

# Statistical Selection of Relevant Subspace Projections for Outlier Ranking

Emmanuel Müller<sup>•</sup>, Matthias Schiffer<sup>◊</sup>, Thomas Seidl<sup>◊</sup>

<sup>•</sup>*Karlsruhe Institute of Technology (KIT), Germany*  
emmanuel.mueller@kit.edu

<sup>◊</sup>*RWTH Aachen University, Germany*  
{mschiffer, seidl}@cs.rwth-aachen.de

**Abstract**—Outlier mining is an important data analysis task to distinguish exceptional outliers from regular objects. For outlier mining in the full data space, there are well established methods which are successful in measuring the degree of deviation for outlier ranking. However, in recent applications traditional outlier mining approaches miss outliers as they are hidden in subspace projections. Especially, outlier ranking approaches measuring deviation on all available attributes miss outliers deviating from their local neighborhood only in subsets of the attributes.

In this work, we propose a novel outlier ranking based on the objects deviation in a statistically selected set of relevant subspace projections. This ensures to find objects deviating in multiple relevant subspaces, while it excludes irrelevant projections showing no clear contrast between outliers and the residual objects. Thus, we tackle the general challenges of detecting outliers hidden in subspaces of the data. We provide a selection of subspaces with high contrast and propose a novel ranking based on an adaptive degree of deviation in arbitrary subspaces. In thorough experiments on real and synthetic data we show that our approach outperforms competing outlier ranking approaches by detecting outliers in arbitrary subspace projections.

## I. INTRODUCTION

Outlier mining has become an important data mining task to detect inconsistent or suspicious objects in large databases. For recent applications, outlier mining as an unsupervised learning task is important for consistency checks of sensor network measurements, fraud detection in financial transactions, emergency detection in health surveillance and many more. As measuring and storing of data has become very cheap, in all of these applications, objects are described by many attributes. However, for each object only subsets of relevant attributes provide the meaningful information, the residual attributes are irrelevant for this object. For example in health surveillance, for one patient attributes such as “age” and “skin humidity” might be important to detect the abnormal “dehydration” status of this patient. Other attributes such as “heart beat rate” are irrelevant for the detection of this outlier, but are relevant for the detection of abnormal patients with a heart disease. All of these attributes are required for some outlier detection, but each outlier occurs only in subsets of these attributes. Thus, the distinction between outliers and regular objects is heavily hindered by considering all available attributes as typically done in traditional outlier mining methods.

Traditional techniques are well established for outlier mining in the full space, but miss outliers which are hidden in subspace projections. Thus, our general aim is to develop a novel outlier ranking based on object deviation in subspace projections. We focus on outlier ranking a special research field of outlier mining which sorts objects according to their local degree of deviation. Local outlier rankings use the local neighborhood around each object to report an ordered list presenting the most outlying object first. They provide more information than just the binary decision about being an outlier or not. Outlier rankings provide for each object the extent of *outlierness*. However, traditional outlier rankings using outlierness measures in full space are not appropriate for outliers hidden in subspaces. In the full space all objects appear to be alike so that traditional outlier rankings cannot distinguish the outlierness of objects any more.

In this work we measure outlierness (the degree of deviation) of an object in projections of the database taking only subsets of attributes into account. Hence, we can successfully detect an outlier in a set of relevant subspace projections in which this outlier stands out from its surrounding objects. We model various behaviors of one object in different projections of the database. An object might show high deviation compared to its neighborhood in one subspace. In addition, the same object might cluster with some other objects in a second subspace or might not show up as an outlier in a third scattered subspace where all objects seem to be outliers. To illustrate this, we have depicted several projections of a toy example with two hidden outliers in Figure 1. Please note, that each object requires the detection of individual subspace projections to detect its outlier properties. This is in contrast to related paradigms such as subspace clustering [14] or dimensionality reduction techniques [10]. Global dimensionality reduction techniques such as principal components analysis provide only a single projection for all objects. In contrast, we aim at detecting multiple relevant subspaces per object. Subspace clustering detects multiple projections, however, focuses on subspaces for groups of clustered objects. For outlier ranking, the focus is on the individual objects and subspaces in which an outlier is highly deviating from its local neighborhood. Thus, outlier ranking in subspaces poses novel challenges not tackled by these research topics.

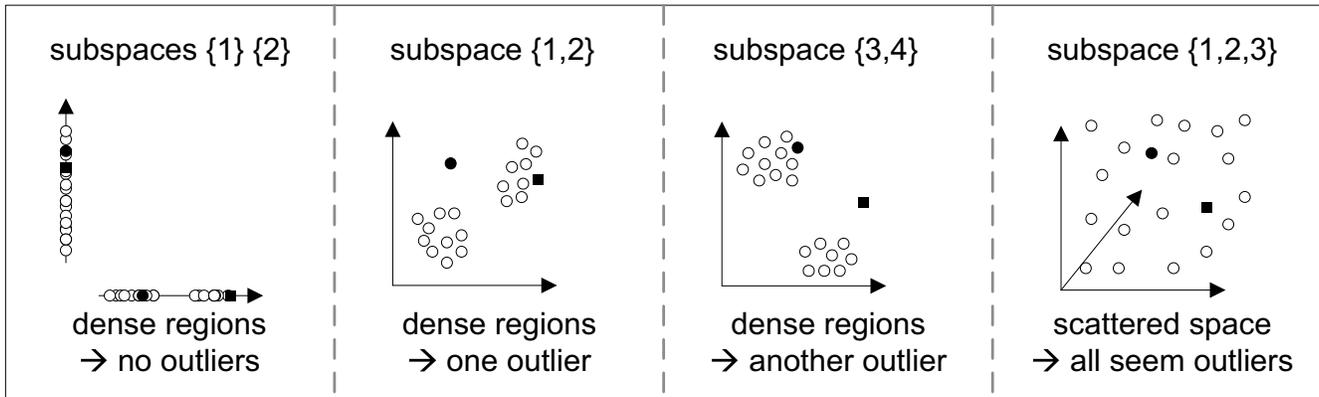


Fig. 1. Example: Outliers in arbitrary subspaces

In this work, we focus on two key properties for outlier ranking in subspace projections. First, for each object we statistically select a set of projections for outlier ranking, we call these *relevant subspaces*. In relevant subspaces the neighborhood of an object is clustered and the object is an outlier if it deviates from these clustered objects. In case these relevant subspaces inhabit different clusters in terms of objects, some objects might be members of a cluster in one subspace while being outliers in others at the same time. In contrast, in irrelevant subspaces the neighborhood of an object is distributed uniformly random such that all objects seem to be outliers. Overall, our outlier ranking is confronted with arbitrary subspaces, while only very few are relevant and contribute to a distinction between clustered objects and outliers. Second, object deviation increases with the number of attributes in a relevant subspace. As distances between objects grow more and more alike due to the “curse of dimensionality” [5], objects are clustered in dense regions in one dimensional subspaces while objects are scattered in higher dimensional spaces (cf. Figure 1). In general, the deviation of objects is highly influenced by the number of attributes in the considered subspaces. Thus, for outlier ranking in subspace projections, we have to cope with two major challenges:

- Outliers appear only in relevant subspaces.
- Incomparable deviation in different subspaces.

To tackle both of these challenges, we propose *OUTRES* a new method for outlier ranking in relevant subspace projections. For our outlier ranking we consider only a selection of non-uniformly distributed projections. We exclude uniformly random distributed subspaces by a statistical test, as they hinder the distinction between outliers and regular objects. Furthermore, we propose a novel adaptive outlier ranking measuring comparable degrees of deviation for objects in arbitrary subspaces. We define outliers to be objects highly deviating from the estimated density in their local subspace neighborhood. Overall in contrast to traditional outlier ranking approaches, our outlier ranking considers deviation in subspaces and adapts to the number of attributes in each considered subspace.

## II. RELATED WORK

In general, outliers are objects that deviate from the rest of the data to a great extent. However, there have been various outlier models proposed in the literature. We categorize these models into two paradigms, traditional outlier mining methods and subspace outlier mining techniques.

