace defined by* $Q(\overline{X})$. This is denoted by $G(\overline{X})$. The value of $G(\overline{X})$ is affected by the number of dimensions in $Q(\overline{X})$. The *subspace outlier degree* $SOD(\overline{X})$ of a data point is defined by normalizing this distance $G(\overline{X})$ by the number of dimensions in $Q(\overline{X})$.

$$SOD(\overline{X}) = \frac{G(\overline{X})}{|Q(\overline{X})|}$$

It remains to explain how the reference set $S(\overline{X})$ is generated with the use of distances. This may sometimes turn out to be a challenge, since the concept of proximity is itself hard to define in full dimensional space. Therefore, there is a circularity in using full dimensional distances to pick the reference set. The work [273] uses a shared nearest neighbor approach in order to compute this locality.

This work tries to find the outliers in a *single* subspace of the data, on the basis of local analysis. In practice, the deviations may be hidden in unusual subspaces which are not evident from the 1-d variance statistics of the reference set. Therefore, if the wrong subspace is selected by aggregate analysis, it is quite likely that many outliers may be missed. Furthermore, since the different dimensions in the data may combine to provide unusual results, it is sometimes more helpful to evaluate the locality of a data point in a subspace by examining the data distribution in the entire subspace, rather than examining the different dimensions independently from one another.

3.2 Finding Distance-based Outlying Subspaces

Most of the methods for outlier detection attempt to search for relevant subspaces in order to find outliers. However, some recent methods [499–501] are designed for finding the outlying subspaces *for a given data point*. Thus, the causality in this case is the other way around, where subspaces are determined from points.

A system called *HOS-Miner* was presented in [499]. According to this work, the definition of the outlying subspace for a given data point $\overline{X}$ is as follows:

DEFINITION 5.1 *For a given data point* $\overline{X}$*, determine the set of subspaces such that the sum of its* k*-nearest neighbor distances in that subspace is at least* δ*.*

This approach does not normalize the distances with the number of dimensions. Therefore, a subspace becomes more likely to be outlying with increasing dimensionality. This definition also exhibits closure

properties in which any subspace of a non-outlying subspace is also not outlying. Similarly, every superset of an outlying subspace is also outlying. Clearly, only *minimal* subspaces which are outliers are interesting. The method in [499] uses both downward- and upward-closure properties to prune off subspaces which are either not relevant or not interesting. An X-Tree is used in order to perform the indexing for performing the k-nearest neighbor queries in different subspaces efficiently. It should be noted that while the closure properties result in better efficiency and algorithmic convenience, they do not necessarily imply greater effectiveness. As the earlier example with the young diabetes patient illustrated, true outliers are often hidden in subspaces of the data, which cannot be inferred from their lower or higher dimensional projections.

In order to further improve the efficiency of the learning process, the work in [499] uses a random sample of the data in order to learn about the subspaces before starting the subspace exploration process. This is achieved by estimating a quantity called the *Total Savings Factor (TSF)* of the outlying subspaces. These are used to regulate the search process for specific query points and prune the different subspaces in an ordered way. Furthermore, the TSF values of different subspaces are dynamically updated as the search proceeds. It has been shown in [499] that such an approach can be used in order to determine the outlying subspaces of specific data points efficiently. Numerous methods for using different kinds of pruning properties and genetic algorithms for finding outlying subspaces are presented in [500, 501].

4. Combining Outliers from Multiple Subspaces

One of the major challenges of subspace analysis is that a given data point may show very different behavior in terms of its outlier degree in different subspaces. This also corresponds to the fact that the *outlier scores* from different subspaces may all be very different. These need to be combined into a unified outlier score. This principle is generally related to that of ensemble-analysis, which was discussed in Chapter 1. A variety of methods have been proposed for examining different subspaces for outlier ranking.

4.1 Random Subspace Sampling

The simplest method for combining outliers from multiple subspaces is the use of random subspace sampling. In the work in [289], an approach called *feature bagging* is used, which is analogous to the ensemble technique often used in data classification. This approach also falls in the class of *independent ensembles* introduced in Chapter 1.

The broad approach is to repeatedly apply the following two steps:

- Randomly select between $(d/2)$ and d features from the underlying data set in iteration t in order to create a data set D_t in the tth iteration.
- Apply the outlier detection algorithm O_t on the data set D_t in order to create score vectors S_t.

In principle, the outlier detection algorithm O_t used for the tth iteration could be different. However, the work in [289] uses the LOF algorithm for all the iterations.

At the end of the process, the outlier scores from the different algorithms need to be combined. There are two distinct methods which are used in order to combine the different subspaces:

- *Breadth-first Approach:* In this approach, the ranking of the algorithms is used for combination purposes. The top-ranked outliers over all the different executions are ranked first, followed by the second-ranked outliers (with repetitions removed), and so on. Minor variations could exist because of tie-breaking between the outliers within a particular rank.
- *Cumulative Sum Approach:* The outlier scores over the different algorithm executions are summed up. The top ranked outliers are reported on this basis.

It was shown in [289] by synthetic data analysis, that combining methods are important when some of the features are noisy. In such cases, full-dimensional algorithms are unable to distinguish the true outliers from the normal data, because of the additional noise. Improvements over the base LOF-approach were also observed with the use of real-data analysis. At first sight, it would seem that random subspace sampling [289] does not attempt to optimize the discovery of subspaces to finding rare instances at all. Nevertheless, it does have the paradoxical merit that it is relatively efficient to sample subspaces, and therefore a large number of subspaces can be sampled in order to improve robustness. The robustness resulting from multiple subspace sampling is clearly a very desirable quality, as long as the combination function at the end recognizes the differential behavior of different subspace samples for a given data point. In a sense, this approach implicitly recognizes the difficulty of detecting relevant and rare subspaces for the outlier detection problem, and therefore approaches the problem by sampling as many subspaces as possible in order to reveal the rare behavior. From a conceptual perspective, this approach is similar to that of harnessing the

power of many weak learners to create a single strong learner in classification problems. The approach has been shown to show consistent performance improvement over full dimensional methods for many real data sets in [289]. This approach may also be referred to as the *feature bagging method* or *random subspace ensemble method*. This approach is likely to have significant potential for improving subspace analysis, by experimenting with different choices of combination functions.

The work in [310] designs the concept of *isolation forest*, which derives its motivation from another ensemble technique known as *random forests*, which are commonly used in classification. In this case, the data is recursively partitioned by axis-parallel cuts along randomly selected attributes, so as to isolate different kinds of instances from one another. In such cases, the tree branches containing outliers are noticeably less deep, because these data points are quite different from the normal data. Thus, data points which have noticeably shorter paths in the branches of different trees are more likely to be outliers. The different branches correspond to different local subspace regions of the data, depending on how the attributes are selected for splitting purposes. The smaller path methods correspond to lower dimensionality of the subspaces in which the outliers have been isolated. The final combination step is performed by using the path lengths of the data points in the different samples. One major challenge of using such an approach is that when the dimensionality of the data increases, an incorrect choice of attribute for splitting at the higher levels of the tree is more likely to mislead the detection approach. Nevertheless, the approach is efficient in determining each subspace sample, and the use of multiple subspace samples is a desirable quality of the approach.

4.2 Selecting High Contrast Subspaces

The subspace ensemble method [289] discussed in the last section randomly samples subspaces. If many dimensions are noisy, at least a few of them are likely to be included in each subspace sample. This implies that a larger number of subspace samples will be required in order to obtain more robust results. Therefore, it is natural to ask whether it is possible to perform a pre-processing in which a smaller number of *high-contrast* subspaces are selected.

In the work proposed in [256], the outliers are found only in these high-contrast subspaces, and the corresponding scores are combined together. Thus, this approach decouples the subspace search as a a generalized pre-processing approach from the outlier ranking of the individual data points. The approach discussed in [256] is quite interesting because

of its pre-processing approach to finding relevant subspaces in order to reduce the irrelevant subspace exploration. While the high contrast subspaces are obtained using aggregation-based methods, the aggregation behavior is only used as hints in order to identify multiple subspaces for greater robustness. The assumption here is that rare events are *statistically more likely* to occur in subspaces where there is significant non-uniformity and contrast. The final outlier score combines the results over different subspaces. The insight in the work of [256] is to combine subspace selection and multiple subspaces analysis in order to determine the relevant outlier scores. Therefore, the risk of not picking the correct subspace is reduced. This approach has been shown to work well in [256] over the random subspace sampling method.

The conditional probability for an an attribute value along any particular dimension $P(x_1|x_2 \ldots x_d)$ is the same as its unconditional probability $P(x_1)$ for the case of uncorrelated data. High-contrast subspaces are likely to violate this assumption because of non-uniformity in data distribution. In our earlier example of the young diabetes patients, this corresponds to the unexpected rarity of the *combination* of youth and the disease. The idea is that subspaces with such unexpected non-uniformity are more *likely* to contain outliers, though it is treated only as a weak hint for pre-selection of one of multiple subspaces.

A variety of tests based on the student's t-distribution can be used in order to measure the deviation of this sample from the basic hypothesis of independence. This provides a measure of the non-uniformity of the subspace, and therefore provides a way to measure the quality of the subspaces in terms of their propensity to contain outliers. A bottom-up *Apriori* style [29] approach was proposed in order to determine the relevant projections. In this approach the subspaces are continuously extended to higher dimensions for testing. Details of the approach are available in [256].

4.3 Local Selection of Subspace Projections

The work in [337] uses *local* statistical selection of relevant subspace projections in order to determine outliers. In other words, the selection of the subspace projections is optimized to specific data points, and therefore the locality of a given data point matters in the selection process. For each data point $\overline{X}$, a set of subspaces is identified, which are considered *high contrast* subspaces from the perspective of outlier detection. However, this exploration process uses the high contrast behavior as statistical *hints* in order to explore *multiple* subspaces for robustness, since a single subspace may often miss the true projection.

Algorithm *OUTRES*(Data Point: $\overline{X}$
Subspace: S);
begin
for each attribute i not in S
if $S_i = S \cup \{i\}$ passes non-uniformity test **then**
begin
Compute $OS(S_i, \overline{X})$;
$O(\overline{X}) = OS(S_i, \overline{X}) \cdot O(\overline{X})$;
$OUTRES(\overline{X}, S_i)$;
end
end

Figure 5.2. The OUTRES Algorithm

The *OUTRES* method [337] examines the density of lower dimensional subspaces in order to identify relevant projections. The basic hypothesis, is that for a given data point $\overline{X}$ it is desirable to determine subspaces in which the data is sufficiently non-uniformly distributed in its locality. In order to characterize the distribution of the locality of a data point, the work in [337] computes the density of the locality of data point $\overline{X}$ in subspaces S as follows:

$$den(S, \overline{X}) = |\mathcal{N}(\overline{X}, S)| = |\{\overline{Y} : dist(\overline{X}, \overline{Y} \leq \epsilon\}|$$

This is the simplest possible definition of the density, though other more sophisticated methods such as kernel density estimation [409] are used in *OUTRES* in order to obtain more refined results. Kernel density estimation is also discussed in Chapter 4. A major challenge here is in comparing the subspaces of varying dimensionality. This is because the density of the underlying subspaces reduces with increasing dimensionality. It has been shown in [337], that it is possible to obtain comparable density estimates across different subspaces of different dimensionalities, by selecting the bandwidth of the density estimation process according to the dimensionality of the subspace.