### a) Traditional Methods:

Different models have been proposed modeling deviation globally e.g. in distance-based [12], cluster-based [9] or statistical-based [4] outlier mining methods. However, such techniques suffer from difficulties in parametrization, as the extent of deviation is usually hard to quantify globally. This has led to outlier ranking based on the local degree of deviation for each object, as in the well established local outlier factor (LOF) approach [6] or its extension (LOCI) based on local deviation [24]. Further extensions have been proposed based on this general local outlier factor idea. The most recent approach proposes an angle based outlier factor (ABOF) [15]. Based on the assumption, that angles between objects are more stable than distances, ABOF computes for each object an angle range to the residual objects. However, all of these outlier ranking methods base on the full space, and thus, fail to separate outliers from regular objects in subspace projections.

### b) Subspace Methods:

In contrast, recent approaches consider subspace projections for outlier ranking. The key property for all of these approaches is the appropriate choice of considered subspaces. As the most basic approach a random choice of subspaces has been proposed for outlier ranking in subspaces by RPLOF [16] and Isolation Forest [17]. However, clearly such simple heuristics might miss outliers due to the random selection of subspaces. Recently, a more meaningful selection has been proposed that selects only one subspace spanned as a hyperplane by a set of reference points (SOD) [13]. Its general hypothesis states that outliers deviate within this hyperplane. However, SOD determines the outlierness of an object only in this single subspace, if objects deviate in two or more subspaces SOD is unable to distinguish between their outlier factors.

Recent approaches base on subspace clustering as a related paradigm. As first approach one based on the subspace clustering method CLIQUE [2] to derive objects deviating from subspace clusters [1]. However, designed as a binary decision this approach does not provide an outlier ranking. As first outlier ranking based on subspace clusters a ranking function using cluster properties as indicators for outliers has been proposed (OutRank) [20]. However, based on the aggregated information of clusters one ignores the actual deviation of each object in the considered subspaces.

Overall, traditional full space outlier ranking methods simply compute object deviation in one fixed space and thus miss outliers in subspace projections. In contrast, outlier ranking in subspace projections take arbitrary projections into account. However, none of the proposed subspace methods considers a meaningful selection of relevant subspaces. Furthermore, they all ignore the incomparable deviation of objects in different subspaces for their outlierness measures.

### III. OUTLIER RANKING IN SUBSPACES

Our general idea is to measure deviation of each object in a set of relevant subspace projections. In contrast to traditional outlier ranking approaches, we consider for each object its deviation in multiple subspaces. This ensures to find objects deviating in projections of the data, but, it also poses new major challenges for outlier ranking. In the following we start with some basic notions and provide a formalization of these challenges before we propose our novel selection of significantly non-uniformly distributed subspaces and our novel adaptive outlierness measure.

#### A. Notions and Challenges

In general, the aim of outlier ranking is to provide a sorting of all objects  $o$  given in a database  $DB$ . Technically, one ranks according to the *degree of deviation* measured by a ranking function  $r : DB \rightarrow \mathbb{R}$ . The ranking function provides a real valued measure of the objects' outlierness. Ranking functions can be defined arbitrarily based on the object's features  $o = (o_1, \dots, o_d)$ . In contrast to traditional approaches that measure the degree of deviation in the full  $d$ -dimensional space  $D = \{1, \dots, d\}$ , we measure deviation in subspace projections  $S \subseteq D$ . Thus, we ensure to find outliers hidden in any possible subspace projection.

The general challenge for outlier ranking approaches, is to provide a meaningful ranking function which achieves to distinguish between an outlier object  $o$  and a regular object  $p$  by providing a clear distinction:  $r(o) \ll r(p)$ . However, traditional outlier ranking functions fail for outliers hidden in subspaces as they provide for all objects very similar ranking values  $r(o) \approx r(p) \forall o, p \in DB$ . This can be explained by the scattered full space of such data sets. While each object has only a subset of relevant attributes, the residual attributes provide more or less random values. Considering distances between objects using all of these attributes one observes an effect termed the "curse of dimensionality". As traditional ranking functions consider all attributes for distance

computation  $dist_D(o, p)$  they cannot distinguish outliers from regular objects. For the typically used Euclidean distance

$$dist_D(o, p) = \sqrt{\sum_{i \in D} (o_i - p_i)^2}$$

, distances between  $o \in DB$  and any residual objects grow more and more alike with increasing number of attributes  $|D| \rightarrow \infty$ :

$$\lim_{|D| \rightarrow \infty} \frac{\max_{p \in DB} dist_D(o, p) - \min_{p \in DB} dist_D(o, p)}{\min_{p \in DB} dist_D(o, p)} \rightarrow 0$$

As consequence, ranking values based on these full space distances become meaningless:

$$\lim_{|D| \rightarrow \infty} r(o) - r(p) \rightarrow 0 \Rightarrow r(o) \approx r(p) \forall o, p \in DB$$

Although outliers do not show up in full space, they deviate in subspace projections. Thus, we cope with the curse of dimensionality by considering the outlierness of each object in a selection of relevant subspaces. This set of relevant subspaces  $RS(o)$  is selected individually for each object  $o$  such that these subspaces provide a high contrast between  $o$  and its surrounding neighborhood. We measure the outlierness  $score(o, S)$  by restricting distance functions  $dist_S(o, p)$  to the subspace dimensions in  $S$ . The overall ranking value  $r(o)$  of an object  $o$  is then simply computed by aggregating its outlierness in all relevant subspaces:

#### Definition 1: Subspace Ranking Function

The overall ranking value  $r(o)$  of an object  $o \in DB$  w.r.t. a set of relevant subspaces  $RS(o)$  and an outlierness measure  $score(o, S)$  is defined as:

$$r(o) = \prod_{S \in RS(o)} score(o, S)$$

As aggregation of all outlierness measures in different subspaces one could use several meaningful functions. Please note, that we use scoring values in the range of  $0 \dots 1$  with outliers represented by low scores. Thus, the minimum over all scorings would provide a meaningful aggregation. However, this would highlight only the outlierness in one single subspace. Using the sum of scores as aggregation would lead to low contrast as objects found in clusters with high score values would blur the overall ranking value. In contrast, we use the product incorporating outlier properties from different subspaces such that low scores in multiple subspaces highlight an object as clear outlier providing high contrast between outliers and regular objects.

While traditional ranking functions consider the outlierness of an object only in the full space  $D$ , we aim at considering outlierness in a set of subspaces  $RS(o) \subseteq \mathcal{P}(D)$  out of the powerset of possible subspace projections. This is meaningful

as outliers might be hidden in multiple subspace projections. However, two novel challenges arise:

- How to choose the set of relevant subspaces  $RS(o)$  for meaningful outlier ranking
- How to achieve comparable outlierness values  $score(o, S)$  over multiple subspaces  $S \in RS(o)$ .

To tackle these challenges, our key hypothesis is that outliers can be distinguished in local neighborhoods of non-uniformly distributed subspaces. We base on the idea of local outlier ranking as proposed by full space approaches [6], [24]. According to this outlier mining paradigm, we define outliers as objects that are highly deviating from their local neighborhood. Traditional methods already observed varying density distributions and proposed local outlierness measures to tackle outlier ranking in full space. For subspace outlier ranking we observe an additional factor for varying densities not yet addressed in the literature. Different subspace provide highly varying densities ranging from densely clustered subspaces up to uniformly distributed subspace. To tackle this variance in densities we propose a statistical selection of subspaces and our novel adaptive outlierness measure for outlier ranking in relevant subspaces.