Furthermore, the work in [337] uses statistical techniques in order to meaningfully compare different subspaces. For example, if the data is uniformly distributed, then the number of data points lying within a distance ϵ of the data point should be regulated by the fractional volume of the data in that subspace. Specifically, the fractional parameter defines a binomial distribution characterizing the number of points in that volume, if that data were to be uniformly distributed. Of course, one is really interested in subspaces which deviate significantly from this

behavior. The (local) relevance of the subspace for a particular data point $\overline{X}$ is computed using statistical testing. The two hypothesis are as follows:

- Hypothesis H_0: The local subspace neighborhood $\mathcal{N}(\overline{X}, S)$ is uniformly distributed.
- Hypothesis H_1: The local subspace neighborhood $\mathcal{N}(\overline{X}, S)$ is not uniformly distributed.

The Kolmogorov-Smirnoff goodness of fit test [424] is used to determine which of the afore-mentioned hypothesis are true. It is important to note that this process provides an idea of the *usefulness* of a subspace, and is used in order to enable a *filtering condition* for removing irrelevant subspaces from the process of computing the outlier score of a specific data point. A subspace is defined as relevant, if it passes the hypothesis condition H_1. In other words, outlier scores are computed using a combination of subspaces which *must* satisfy this relevance criterion.

In order to combine the scores which are obtained from multiple *relevant* subspaces, the work in [337] uses the product of the outlier scores obtained from different subspaces. Thus, if $S_1 \ldots S_k$ be the different abnormal subspaces found for data point $\overline{X}$, and if $O(S_i, \overline{X})$ be the outlier score from subspace S_i, then the overall outlier score $OS(\overline{X})$ is defined as follows:

$$OS(\overline{X}) = \prod_i O(S_i, \overline{X})$$

It is evident that *low scores* represent a greater tendency to be an outlier. The advantage of using the product over the sum, is that the latter is dominated by the high scores, as a result of which a few subspaces containing normal behavior will dominate the sum. On the other hand, in the case of the product, the outlier behavior in a small number of subspaces will be greatly magnified. This is particularly appropriate for the problem of outlier detection. So far, it has not been discussed, how the actual subspaces $S_1 \ldots S_k$ are determined. This will be achieved with a careful subspace exploration.

In order to actually define the outlier score, subspaces are considered significant for particular objects only if their density is at least two standard deviations less than the mean value. This is essentially a filter condition for that subspace to be considered deviant. Thus, the deviation $dev(\overline{X}, S_i)$ of the data point $\overline{X}$ in subspace S_i is defined as the ratio of the deviation of the density of the object from the mean density, divided by two standard deviations.

$$dev(S_i, \overline{X}) = \frac{\mu - den(S_i, \overline{X})}{2 \cdot \sigma}$$

The outlier score of a data point in a subspace is the ratio of the density of the point in the space to its deviation, if it satisfies the filter condition of the density being at least two standard deviations less than the mean. Otherwise the outlier score is considered to be 1, and it does not affect the overall outlier score in the product function defined earlier for combining different subspaces. Thus, for the points satisfying the filter condition, the outlier score $OS(S_i, \overline{X})$ is defined as follows:

$$O(S_i, \overline{X}) = \frac{den(S_i, \overline{X})}{dev(S_i, \overline{X})}$$

An observation in [337] is that subspaces which are either very low dimensional (eg. 1-d subspaces) or very high dimensional are not very informative from an outlier detection perspective. A recursive exploration of the subspaces is performed, where an additional attribute is included in the subspace for statistical testing. Therefore, the work in [337] uses recursive processing in which the subspaces are built in recursive fashion. When an attribute is added to the current subspace S_i, the non-uniformity test is utilized to determine whether or not that subspace should be used. Otherwise, this subspace is discarded.

The overall algorithm uses a recursive subspace exploration procedure in order to measure the outlierness of any particular object. Note that the entire recursive algorithm uses the data point $\overline{X}$ as input, and therefore the procedure needs to be applied separately *for each data point*. For any given subspace, an attribute is incrementally added. Then, the non-uniformity test is applied to determine if it is relevant. If it is not relevant, then the subspace is discarded. Otherwise, the outlier score $O(S_i, \overline{X})$ in that subspace is computed for the data point, it is multiplied with the current value of $OS(\overline{X})$. Since the outlier scores of subspaces, which do not meet the filter condition are set to 1, they do not affect the density computation in this multiplicative approach. The procedure is then recursively called in order to explore the next subspace. Thus, such a procedure potentially explores an exponential number of subspaces, though the real number is likely to be much smaller in practice. This is because of the non-uniformity test, which prunes off large parts of the recursion tree during the exploration. The overall algorithm for subspace exploration for a given data point $\overline{X}$ is illustrated in Figure 5.2.