As a density-based approach OUTRES measures its outlierness  $score(o, S)$  according to the density  $den(o, S)$  of an object in subspace  $S$ . Low density values on an object indicate its outlierness and lead to low scores. However, objects might be outliers in multiple subspaces, thus, a meaningful outlierness measure has to be comparable over different subspaces. We will propose an instantiation of our outlierness function  $score(o, S)$  in Section III-C, where we will also give details about the underlying adaptive density  $den(o, S)$  for a comparable outlierness measurement. As a flexible approach any density estimation method could be used in our outlierness measure. Thus, for the general discussion in this section let us assume the typically used density in a local neighborhood defined as the number of objects in  $\varepsilon$ -distance to the object  $o$  in subspace  $S$ :

$$den(o, S) = |\mathcal{N}(o, S)| = \{p \mid dist_S(o, p) \leq \varepsilon\}$$

Based on this common density instantiation one can formally derive two major challenges for an outlier ranking in subspaces:

#### Challenge 1: Comparability of Outlierness

Outlierness measures are not comparable over multiple subspaces if: For subspace  $S, T \subseteq D$  with  $T \subset S$

- due to curse of dimensionality
- $\Rightarrow \forall p \in DB : dist_S(o, p) \geq dist_T(o, p)$
- $\Rightarrow den(o, S) \leq den(o, T)$
- $\Rightarrow score(o, S) \leq score(o, T)$
- $\Rightarrow$  outlierness is biased w.r.t. dimensionality

As density drops for increasing dimensionality, outlierness measures based on density in subspace projections are biased w.r.t. the dimensionality of the considered subspaces. As stated in Challenge 1, considering a subspace  $T$  and one of its higher

dimensional projections  $S$ , density drops from  $T$  to  $S$ . Thus, overall aggregation (cf. Def. 1) of outlierness is hindered by incomparable measures. Please keep in mind that we set low values in  $score(o, S)$  for highly deviating objects as we sort our ranking in ascending order. With such an incomparable measure, subspaces with many attributes would dominate the ranking value and outliers in low dimensional projections could not show up in the overall ranking. Thus, as we take multiple subspaces into account we have to provide an adaptive outlierness measure with comparable outlierness in arbitrary subspace projections to achieve a fair ranking of objects in any subspace.

#### Challenge 2: Relevance of Subspaces

A subspace  $S$  hinders the distinction of outliers if:

- $S$  is distributed uniformly random
- $\Rightarrow \forall o, p, q \in DB : dist_S(o, q) \approx dist_S(p, q)$
- $\Rightarrow \forall o, p \in DB : den(o, S) \approx den(p, S)$
- $\Rightarrow \forall o, p \in DB : score(o, S) \approx score(p, S)$
- $\Rightarrow$  distinction of outliers is hindered

Obviously the full space  $D$  is such an irrelevant subspace for increasing number of attributes  $|D| \rightarrow \infty$

With decreasing density, one reaches subspaces with uniformly distributed objects where outliers do not show up. Including such an irrelevant subspace projection  $S$  into a ranking function yields very similar ranking values for all objects. Thus, our key property for the set of relevant subspaces is to exclude subspaces which are distributed uniformly random.

#### B. Selection of Relevant Subspaces

First, we propose a statistical selection of the set of *relevant subspaces*  $RS(o)$  that can distinguish between the object  $o$  and its local neighborhoods in the selected subspaces. As motivated by Challenge 2, such a distinction based on the objects density is not possible in scattered subspaces that show uniformly random distributed data due to the low contrast between outliers and regular objects. Thus, we propose to exclude such scattered subspaces from outlier ranking by testing the underlying distribution in the local neighborhood  $\mathcal{N}(o, S)$ .

Our test is based on a statistical significance test aiming at reducing the probability that a uniformly distributed subspace passes into the set of relevant subspaces. We test against the null hypothesis that data is uniformly distributed with  $|\mathcal{N}(o, S)| \sim Binomial(|DB|, vol(\mathcal{N}(o, S)))$ . W.l.o.g. we assume that data is normalized to  $0 \dots 1$  such that the volume of  $\mathcal{N}(o, S)$  provides us the probability of observing one object in this neighborhood. As given for uniformly distributed data, the expected number of objects is then  $|DB| \cdot vol(\mathcal{N}(o, S))$ . Each object has equal probability of being in the neighborhood depending only on the neighborhoods volume. Based on this null hypothesis, we define  $H_0$  ( $S$  is irrelevant) and  $H_1$  ( $S$  is a relevant subspace) for object  $o$ . As uniformly distributed subspaces hinder the detection of meaningful outliers, this definition ensures with a given significance level  $\alpha$  that

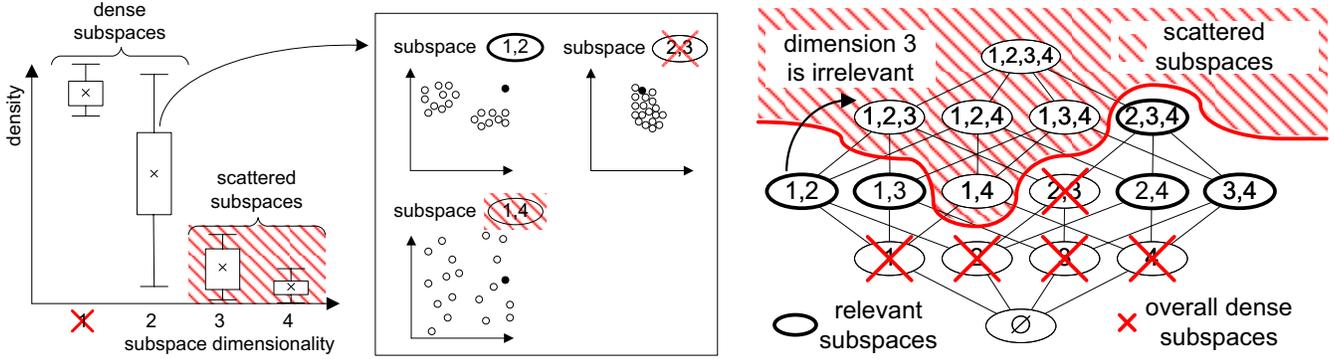


Fig. 2. Relevant subspace projections for outlier mining

uniformly distributed subspaces are only included into the ranking with a very low probability of less than  $\alpha$ . While for uniformly distributed data one expects  $|DB| \cdot \text{vol}(\mathcal{N}(o, S))$  many objects in the neighborhood, a relevant subspace should contain significantly more objects.

**Definition 2: Relevance Test**

For subspace  $S$  and neighborhood  $\mathcal{N}(o, S)$  we define hypotheses  $H_0$  and  $H_1$ :

- $H_0$  :  $S$  is distributed uniformly random in  $\mathcal{N}(o, S)$
- $H_1$  :  $S$  is distributed non-uniformly in  $\mathcal{N}(o, S)$

ensuring significantly low first error:

$$P(H_0 \text{ is rejected} \mid H_0 \text{ is true}) \leq \alpha$$

As statistical tool for testing uniform distribution we use the Kolmogorov-Smirnov goodness of fit test for the uniform distribution [28]. In recent mining tasks this test has shown good performance for subspace cluster detection [18]. In contrast to subspace clustering where one is interested in sets of objects grouped in a certain subspace, we select relevant subspaces for each object to distinguish between outliers and regular objects. For good ranking quality, we have to ensure that such uniform subspaces are only included in very rare cases by setting a low  $\alpha$  value. As significance level for the statistical hypothesis test we set  $\alpha = 0.01$ . Thus, the probability of wrongly rejecting the hypothesis  $H_0$  (the subspace is uniformly distributed) is only 1%, i.e. for one out of hundred uniform subspaces the test will make an error and state that this subspace is relevant. We will show the influence of the  $\alpha$  parameter for the overall outlier ranking quality in Section IV.

As illustrated on the left side of Figure 2, we exclude uniformly distributed subspaces for each object  $o$  individually. Starting considering 1d projections first, typically these low dimensional projects are uniformly dense. The whole database seems to be one dense region. Furthermore, outliers in 1d projections could be easily detected as pre-processing. By including more and more dimensions, due to correlations of the data, the database diverts in multiple dense regions. Density shows a high variance between dense clusters and deviating outliers. For our outlier ranking we only take the

outlierness of objects in these relevant subspaces into account. Adding even more dimensions the subspaces become scattered like the full space. All objects seem to be outliers.

For a toy example we have depicted three subspaces where a single hidden outlier is clearly deviating only in the relevant subspace  $\{1, 2\}$ . Measuring outlierness in this subspace yields a clear distinction between this outlier and its local neighborhood. Please note, it is crucial to exclude the scattered subspace  $\{1, 4\}$  as all objects seem outliers. Thus, outlierness measures would lose their contrast as low score values are provided for all objects. In contrast, subspace  $\{2, 3\}$  is not excluded by the relevance test as dense regions with high scores do not affect the contrast of Def. 1.

Our general idea is to include only subspaces which are distributed significantly different then the uniformly random distribution. Hence, we exclude the subspaces that do not provide any distinction between objects and hinder our outlier detection. A key observation for relevant subspaces is that with increasing number of attributes in a subspace  $S$ , one reaches subspaces with uniformly distributed objects where outliers do not show up any more. By including more and more attributes distances between objects grow more and more alike [5]. Thus, the selection of relevant subspaces can be reduced to the selection of significant attributes to be included in a given subspace projection  $S$ , as stated in the following corollary:

**Corollary 1: Uniformly distributed subspaces**

Let  $S = \{d_1, \dots, d_k\}$  be a subspace. Then it holds true:

$$S \text{ uniformly distributed} \Rightarrow \begin{aligned} & d_1 \text{ uniformly distributed} \\ & \wedge \dots \wedge \\ & d_k \text{ uniformly distributed} \end{aligned}$$

Consequently, by testing each attribute  $d_i$  we can assure that no uniformly distributed subspace is included in the set of relevant subspaces  $RS(o)$ . Moreover, we discard a subspace based on these insights as soon as at least one attribute is distributed uniformly random.

We base on statistical tests to detect significant subspaces by excluding uniformly distributed attributes from further consideration. We perform an incremental processing of the subspaces including in each step an additional attribute for the

considered subspace  $S$ . By adding attribute  $d_i$  to  $S$  we check if objects are uniformly distributed in  $d_i$ . We call an attribute  $d_i$  *relevant* for outlier ranking, if objects are significantly non-uniformly distributed. In contrast to other attributes, a relevant attribute might be added to  $S$  while preserving the clustered regions of subspace  $S$  also in subspace  $S \cup d_i$ . Summing up, we detect meaningful outliers by searching subspaces consisting only of relevant attributes containing clustered regions from which outliers can deviate.

Furthermore, as we aim at detecting outliers that deviate from clustered objects in their local neighborhood we check uniform distribution according to this neighborhood  $\mathcal{N}(o, S)$  and not w.r.t. the entire subspace. Based on this, our outlier ranking is computed for each object w.r.t. its subspace neighborhood, instead of the whole  $DB$ . Hence, the choice of relevant subspaces occurs strictly on the basis of the object locality. This is in contrast to subspace search approaches [25], which provide global subspace estimations supporting clustering with interesting projections. Compared to such approaches, the main advantage of our subspace selection is in the selection of locally significant subspaces for each object taking local deviation into account.

To illustrate the effects of relevant subspace selection, we depict a subspace lattice with all possible subspaces of a 4d data space in Figure 2. Starting considering 1d projections first, typically these low dimensional projects are uniformly dense, and thus, they do not affect our outlier ranking. In the relevant subspace projections highlighted in bold, we find the detect outliers deviating from their local neighborhood. As depicted we prune the higher-dimensional subspaces. For example, adding dimension 3 to subspace  $\{1, 2\}$  results in such an irrelevant subspace. Incrementally using the statistical test for each dimension we detect the irrelevant dimensions and stop further processing of higher dimensional subspaces. Formally, we define relevant subspaces  $RS(o)$  in Definition 3 to be the set of subspaces that are significantly non-uniformly distributed.

**Definition 3: Set of Relevant Subspaces**

The set of relevant subspaces contains subspaces that are significantly non-uniformly distributed:

$$RS(o) = \{S \in \mathcal{P}(D) \mid S \text{ passes } H_1\}$$

Only these subspaces are considered for our outlier ranking (cf. Def. 1).

Computing the outlierness of an object in its relevant subspaces  $RS(o)$  yields a high contrast between the object and its local neighborhood. Although we have excluded the irrelevant subspaces a major challenge remains. Subspaces in  $RS(o)$  have arbitrary dimensionality and show highly varying density values as motivated in Challenge 1.

*C. Adaptive Outlierness in Subspaces*

For a meaningful outlier ranking based on outlierness in multiple subspace projections the definition of  $score(o, S)$  has

to provide an adaptive outlierness measure as the overall ranking combines object properties out of very different subspaces  $S \in RS(o)$ . We propose such an *adaptive outlierness* measure by defining an *adaptive density* and a *local deviation* for each object.

1) *Adaptive object density*: As formalized in Challenge 1, measuring density in multiple subspaces leads to a challenging task, namely the strong dependence of densities on the number of attributes in the considered subspaces. For two subspaces  $S, T \subseteq D$  with  $T \subset S$  a simple counting of objects in a fixed neighborhood yields  $den(o, S) \leq den(o, T)$ . The main problem for density-based mining of different subspaces is the fixed neighborhood [3], [19], [23]. As distances between objects grow with increasing number of attributes, a fixed neighborhood  $\mathcal{N}(o, S) = \{p \mid dist_S(o, p) \leq \varepsilon\}$  becomes empty. All objects tend to have higher distance than the fixed  $\varepsilon$  parameter. To tackle this general problem of density estimation in arbitrary subspaces, we propose an adaptive density using a variable neighborhood. By increasing the neighborhood distance  $\varepsilon$  with increasing number of attributes, our density measure can automatically adapt to the expected data distribution. Thus, different subspaces become comparable and outlierness based on density estimation can automatically adapt to the number of attributes.

In general, we propose an adaptive neighborhood  $\mathcal{AN}(o, S)$  based on a variable  $\varepsilon(|S|)$  range.

$$\mathcal{AN}(o, S) = \{p \mid dist_S(o, p) \leq \varepsilon(|S|)\}$$

The general idea is to derive the variable range out of a common observation in subspace projections. While increasing the number of attributes in a subspace projection the volume of a fixed neighborhoods decreases significantly compared to the overall volume of the subspace. For example, consider the volume of the neighborhood covering the whole data range  $0 \dots 1$  of one attribute with  $\varepsilon = 0.5$ . If one keeps the neighborhood range fixed, the volume in subspace  $S$  is given by

$$vol(\mathcal{N}(o, S)) = \frac{\pi^{|S|/2}}{\Gamma(|S|/2 + 1)} \cdot 0.5^{|S|}$$

with the gamma function  $\Gamma(n + 1) = n \cdot \Gamma(n)$ ,  $\Gamma(1) = 1$ ,  $\Gamma(1/2) = \sqrt{\pi}$ . The volume decreases with increasing number of attributes.

$$vol(\mathcal{N}(o, S)) \ll vol(\mathcal{N}(o, T)) \text{ for } |T| < |S|$$

Thus, the expectation of detecting objects in such neighborhoods is decreasing as well, resulting in very low density estimations. Our variable neighborhood range adapts to this phenomenon. While for 1-dimensional subspaces neighborhood ranges are typically set to lower values  $\varepsilon \leq 0.5$ , in higher-dimensional subspaces the range should be increased with the number of dimensions. By increasing the range we ensure that the expected number of objects remains constant. Thus, we provide a comparable density estimation in arbitrary

subspaces. In our prior work, we have shown that such an adaptive neighborhood can be of general benefit for other mining paradigms as well, e.g. for density-based subspace clustering [19].

2) *Instantiation of adaptive density*: In the following we instantiate the basic idea of adaptive neighborhoods to a specific density estimation technique. As a flexible outlier model, OUTRES could be used with any density measure such as the simple counting of objects in the objects neighborhood (cf. Section III-A). However, we base our density measure on more enhanced and well established density estimation techniques [26]. As the overall density distribution of the data is not known in advance, density  $den(o, S)$  of an object  $o$  can be estimated by using kernel density estimators.

Each object  $o$  contributes to the overall density by a local impact defined by a kernel function  $K(x)$  with  $x = \frac{dist_S(o,p)}{h}$  being the scaled distance of any other object  $p$  to the object  $o$ . The bandwidth parameter  $h$  is used to scale the influence of each object to a maximal distance of  $h$ . The overall density for an object  $o$  is then simply the sum of kernel function over all objects in the database. As kernel function we use the Epanechnikov Kernel

$$K_e(x) = (1 - x^2), \quad x < 1$$

, providing optimal density estimation according to the mean integrated squared error [26]. Concludingly,  $den(o, S)$  is calculated by the formula:

$$den(o, S) = \frac{1}{|DB|} \sum_{p \in DB} K_e \left( \frac{dist_S(o,p)}{h} \right)$$

Since objects being farther away than  $h$  from a certain object do not contribute to its density, we obtain a local density on which our outlier detection is based.