5. Generalized Subspaces

A significant amount of success has been achieved for finding outliers in axis-parallel subspaces in recent work. While these methods are effective for finding outliers in cases where the outliers naturally deviate in

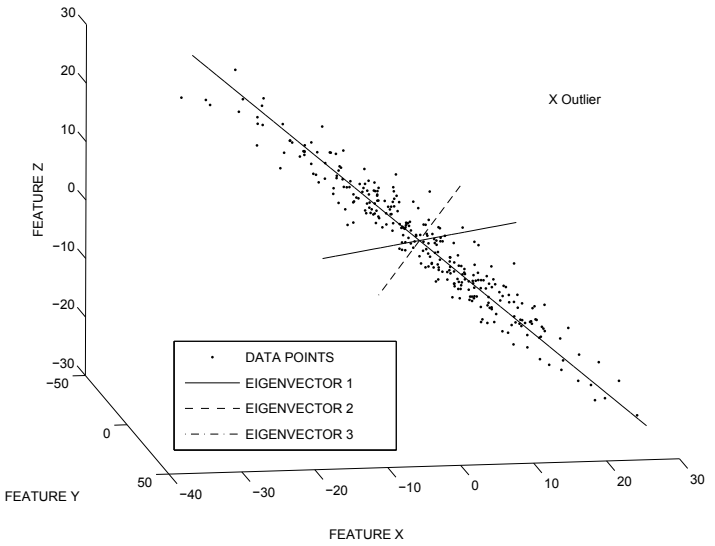


Figure 5.3. The example of Figure 3.4 re-visited: Global PCA can discover outliers in cases, where the entire data is aligned along lower dimensional manifolds.

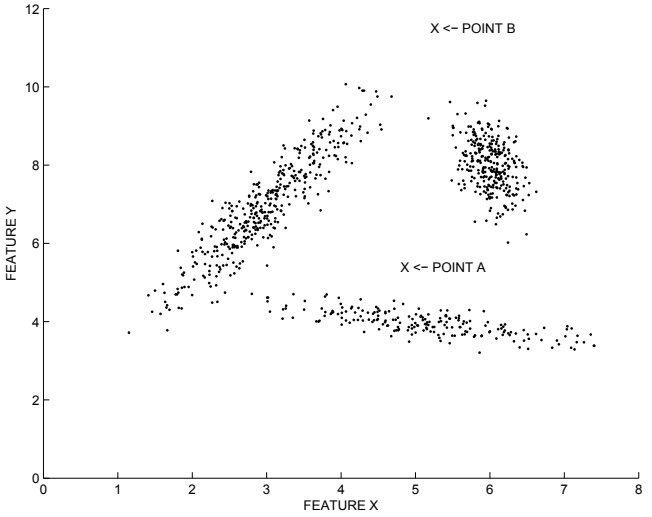


Figure 5.4. The example of Figure 2.7 revisited: Outliers are best discovered by determining deviations from local PCA-based clusters. Neither axis-parallel subspace outliers nor global-PCA can capture such clusters.

specific subspaces from the clusters, they are not very useful for finding clusters in cases where the points are aligned along lower-dimensional manifolds of the data. For example, in the case of Figure 5.4, no 1-dimensional subspace analysis from the 2-dimensional data can find the outliers. On the other hand, it is possible to find *localized* 1-dimensional correlated subspaces so that most of the data aligns along these localized 1-dimensional subspaces, and the remaining deviants can be classified as outliers.

These algorithms are generalizations of the following two classes of algorithms:

- The PCA-based linear models discussed in Chapter 3 find the *global* regions of correlation in the data. For example, in the case of Figure 5.3, the outliers can be effectively identified by determining these global directions of correlation. However, no such *global* directions of correlation exist in the case of Figure 5.4.
- The axis-parallel subspace outliers discussed earlier in this chapter can find deviants, when the data is naturally aligned along low dimensional axis-parallel subspace clusters. However, this is not the case in Figure 5.4, where the data is aligned along arbitrary directions of correlation.

This problem can be partially addressed with the use of generalized projected clustering methods, where the clusters are determined in arbitrarily aligned subspaces of the data [7]. The method discussed in [7] has a built-in mechanism in order to determine the outliers *in addition to* the clusters. Such outliers are naturally data points which do not align with the clusters. However, the approach is not particularly optimized for finding the outliers, because the primary purpose of the method is to determine the clusters. The outliers are discovered as a side-product of the clustering algorithm, rather than as the primary goal. Therefore, the approach may discover the weaker outliers, which correspond to the noise in the data. Similarly, the approach in [132] is focussed on determining the noise in the data for improving mixture modeling of probabilistic PCA algorithms. In order to determine the outliers which are optimized to the locality of a particular data point, it is critical to determine localized subspaces which are optimized to the data point $\overline{X}$, which is being evaluated for its outlier score. The determination of such subspaces is non-trivial, since it often cannot be inferred from locally aggregate properties of the data, for detecting the behavior of *rare* instances.

Another method was recently proposed in [274] for finding outliers in generalized subspaces of the data. The main difference from earlier gen-

eralized subspace clustering methods is that local reference sets are used for local correlation analysis. For a given data point $\overline{X}$, this method finds the full-dimensional k-nearest neighbors of $\overline{X}$. This provides a reference set S with mean vector $\overline{\mu}$. The PCA approach of Chapter 3 is applied to the covariance matrix $\Sigma(S)$ of the *local* reference set S in order to determine the key eigenvectors $\overline{e_1} \ldots \overline{e_d}$, in increasing order of variance, with corresponding eigenvalues $\lambda_1 \leq \lambda_2 \ldots \leq \lambda_d$. The discussion in section 3 of Chapter 3 performs these same steps [406] except that they are performed on a *global* basis, rather than on a local reference set S. Even if all d dimensions are included, it is possible to create a normalized outlier score of a data point $\overline{X}$, to the centroid $\overline{\mu}$ of the data with the use of local eigenvalue scaling, as discussed in Chapter 3:

$$Score(\overline{X}) = \sum_{j=1}^{d} \frac{|(\overline{X} - \overline{\mu}) \cdot \overline{e_j}|^2}{\lambda_j} \tag{5.1}$$

As discussed in section 2.2.2 of Chapter 2, this can be approximately modeled as a χ^2 distribution with d degrees of freedom for each data point, and the outlier scores of the different data points can be reasonably compared to one another. Such an approach is used in [406] in the context of global data analysis. The survey paper of Chandola et al. [107] provides a simpler exposition. The work in [274] uses a similar approach with the use of a local reference set, selected with the use of full dimensional k-nearest neighbor distances.

Eigenvectors with large values of λ_i will usually not contribute much to the score, though as discussed below, this may not always be the case. Such directions are pruned from the score. The δ eigenvectors[2] with the smallest eigenvalues are picked for the computations above. Correspondingly, the pruned score is defined on the basis of the first $\delta \leq d$ eigenvectors only with the smallest eigenvalues.

$$Score(\overline{X}, \delta) = \sum_{j=1}^{\delta} \frac{|(\overline{X} - \overline{\mu}) \cdot \overline{e_j}|^2}{\lambda_j} \tag{5.2}$$

How should the value of δ be determined for a particular data point $\overline{X}$? The score is a χ^2-distribution with δ-degrees of freedom. It was observed in [274] that the value of δ can be parameterized, by treating the χ^2 distribution as a special case of the Γ distribution.

$$Score(\overline{X}, \delta) \sim \Gamma(\delta/2, 2)$$

[2]The work in [274] uses δ as the number of *longest* eigenvectors, which is only a notational difference, but is noted here to avoid confusion.

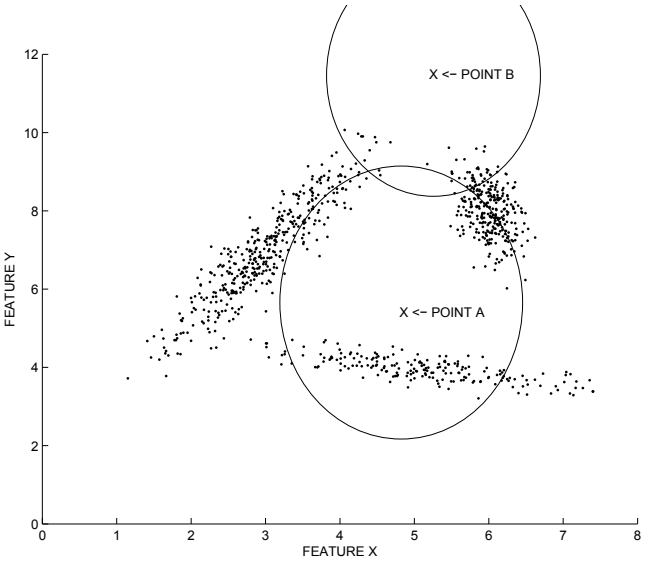


Figure 5.5. Local reference set may sometimes contain points from multiple generating mechanisms

The optimal value of δ is picked specifically for each data point, by picking the value of δ in order to determine the maximal unlikely deviation based on this model. This is done by using the cumulative density function of the aforementioned distribution. While this value can be directly used as an outlier score, it was also shown in [274], how this score may be converted into a more intuitive probability value.

This approach has several issues:

- A *single subspace* has been used by this approach for finding the outliers with the use of the local reference set S. If the local reference set S is not accurately determined, then this will not provide the proper directions of local correlation. The use of a single subspace is risky, especially with the use of weak aggregation-based hints, because it is often possible to unintentionally remove relevant subspaces. This can have drastic effects. The use of multiple subspaces may be much more relevant in such scenarios, such as the methods proposed in [289, 256, 337, 341].

- There is an inherent circularity in identifying the reference set with the use of full dimensional k-nearest neighbor distances, especially if the distances are not meaningfully defined in full dimensionality. The choice of points in the reference set and the choice of the subspace clearly impact each other in a circular way. This is a classical

"chicken and egg" problem in subspace analysis, which was first pointed out in [5]. The analysis in such cases needs to be *simultaneous* rather than *sequential*. As is well known, the most robust techniques for handling circularity in virtually all problem domains (eg. the EM algorithm and many projected clustering methods) use iterative methods, so that the point-specific and dimension-specific aspects of the problem are able to interact with one another. This is however, not the case in [274], where a sequential analysis is used.

In particular, it may happen that many locally irrelevant features may be used during the determination of the local reference set, when full dimensional distances are used. This set could therefore contain data points from multiple generating mechanisms, as illustrated in Figure 5.5. When the number of irrelevant features is unknown, a specific number of points in the reference set will not be able to avoid this problem. The use of a smaller reference set size can reduce the chance of this happening to some extent, but can never guarantee it, especially when many irrelevant features are used. On the other hand, reducing the reference set size can also result in a correlation hyperplane, whose eigenvalue statistics overfit an artificially small set of reference points.

- An interesting question arises, as to whether it is necessary to select a particular set of dimensions in a hard way, since the eigenvalues in the denominator of Equation 5.1 already provide a soft weighting to the importance (or relevance) of the different dimensions. For example, if for a large value of λ_i, a data point shows even larger deviations along that direction, such an outlier would either be missed by dimension pre-selection, or would include other less relevant dimensions. An example is the outlier B in Figure 5.5, which is aligned along the longer eigenvector, and therefore the longest eigenvector is the *most informative* about its outlier behavior. In particular, the method of picking the δ smallest eigenvectors implicitly assumes that the relevance of the attributes are ordered by eigenvalue magnitude. While this may generally be true for aggregation-based clustering algorithms, it is very often not true in outlier analysis because of the unusual nature of outliers. The possibility of outliers aligning along long eigenvectors is not uncommon at all, since two highly correlated attributes may often show highly deviant behavior of a similarly correlated nature. This example also shows, how *brittle* the rare nature of outlier analysis is to aggregation-based measures. This is because of the varying

causes of rarity, which cannot be fully captured in aggregation statistics. This is relevant to our discussion in the introduction section, that straightforward generalizations of subspace selection methods from clustering (based on aggregates), are often not appropriate or optimized for (the rare nature of) outlier analysis. One advantage of using all the dimensions is that it reduces to a local Mahalanobis distance with the same dimensionality, and allows better comparability in the scores across different outliers. In such cases, intuitive probability values may be derived more simply from the $\chi^2(d)$ distribution.