In contrast to the simple counting of objects in the neighborhood given by  $|\mathcal{AN}(o, S)|$ , kernel density estimation has major benefits due to the weighted influence of each object. The sum of Epanechnikov Kernels provides a theoretically sound density definition. However, for density estimation in arbitrary subspace projections the fixed bandwidth  $h$  shows similar drawbacks to the fixed  $\varepsilon$  range. For comparable outlieriness over arbitrary subspaces, we propose to adapt the density by a variable kernel bandwidth  $\varepsilon(|S|)$ . As the true underlying density distribution is unknown, we only use the dimensionality of the space to derive the bandwidth for adaptive density estimation. For a fixed space with dimensionality  $d$  optimal bandwidth  $h_{optimal}(d)$  is given by the following formula:

$$h_{optimal}(d) = \left( \frac{8 \cdot \Gamma(\frac{d}{2} + 1)}{\pi^{\frac{d}{2}}} \cdot (d + 4) \cdot (2\sqrt{\pi})^d \right) \cdot n^{\frac{-1}{d+4}}$$

where  $n = |DB|$  is the database size and the gamma function as in the computation of the neighborhood volume [26].

As motivated in the previous paragraph, optimal bandwidth is computed based on the expected number of objects in a neighborhood. The formula for optimal bandwidth can simply

be seen as the optimal radius of an  $\varepsilon$ -sphere such that one yield statistically optimal density estimation results. For density estimation in subspaces this means that one has to choose a bandwidth for each individual subspace. Assuming that  $n$  is fixed in a static database, we observe  $h_{optimal}(d)$  to be a monotonically increasing function. So intuitively, for increasing dimensionality the influence (bandwidth) of each object is increased as well in order to maintain optimal density estimates, while the data space is becoming sparse. For comparable outlieriness we use the optimal bandwidth to adapt density estimation in arbitrary subspaces. By the user parameter  $\varepsilon$  we allow the user to quantify a notion of locality and adjust this value for arbitrary subspaces based on the optimal bandwidth. Formally, the bandwidth for a given number of attributes in a subspace  $S$  is defined by Definition 4:

**Definition 4: Adaptive neighborhood**

For a subspaces dimensionality  $|S|$ ,  $|S| \geq 2$ , the adaptive neighborhood  $\varepsilon(|S|)$  is defined by

$$\varepsilon(|S|) = \varepsilon \cdot \frac{h_{optimal}(|S|)}{h_{optimal}(2)}$$

Thus, we simply scale the given starting bandwidth  $\varepsilon$  from 2d space up to full data space and use these value for density estimation. In contrast to the fixed bandwidth in kernel density estimation we use our adaptive neighborhood as variable bandwidth for each individual subspace. Consequently, our automatic bandwidth adaption ensures comparable density estimates for arbitrary dimensional subspaces.

3) *Local object deviation*: For our outlier ranking based on deviations of density we first compute the density for each object and compare it with the local (average) density in a relevant subspace. By that, our approach is able to detect objects highly deviating from the residual data in a relevant subspace, i.e., objects having exceptionally low densities. While our adaptive density ensures comparability over multiple subspaces, our local deviation ensures meaningful outlieriness values inside one subspace. Hence, in addition to the adaptive density, we ensure to highlight an outlier with very low density compared to its local neighborhood in the considered subspace.

Having such a comparable density estimation, an outlier can be detected as an object showing significantly low density. As we aim at a local outlieriness we measure deviation based on an adaptive threshold. As first filter step we select only objects with significantly low density

$$den(o, S) < \mu - 2 \cdot \sigma$$

compared to  $\mu$  and  $\sigma$  as the mean and standard deviation of  $den(o, S)$  in the neighborhood of object  $o$ . From statistical observations, only very rare objects deviate more than two standard deviations from the mean value (cf. Chebyshev's inequality [8]). As statistical probability for such objects is low (e.g. for normal distributed data it is less than 2.1%), their outlieriness has to be high. Using mean  $\mu$  and standard deviation  $\sigma$  of the estimated (local) density we ensure to be

adaptive to varying density. Object deviation is then given by the following definition.

**Definition 5: Object deviation**

The deviation of an object  $o$  with respect to mean and standard deviation of the estimated density:

$$dev(o, S) = \frac{\mu - den(o, S)}{2 \cdot \sigma}$$

An object shows high deviation if its density compared to the average density  $\mu$  in its neighborhood  $\mathcal{AN}(o, S)$  is significantly low.

4) *Adaptive Outlierness*: Overall the outlierness of an object  $o$  has to fulfill two major requirements. First, it has to be adaptive to arbitrary dimensional subspaces. Thus, based on our adaptive object density we propose an adaptive outlierness which is comparable for different subspaces (cf. Sec. III-C2). Second, our adaptive outlierness has to cope with object deviation considering statistically deviation from the mean value (cf. Sec. III-C3). Incorporating both aspects in our adaptive outlierness measure we define  $score(o, S)$  as follows:

**Definition 6: Adaptive Outlierness**

The outlierness of an object  $o$  in subspace  $S$  is derived by its density and its deviation in this subspace:

$$score(o, S) = \begin{cases} \frac{den(o, S)}{dev(o, S)} & , \text{ if } dev(o, S) \geq 1 \\ 1 & , \text{ else.} \end{cases}$$

Our novel outlierness incorporates both aspects derived by density and the deviation of each object: low density and high deviation are both indicates for high outlierness. Highly deviating objects show up by  $dev(o, S) \geq 1$  as density is significantly low compared to mean and standard deviation.

Overall we cope with the different behaviors of objects in different subspaces: Scattered irrelevant subspaces are excluded by our relevance testing (cf. Sec. III-B). Objects in a dense subspace  $S$  result only in high density and almost no deviation such that we set  $score(o, S) = 1$  they do not affect the ranking value (cf. Def. 1). Only if objects show up with low density or high deviation in a relevant subspace they contribute to the overall ranking value (cf. Def. 6).

**D. Computation of OUTRES**

For the overall computation of our outlier ranking, first of all one has to select the relevant subspaces for each object. In a naive solution, one would test for each object  $o \in DB$  its local neighborhood  $\mathcal{AN}(o, S)$  in arbitrary subspaces  $S \in \mathcal{P}(D)$ . As the number of possible subspaces increases exponentially with the number of given attributes in  $D$  this selection is obviously not practically feasible. For each object, one would compute  $RS(o)$  out of  $2^{|D|}$  possible subspaces by using our relevance test. Furthermore, we require for all objects a density computation which yields a quadratic complexity w.r.t. the number of objects. Overall the complexity of outlier ranking based on relevant subspaces would be  $O(|DB| \cdot (2^{|D|} \cdot |DB|))$ .

Thus, for an efficient processing we propose an approximative selection of relevant subspaces based on a pruning heuristic. We process subspaces bottom-up and prune based

on the observation that having reached a sparse subspace with uniformly distributed data we may stop processing, as data is scattered even more in higher dimensional projections. As most important property, this pruning ensures to exclude all irrelevant subspace. This provides a high contrast for our outlier ranking as all considered subspaces are non-uniformly distributed. As an approximative selection we cannot ensure to include all non-uniformly distributed subspaces, but as highlighted also by our experiments we achieve high quality outlier ranking with this simple pruning heuristic. We even observe some redundant subspaces that actually are selected for  $RS(o)$ , but do not contribute to the overall outlier ranking. Thus, further enhancements for algorithmic solutions of our novel outlier ranking model seem to be promising tasks for future research. Especially, pruning of dense regions that do not show any hidden outliers might lead to even better runtimes. In addition to our pruning of irrelevant subspaces one might add a second filter to prune some of the relevant subspaces that might contribute only scores equal to 1 for the outlierness measure. Such an enhanced filtering could safely exclude further parts of the search space without any quality losses.