The high dimensional case is an extremely difficult one, and it is understandable that no given method will be able to solve these problems perfectly. It should also be pointed out that the iterative EM algorithm discussed in Chapter 2 will be able to discover the local directions of correlation along with outliers which have low fit value to the model. These may sometimes include weak outliers, which are not always interesting. Given that direct discovery of optimal subspaces in a given locality is much more difficult in outlier analysis, a possible line of work would be to use a two-phase approach of first finding the weak outliers, and then determining the strong ones among them by more detailed analysis. For example, it may be possible to use this pre-filtered set of weak outliers for intensive ensemble-based subspace exploration. Combining pre-filtered data points with pre-filtered high-contrast subspaces may provide an interesting direction of future exploration. A significant scope still exists for further improvement of the techniques designed in this area.

6. Discussion of Subspace Analysis

While subspace outlier analysis seems to be the only meaningful method for high dimensional outlier detection, the approach faces a number of challenges, a lot of which are computational in nature. In the high-dimensional case, a small number of deviant subspaces may remain hidden out of a large number of possibilities. This can create unprecedented challenges for outlier analysis. The combinatorial nature of the problem necessitates the design of more efficient algorithms which can perform an ordered exploration of these spaces. In spite of the recent advances in the literature, the design of efficient algorithms for the high dimensional subspace exploration scenario remains a challenge. This is of course an inherent property of high-dimensional data, in which the curse of dimensionality impacts the results both from a qualitative and efficiency perspective.

The second challenge arises from the fact that a subspace exploration technique reports a number of different possibilities for the projections. In such cases, it remains a challenge to combine the results from these deviant subspaces, and rank the resulting outliers effectively. This is of course an opportunity as well, since the results from multiple subspaces may provide more robust outliers. Therefore, significant advancements are required in *ensemble analysis* for outlier detection.

It has been claimed in [514] as an apparently new insight, that the major reason for difficulty in high dimensional outlier analysis is not the concentration of distances, but the masking effects of the locally noisy and irrelevant nature of some of the dimensions, and that the literature has failed to discuss the impact of locally relevant dimensions. This is an incorrect assertion, since both the aspects of local feature selection (relevance) and distance concentration have been studied extensively in the literature. While it is true that noisy and irrelevant attributes mask the outliers, the observation is certainly not new, and the two factors of distance concentration and local feature relevance are closely related. The original work in [4] (and virtually every other subsequent work [289, 256, 337] on this topic) provides a pictorial illustration and a fairly detailed discussion of how (locally) irrelevant attributes mask outliers in different feature-specific views of the data. As stated in [4]: "... *by using full dimensional distance measures it would be difficult to determine outliers effectively because of the averaging behavior of the noisy and irrelevant dimensions. Furthermore, it is impossible to prune off specific features a-priori, since different points may show different kinds of abnormal patterns, each of which use different features or views.*" The ineffectiveness of *global* feature selection in high dimensional data in fact forms the motivating reason for subspace analysis, which can be considered a *local* feature selection method, or a *local* dimensionality reduction method [7, 95]. These connections of local subspace analysis to the ineffectiveness of global feature selection in high dimensional data were explicitly discussed in detail in the motivational discussion of one of the earliest works on subspace analysis [5]. At this point, these results are well known and established[3] wisdom. While it is possible to reduce the distance concentration effects by carefully calibrating the fraction of informative dimensions, such cases are (usually) not interesting for subspace analysis.

[3]Some of the earliest methods even refer to these classes of techniques as local dimensionality reduction [95] in order to emphasize the enhanced and differential local feature selection effect, which arises as a result of different generating mechanisms.

Distance concentration and (too many) irrelevant attributes are closely related. The interesting cases for subspace analysis (typically) show some levels of both properties. Even limited levels of distance concentration impact the effectiveness of full dimensional distance-based algorithms, and this impact is therefore important to examine in outlier analysis. It should be noted that noisy and irrelevant attributes are more likely to lead to concentration of distances. For example, for the case of uniformly distributed data, where all attributes are noisy, the concentration effect is extreme, and an outlier deviating along *a relatively small number of dimensions* will be hard to discover by full dimensional methods. In such cases, from a full dimensional distance-based or density-based perspective, all data points have almost equally good outlier scores, and this can be equivalently understood in terms of *either* locally irrelevant features or distance concentration effects. Of course, real data sets are not uniformly distributed, but *both* irrelevant features and concentration effects are present to varying degrees in different data sets. The general assumption for subspace analysis is that the addition of more dimensions often does not add *proportionally* more information for a particular outlier. The challenging outliers are often defined by the behavior of a small number of dimensions, and when the point-specific information does not increase substantially with data dimensionality, even modest concentration effects will have a negative impact on full dimensional algorithms. The more the number of irrelevant attributes, the more erroneous the computations for full-dimensional distance-based methods. An extreme example at the other end of the spectrum is where an outlier shows informative and deviant behavior in every dimension, and therefore outlier characteristics grow *stronger* with increasing dimensionality. However, in this rather uninteresting case, since the outlier shows *both* many relevant features *and* also typically does not conform to the distance concentration behavior of the remaining data, a trivial full dimensional distance-based algorithm would find it easily in most cases. In general, cases where the informative dimensions also increase significantly with data dimensionality, are not as interesting for subspace analysis because the full dimensional masking behavior becomes less prominent in this easier case. Subspace analysis does not exclude the possibility that the more obvious deviants may also be found by full dimensional analysis.

Outliers, by their very rare nature, may often be hidden in small combinations of dimensions in a high dimensional data set. Subspace analysis is interesting for such scenarios. On the other hand, when more dimensions do add (significantly) more information, then this becomes an easy case for analysis, which no longer remains interesting. In the

former case, the vast majority of noisy dimensions make all data points appear as outliers from a density-based or data sparsity perspective.

To summarize, subspace outlier analysis is one of the most challenging problems because of the rare and unusual nature of outliers. In order to design meaningful algorithms, the following principles need to be kept in mind.

- Aggregation-based methods for subspace analysis only provide very weak hints for outlier analysis as compared to clustering algorithms. A direct exploration of rare regions is possible, though it is computationally challenging because of combinatorial explosion [4]. As a result, it becomes necessary to use heuristic methods.

- Aggregation-based methods may be usable, if caution is utilized in recognizing the fact that a given subspace derived from such methods may not always include the relevant dimensions. Exclusion of relevant dimensions has more drastic effects than inclusion of many irrelevant dimensions. Where possible, subspace ensembles should be used in order to combine the weak hints derived from the different subspaces, if aggregation-based measures are used.

- The individual component of an ensemble should be designed with efficiency considerations. This is because the ability to execute the individual component more number of times within a fixed time frame, eventually provides more robustness.

7. Conclusions and Summary

Subspace methods for outlier detection are used in cases, where the outlier tendency of a data point is diluted by the noise effects of a large number of locally non-informative dimensions. In such cases, the outlier analysis process can be sharpened significantly by searching for subspaces in which the data points deviate significantly from the normal behavior. The earliest work on subspace outlier detection used evolutionary search methods in order to determine abnormal lower dimensional projections of the data. A number of subsequent methods have also been designed for determining multiple relevant subspaces for a candidate outlier, and then combining the results from different subspaces in order to create a more robust ensemble-based ranking. It is also possible to determine the outliers in arbitrarily oriented subspaces of the data. Such methods are able to exploit the local correlations in the data in order to determine relevant outliers.

Outlier analysis is the most difficult problem among all classes of subspace analysis problems. This difficulty arises out of the rare nature

of outliers, which makes direct statistical analysis more difficult. Since subspace analysis and local feature selection are related, it is noteworthy that even for global feature selection, there are few known methods for outlier analysis, as compared to clustering and classification algorithms. The reason is simple: enough statistical evidence is often not available for the analysis of rare characteristics. Robust statistics is all about *more* data, and outliers are all about *less* data and statistical non-conformity with most of the data! Regions and subspaces containing statistical conformity tell us very little about the complementary regions of non-conformity in the particular case of high-dimensional subspace analysis, since the *potential* domain of the latter is much larger than the former. In particular, a local subspace region of the greatest aggregate conformity does not necessarily reveal anything about the rare point with the greatest statistical non-conformity.

While it is doubtful that the more difficult variations of the problem will ever be fully solved, or will work completely in all situations, it may be possible to design methods which work in many important scenarios. There are many merits in being able to design such methods, because of the numerous insights they can provide in terms of identifying the causes of abnormality. The main challenge is that outlier analysis is so brittle, that it is often impossible to make confident assertions about inferences drawn from aggregate data analysis. The issue of efficiency seems to be closely related to that of effectiveness in high dimensional outlier analysis. This is because the search process for outliers is likely to require exploration of multiple local subspaces of the data in order to ensure robustness. With increasing advances in the computational power of modern computers, there is as yet hope that this area will become increasingly tractable for analysis.

8. Bibliographic Survey

In the context of high-dimensional data, there are two distinct lines of research, one of which investigates the *efficiency* of high dimensional outlier detection [46, 185, 467], and the other investigates the more fundamental issue of the *effectiveness* of high dimensional outlier detection [4, 273]. Unfortunately, the distinction between these two lines of work is sometimes blurred in the literature, even though these are clearly different lines of work with very different motivations. It should be noted that the methods discussed in [46, 185, 467] are all *full dimensional methods*, because outliers are defined on the basis of their full dimensional deviation. While the method of [467] uses projections for indexing, this is

used only as an approximation to improve the efficiency of the outlier detection process.

In the high-dimensional case, the efficiency of (full dimensional) outlier detection also becomes a concern, because most outlier detection methods require repeated similarity search in high dimensions in order to determine the nearest neighbors. The efficiency of these methods degrades because of two factors: (i) the computations now use a larger number of dimensions, and (ii) the effectiveness of pruning methods and indexing methods degrades with increasing dimensionality. The solution to these issues still remains unresolved in the vast similarity search literature. Therefore, it is unlikely that *significantly* more efficient similarity computations could be achieved in the context of high dimensional outlier detection, though some success has been claimed for improving the efficiency of high dimensional outlier detection in methods proposed in [46, 185, 467]. On the whole, it is unclear how these methods would compare to the vast array of techniques available in the similarity search literature for indexing high dimensional data. This chapter does *not* investigate the efficiency issue at all, because the efficiency of a *full dimensional* outlier detection technique is not important, if it does not even provide meaningful outliers. Therefore, the focus of the chapter is on methods which *re-define* the outlier detection problem in the context of lower dimensional projections. It is also noted that an angle-based outlier detection for high-dimensional data has been proposed in [269], though this method has been discussed in the chapter on extreme value analysis (Chapter 2), since this method is not a subspace exploration technique. It is also designed to find specific kinds of outliers which lie at the boundaries of the multivariate data, and is much closer in principle to other multivariate extreme value analysis methods such as depth-based and deviation-based methods.