---

**Algorithm 1 OUTRES( $o, S$ )**

---

```

FOREACH  $i \in D \setminus S$ 
   $S' = S \cup \{i\}$ ;
  IF  $S'$  is relevant // relevance test (cf. Def. 2)
     $den(o, S') = \frac{1}{|DB|} \sum_{p \in \mathcal{AN}(o, S')} K_e(\frac{dist_{S'}(o, p)}{\epsilon(|S'|)})$ ;
     $dev(o, S') = \frac{\mu - den(o, S')}{2 \cdot \sigma}$ ;
    IF  $dev(o, S') \geq 1$  // high deviation (cf. Def. 6)
       $r(o) = r(o) \cdot \frac{den(o, S')}{dev(o, S')}$ ; // aggregation of scoring
    OUTRES( $o, S'$ ); // recursively next subspace
  ELSE
    // break recursion for higher dimensional subspaces

```

---

For the OUTRES algorithm, we test the relevance of subspaces in a bottom-up processing. We base on the observation that objects are dense in low dimensional spaces, while for higher dimensional spaces they diverge until they form an uniformly distributed scattered space. In Figure 2 we show a box-plot for the varying distribution of density in various dimensions as a toy example. Our bottom-up processing starts with low dimensional subspaces and tests step-by-step each subspace. The computation of the ranking  $r(o)$  based on the relevant subspaces of the object  $o$  is given in Algorithm 1. We start for each object  $o$  the recursive processing  $OUTRES(o, \{\})$ , recursion stops if an irrelevant subspace has been detected for object  $o$ . This ensures to exclude all irrelevant subspace projections, which would hinder the distinction of outliers in our ranking. In the worst case (if for all objects all subspaces are relevant) complexity remains  $O(2^{|D|} \cdot |DB|^2)$ , but in practical cases we yield efficient processing as we show in Section IV in addition to the high quality outlier ranking results.

#### IV. EXPERIMENTS

We demonstrate the quality of our OUTRES approach on both synthetic and real world data. We compare OUTRES to the well established LOF [6] and its recent extensions ABOF [15] as full space approaches. Furthermore, we compare against OutRank [20] and SOD [13] as the most recent outlier rankings based on subspace projections.

For comparability, we implemented all algorithms in our open-source framework [21]. By extending the popular WEKA framework we base our work on a widely used data input format for repeatable and expandable experiments. We used original implementations provided by the authors and best-effort re-implementations based on the original papers. We ensure comparable evaluations and repeatability of experiments, as we deploy all implemented algorithms on our website<sup>1</sup>. With our SOREX system [22], we ensure that all of our results will be reproducible, publicly available, and thus, might be used for comparison in future publications.

For fair comparison we base on objective quality measures. We believe that our evaluation setup provides a better quality assessment than showing only some examples of the detected outlier. In contrast to such a commonly used subjective evaluation, we highlight the achieved quality enhancement by three different quality measures. We measure true positive (*TPR*) and false positive (*FPR*) rates visualized in the well established ROC plot. Both of these measures are useful to derive if a ranking detects a high ratio of correct detected outliers (*TPR*) while providing only few non-outlier as detected outliers (*FPR*). However, they only take the ratio of detected outliers and non-outliers into account ignoring more or less the positioning of the objects in the ranking. Thus, we additionally evaluate the results with a ranking coefficient based on Spearman's Ranking Coefficient [27]. In contrast to the ROC plot, ranking coefficients take also the ranking positions of detected outliers into account. This leads to a more fine grained quality measure.

To illustrate the quality of the rankings we use the quality measures for the top- $k$  ranked objects (cf. Definition 7). The *TPR* measure is simply the fraction of found true outliers in the first  $k$  objects

$$\begin{aligned} \text{found true outliers}(R, k) = \\ \{o_{r_1} \dots o_{r_k}\} \cap DB_{\text{hidden outliers}} \end{aligned}$$

compared to the set of hidden outliers  $DB_{\text{hidden outliers}}$  in the database  $DB$ . Analogue, *FPR* is the fraction of found non-outliers in the first  $k$  objects

$$\begin{aligned} \text{found false outliers}(R, k) = \\ \{o_{r_1} \dots o_{r_k}\} \setminus \text{found true outliers}(R, k) \end{aligned}$$

compared to the overall set of non-outlier objects in the database.

##### Definition 7: *TPR and FPR measures*

The true positive rate for the first  $k$  objects of a ranking  $R = \{o_{r_1} \dots o_{r_n}\}$  is defined as:

$$TPR(R, k) = \frac{|\text{found true outliers}(R, k)|}{|DB_{\text{hidden outliers}}|}$$

The false positive rate is defined as:

$$FPR(R, k) = \frac{|\text{found false outliers}(R, k)|}{|DB_{\text{hidden non-outliers}}|}$$

More detailed measures can be derived by ranking coefficients [27]. Spearman's Ranking Coefficient  $SRC(R_1, R_2)$  computes the correlation of two given rankings  $R_1$  and  $R_2$ . We use  $SRC$  to measure the quality for one ranking by comparing it with the optimal ranking  $R_{\text{best}}$ , ranking all outliers first. Furthermore, we normalize with the ranking coefficient for the worst ranking  $R_{\text{worst}}$  having all outliers in the last positions. We define outlier ranking coefficient  $ORC(R, k)$  for the first  $k$  objects in ranking  $R$  as given in Definition 8.

##### Definition 8: **Ranking coefficient measure**

The outlier ranking coefficient for the first  $k$  objects of a ranking  $R = \{o_{r_1} \dots o_{r_n}\}$  is defined as:

$$ORC(R, k) = \frac{SRC(\{o_{r_1} \dots o_{r_k}\}, R_{\text{best}})}{SRC(R_{\text{worst}}, R_{\text{best}})}$$

For these measures the optimal ranking results in  $TPR(R_{\text{optimal}}, k) = 1 \wedge FPR(R_{\text{optimal}}, k) = 0$  and  $ORC(R_{\text{optimal}}, k) = 1$  for  $k = |DB|_{\text{outliers}}$

For non-optimal rankings  $TPR = 1$  is reached for larger  $k$  with  $FPR \gg 0$ , while the  $ORC$  measure does not reach the maximal value of 1 at all for non-optimal rankings. Thus, the  $ORC$  measure is more appropriate for evaluation of outlier rankings. By taking the actual positing of objects into account,  $ORC$  is able to distinguish between two rankings having found the same amount of outliers in the first  $k$  positions. In such a case,  $TPR$  and  $FPR$  show same results as they only consider the object ratio and cannot distinguish between these two rankings. Taking also positioning information into account  $ORC$  shows more fine grained differences in rankings. Especially, one can compare ranking quality by taking the overall  $ORC(R, |DB|)$  for comparison. Thus, after showing all three measures in the first experiment we use only the ranking coefficient measure for comparison in the following experiments.

##### A. Synthetic Data

For scalability experiments, we generate synthetic data following a method proposed in [11], [3] to generate density-based clusters in arbitrary subspaces. In addition, our generator adds outliers deviating from one of these subspace clusters. As there are no global patterns hidden in data, the hidden outliers do not appear in the scattered full space. In our first experiment, we evaluate the quality of the competing approaches

<sup>1</sup><http://dme.rwth-aachen.de/OpenSubspace/SOREX>

on a synthetic data set with 4765 objects represented by four subspace clusters each using 4 out the 16 given attributes and additionally 61 hidden outliers deviating from these clusters. Figure 3(a) illustrates the quality with respect to ROC plot. We observe that all approaches show high increase in true positive rates of detected outliers with only very few false positive. However, all hidden outliers ( $TPR = 1$ ) are found after thousands of considered objects, indicated by  $FPR \gg 0$ . Our novel OUTRES shows best performance compared to LOF, ABOF, SOD and OutRank, as it archives to detect more hidden outliers within the first ranked objects showing both higher  $TPR$  and lower  $FPR$  than the competitors. Comparing ROC plot and ranking coefficient in Figure 3 for the same experiment, we observe that OUTRES outperforms all competing approaches in both quality measures independent of the number of ranked objects. For ranking coefficient it always shows highest correlation with the optimal ranking. As ranking coefficient provides more information about the positioning of the objects than the ROC plot, we use only this measure in the following experiments.

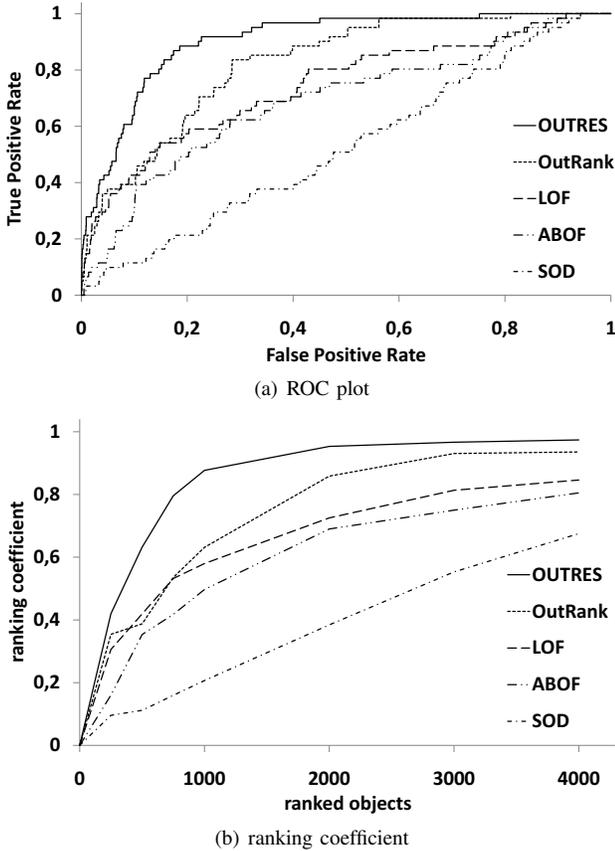


Fig. 3. Ranking quality on synthetic data

In our second experiment, we evaluate the scalability of outlier rankings with respect to the number of given attributes in the database. As outlier ranking in subspace projections aims to detect outliers hidden in any subset of the given attributes, scalability w.r.t. number of given attributes is crucial. We varied the number of attributes from 10 up to 50, while

keeping number of hidden subspace clusters, hidden outliers and database size constant, as in the previous experiment. Figure 4(a) shows the almost constantly high quality of our approach. As we add more and more attributes, hidden outliers disappear in the overall scattered full space. However, as OUTRES investigates only relevant subspace projections it scales w.r.t. number of given attributes. It outperforms all competing approaches in terms of quality.

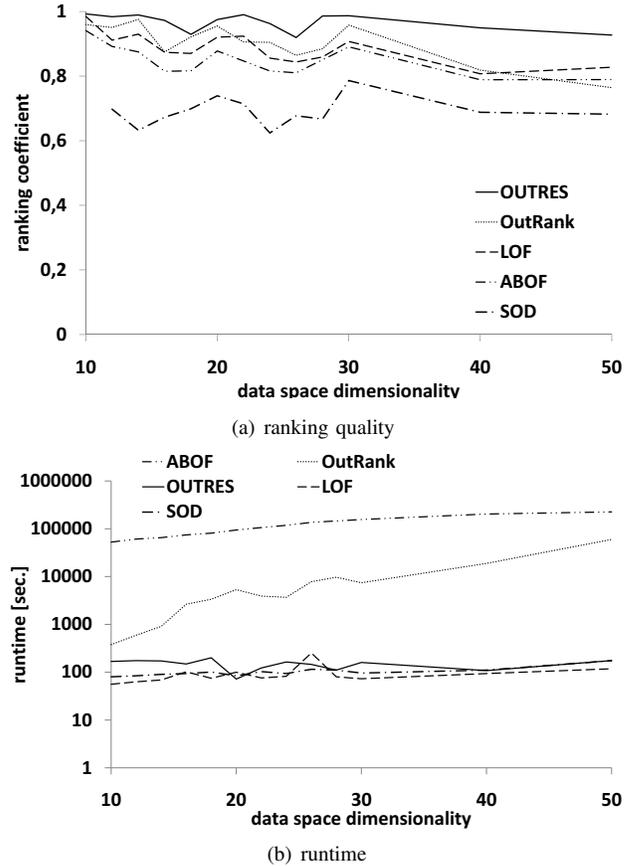


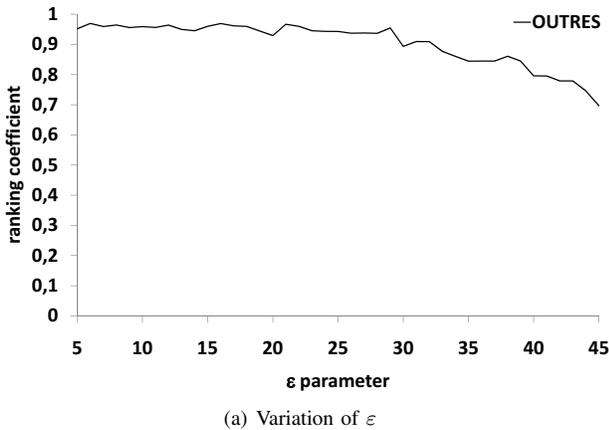
Fig. 4. Scalability w.r.t. number of attributes

In Figure 4(b) we compare the runtimes. In conjunction with the previous quality plot we observe that OUTRES achieves to perform both efficient outlier ranking and a high outlier ranking quality. Although OUTRES has to search for outliers in arbitrary subsets of the attributes, the proposed pruning heuristic by excluding uniformly distributed attributes shows both high quality results but also efficient computation. We skip further scalability experiments, as experiments have shown that database size has less impact on both quality and runtime. However, as an important issue we discuss parametrization of our approach in the following experiments.

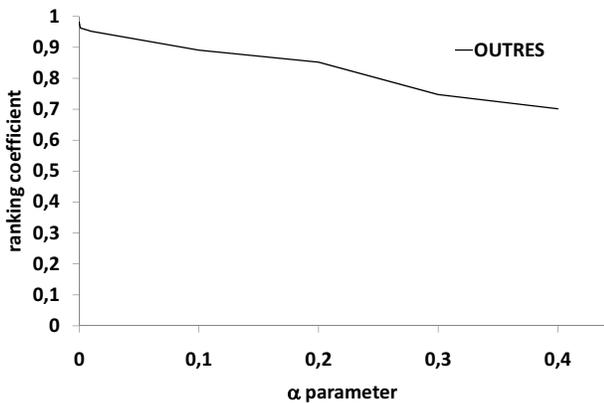
## B. Parametrization

For the two main parameters  $\alpha$  and  $\varepsilon$  we show the robustness of the ranking quality of OUTRES. On the synthetic data set from previous experiment, we varied the neighborhood parameter  $\varepsilon$  from 5 to 45 (data ranges from 0 to 100).

As depicted in Figure 5(a), OUTRES shows a quite robust ranking quality only slightly decreasing for high  $\varepsilon$  values. By increasing the neighborhood around each object density is increasing for all objects. Especially for outliers, density is becoming similar to clustered objects. Overall we achieve a robust approach w.r.t.  $\varepsilon$  due to our automatic adaption of the neighborhood range for the arbitrary subspace projections considered in OUTRES (cf. Def. 4). As default setting of  $\varepsilon$  in our experiments we use  $\varepsilon = 15$  showing best results. In general, with our adaptive neighborhood one can set the usual low  $\varepsilon$  neighborhood ranges in low dimensional subspace. These are increased automatically for higher-dimensional subspaces and provide a high quality density estimation for arbitrary subspace projections.



(a) Variation of  $\varepsilon$



(b) Variation of  $\alpha$

Fig. 5. Robustness of OUTRES w.r.t parameters

For the second parameter  $\alpha$  we observe similar effects. As depicted in Figure 5(b) we observe best ranking quality for low  $\alpha$  settings. For higher  $\alpha$  settings, OUTRES accepts more and more uniform distributed subspaces as relevant subspaces for outlier ranking. As one cannot distinguish between outliers and regular objects in these scattered subspaces, the overall ranking quality decreases. Keeping a low  $\alpha$  setting (default  $\alpha = 0.01$ ), thus, ensures to measure outlierness of objects only in relevant subspace projections.