The problem of subspace outlier detection was first proposed in [4]. In this paper, an evolutionary algorithm was proposed to discover the lower dimensional subspaces in which the outliers may exist. The method for distance-based outlier detection with subspace outlier degree was proposed in [273]. Another distance-based method for subspace outlier detection was proposed in [346]. Some methods have also been proposed for outlier analysis by randomly sampling subspaces and combining the scores from different subspaces [289, 310]. In particular, the work in [289] attempts to combine the results from these different subspaces in order to provide a more robust evaluation of the outliers. These are essentially *ensemble-based* methods, which attempt to improve detection robustness by bagging the results from analyzing different sets of features. The major challenge of these methods is that random sampling may not

work very well, when the outliers are hidden in specific subspaces of the data. The work in [256] can be considered a generalization of the broad approach in [289], where only high contrast subspaces are selected for the problem of outlier detection.

The reverse problem of finding outlying subspaces *from* specific points was studied in [499–501]. In these methods, a variety of pruning and evolutionary methods were proposed in order to speed up the search process for outlying subspaces. The work in [47] also defines the exceptional properties of outlying objects both with respect to the entire population (global properties), and also with respect to particular sub-populations to which it belongs (local properties). Both these methods provide different but meaningful insights about the underlying data. A genetic algorithm for finding the outlying subspaces in high dimensional data is provided in [500]. In order to speed up the fitness function evaluation, methods are proposed to speed up the computation of the k-nearest neighbor distance with the use of bounding strategies. A broader framework for finding outlying subspaces in high dimensional data is provided in [501]. A method which uses two-way search for finding outlying subspaces is proposed in [482]. In this method, full dimensional methods are first used to determine the outliers. Subsequently, the key outlying subspaces from these outlier points are detected and reported. A method for using rules in order to explain the context of outlier objects is proposed in [340].

A number of ranking methods for subspace outlier exploration have been proposed in [337–339]. In these methods, outliers are determined in multiple subspaces of the data. Different subspaces may either provide information about different outliers, or about the same outliers. Therefore, the goal is to combine the information from these different subspaces in a robust way in order to report the final set of outliers. The *OUTRES* algorithm proposed in [337] uses recursive subspace exploration in order to determine all the subspaces relevant to a particular data point. The outlier scores from these different subspaces are combined in order to provide a final value. A tool-kit for ranking subspace outliers has been presented in [338]. A more recent method for using multiple views of the data for subspace outlier detection is proposed in [341]. Methods for subspace outlier detection in multimedia databases were proposed in [51].

Most of the methods for subspace outlier detection perform the exploration in axis-parallel subspaces of the data. This is based on the complementary assumption that the dense regions or clusters are hidden in axis-parallel subspaces of the data. However, it has been shown in recent work that the dense regions may often be located in arbitrarily ori-

ented subspaces of the data [7]. While it has been shown in earlier work that the removal of noise (or weak outliers) improves the effectiveness of generalized subspaces clustering algorithms [7], specific techniques are also required in order to determine outliers in a way which is optimized to the data correlations. Another work in [274] provides an arbitrarily oriented solution for the generalized outlier analysis problem, which extends the correlation-analysis approach proposed in [7] to a method based on local reference sets rather than clusters.

Recently, the problem of outlier detection has also been studied in the context of dynamic data and data streams. The SPOT method was proposed in [498], which is able to determine projected outliers from high dimensional data streams. Thus approach employs a window-based time model and decaying cell summaries to capture statistics from the data stream. A set of top sparse subspaces are obtained by a variety of supervised and unsupervised learning processes. These are used in order detect the projected outliers. A multi-objective genetic algorithm is employed for finding outlying subspaces from training data.

The problem of high dimensional outlier detection has also been extended to other application-specific scenarios such as astronomical data [213], uncertain data [23], transaction data [210] and supervised data [513]. In the uncertain scenario, high dimensional data is especially challenging, because the noise in the uncertain scenario greatly increases the sparsity of the underlying data. Furthermore, the level of uncertainty in the different attributes is available. This helps decide the importance of different attributes for outlier detection purposes. Subspace methods for outlier detection in uncertain data are proposed in [23]. Supervised methods for high-dimensional outlier detection are proposed in [513]. In this case, a small number of examples are presented to user of the outliers. These are then used in order to learn the critical projections which are relevant to the outlierness of an object. The learned information is then leveraged in order to determine the relevant outliers in the underlying data.

9. Exercises

1. Which of the following data points is an outlier in some well chosen two-dimensional projection: $\{ (1, 8, 7), (2, 8, 8), (5, 1, 2), (4, 1, 1), (3, 1, 8) \}$

2. Download the *Arrythmia* data set from the UCI Machine Learning Repository [169]. Write a computer program to determine all distance-based outliers in different 2-dimension projections. Are the outliers the same in different projections?

3. In the *Arrythmia* data set mentioned in the previous exercise, examine the *Age*, *Height* and *Weight* attributes of the *Arrythmia data* set both independently and in combination. Draw a scatter plot of each of the 1-dimensional distributions and different 2-dimensional combinations. Can you visually see any outliers?

4. Write a computer program to determine the subspace outlier degree of each data point in the *Arrythmia* data set for all 1-dimensional projections and 2-dimensional projections. Which data points are declared outliers?

5. Write a computer program to perform subspace sampling of the *Arrythmia* data set, using the approach of [289] by sampling 2-dimensional projections. How many subspaces need to be sampled in order to robustly identify the outliers found in Exercise 2 over different executions of your computer program.

6. Consider a data set with d-dimensions, in which exactly 3 specific dimensions behave in an abnormal way with respect to an observation. How many minimum number of random subspaces of dimensionality $(d/2)$ will be required in order to include all 3 dimensions in the subspace with probability at least 0.99? Plot the number of required samples for different values of $d > 6$.