### C. Real World Data

We analyzed the quality of outlier ranking on three real world data sets (Ionosphere, Breast Cancer and Pendigits) from the UCI repository [7]. All of these data sets provide scattered full spaces, while subsets of the given attributes can be used to distinguish between the hidden patterns and outlying objects. For example, in the Pendigits data set objects are described by  $(x,y)$  positions concatenated in a digit trajectory. Clearly not all of the pen positions are important to detect outlying objects. Some digits deviate significantly in the first position (first two given attributes) from the residual objects starting typically at similar positions (upper left area). Similarly, also in the other data sets outliers can be distinguished from regular objects using subspace projections of the database. For our evaluation measures, we used one of the class labels reduced to 10% of its size as ground truth for hidden outliers. In contrast to adding artificial outliers into the database, such a reduction of the original data distribution seems more natural. The remaining objects of the reduced class show high deviation from other classes and have low density due to the eliminated parts of their own class. In our experiments, we show that outlier ranking approaches successfully detects these very rare hidden observations in subspaces of the given databases.

In Figure 6 we show the ranking coefficients for the real world databases. For all data sets we observe a high ranking quality of OUTRES, outperforming competing approaches by detecting outliers as top ranked objects. Hidden outliers are clearly distinguished by the selection of relevant subspaces. For example, in the pendigits data one object representing the digit four has been first ranked due to its high deviation in the last positions of its trajectory. Overall, our novel OUTRES approach provides highest quality results, while the competing approaches show varying quality over multiple data sets.

## V. CONCLUSION

In this work, we proposed a novel outlier ranking for objects deviating in subspace projections. The OUTRES approach computes local density deviation by looking at a selection of relevant subspaces for each object. Relevance of subspaces is measured by statistical significance tests. Thus, only relevant subspaces that are not distributed uniformly random are used for our outlier ranking. For comparable outlierness measures in different subspaces, we derive an adaptive density measure which automatically adapts to the considered subspace. OUTRES computes an overall high quality outlier ranking by aggregating this adaptive outlierness of objects in relevant subspaces. Our thorough evaluation on both synthetic and real world data shows that OUTRES outperforms competing outlier ranking approaches. Especially, OUTRES achieves to detect outliers hidden in subspace projections.

## ACKNOWLEDGMENT

This work has been supported in part by the UMIC Research Centre, RWTH Aachen University, Germany.

## REFERENCES

- [1] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *SIGMOD*, 2001, pp. 37–46.
- [2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *SIGMOD*, 1998, pp. 94–105.
- [3] I. Assent, R. Krieger, E. Müller, and T. Seidl, "DUSC: Dimensionality unbiased subspace clustering," in *ICDM*, 2007, pp. 409–414.
- [4] V. Barnett and T. Lewis, *Outliers in statistical data*. John Wiley, 1994.
- [5] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is nearest neighbors meaningful," in *IDBT*, 1999, pp. 217–235.
- [6] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *SIGMOD*, 2000, pp. 93–104.
- [7] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [8] G. Hardy, J. Littlewood, and G. Polya, *Inequalities*. Cambridge University Press, 1988.
- [9] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1641–1650, 2003.
- [10] I. Jolliffe, *Principal Component Analysis*. Springer, New York, 1986.
- [11] K. Kailing, H.-P. Kriegel, and P. Kröger, "Density-connected subspace clustering for high-dimensional data," in *SDM*, 2004, pp. 246–257.
- [12] E. Knorr, R. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *VLDB Journal*, vol. 8, no. 3, pp. 237–253, 2000.
- [13] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Outlier detection in axis-parallel subspaces of high dimensional data," in *PAKDD*, 2009, pp. 831–838.
- [14] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM TKDD*, vol. 3, no. 1, pp. 1–58, 2009.
- [15] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in *KDD*, 2008, pp. 444–452.
- [16] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in *KDD*, 2005, pp. 157–166.
- [17] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *ICDM*, 2008, pp. 413–422.
- [18] G. Moise and J. Sander, "Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering," in *KDD*, 2008, pp. 533–541.
- [19] E. Müller, I. Assent, S. Günemann, R. Krieger, and T. Seidl, "Relevant Subspace Clustering: mining the most interesting non-redundant concepts in high dimensional data," in *ICDM*, 2009, pp. 377–386.
- [20] E. Müller, I. Assent, U. Steinhausen, and T. Seidl, "Outrank: Ranking outliers in high dimensional data," in *DBRank Workshop*, 2008, pp. 600–603.
- [21] E. Müller, S. Günemann, I. Assent, and T. Seidl, "Evaluating clustering in subspace projections of high dimensional data," in *VLDB*, 2009, pp. 1270–1281.
- [22] E. Müller, M. Schiffer, P. Gerwert, M. Hannen, T. Jansen, and T. Seidl, "SOREX: Subspace outlier ranking exploration toolkit," in *ECML PKDD*, 2010, pp. 607–610.
- [23] E. Müller, M. Schiffer, and T. Seidl, "Adaptive outlierness for subspace outlier ranking," in *CIKM*, 2010, pp. 1629–1632.
- [24] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in *ICDE*, 2003, pp. 315–326.
- [25] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 90–105, 2004.
- [26] B. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.
- [27] C. Spearman, "The proof and measurement of association between two things," *American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1987.
- [28] M. Stephens, "Use of the Kolmogorov-Smirnov, Cramer-von Mises and related statistics without extensive tables," *Journal of the Royal Statistical Society. Series B*, pp. 115–122, 1970.

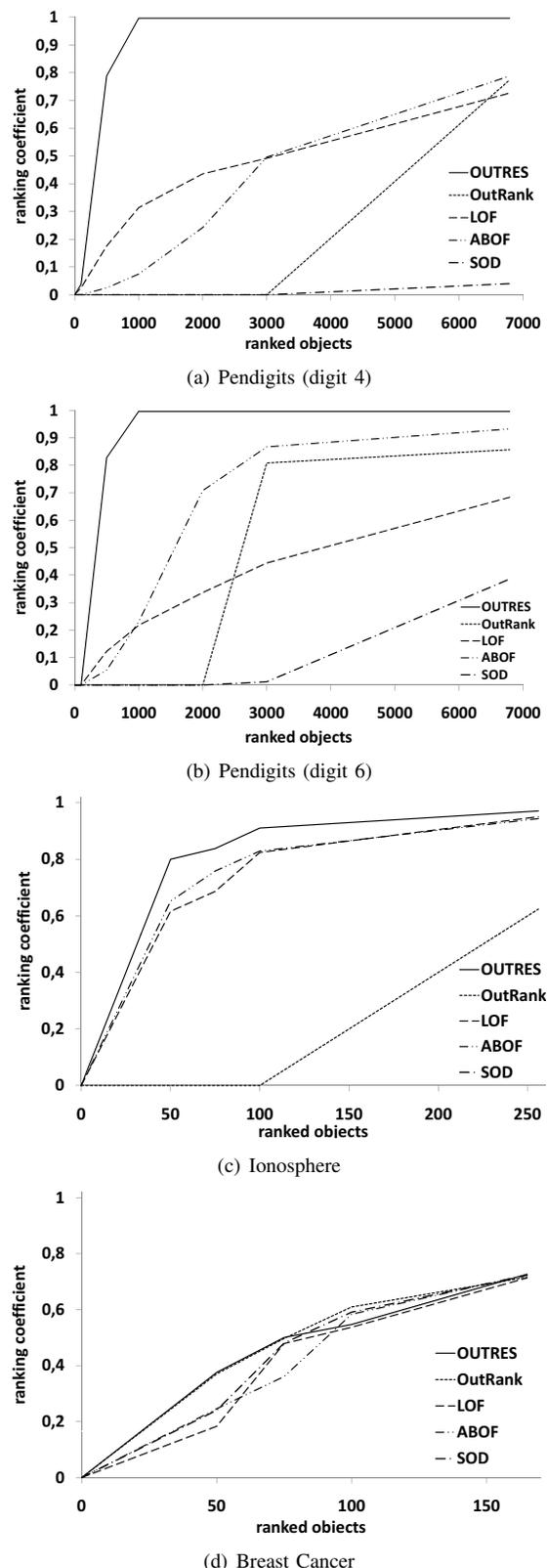


Fig. 6. Ranking quality on real world data