L1-Depth Revisited: A Robust Angle-Based Outlier Factor in High-Dimensional Space

Ninh Pham(✉)

Department of Computer Science,
University of Copenhagen, Copenhagen, Denmark
pham@di.ku.dk

Abstract. Angle-based outlier detection (ABOD) has been recently emerged as an effective method to detect outliers in high dimensions. Instead of examining neighborhoods as proximity-based concepts, ABOD assesses the broadness of angle spectrum of a point as an outlier factor. Despite being a *parameter-free* and robust measure in high-dimensional space, the exact solution of ABOD suffers from the cubic cost $O(n^3)$ regarding the data size n, hence cannot be used on large-scale data sets.

In this work we present a *conceptual* relationship between the ABOD intuition and the L1-depth concept in statistics, one of the earliest methods used for detecting outliers. Deriving from this relationship, we propose to use L1-depth as a variant of angle-based outlier factors, since it only requires a quadratic computational time as proximity-based outlier factors. Empirically, L1-depth is competitive (often superior) to proximity-based and other proposed angle-based outlier factors on detecting high-dimensional outliers regarding both efficiency and accuracy.

In order to avoid the quadratic computational time, we introduce a simple but efficient sampling method named *SamDepth* for estimating L1-depth measure. We also present theoretical analysis to guarantee the reliability of SamDepth. The empirical experiments on many real-world high-dimensional data sets demonstrate that SamDepth with $\sqrt{n}$ samples often achieves very competitive accuracy and runs several orders of magnitude faster than other proximity-based and ABOD competitors. Data related to this paper are available at: https://www.dropbox.com/s/nk7nqmwmdsatizs/Datasets.zip. Code related to this paper is available at: https://github.com/NinhPham/Outlier.

1 Introduction

Outlier detection is the process of detecting anomalous patterns that do not conform to an expected behavior. According to Hawkins [8], an outlier would be

Research supported by the Innovation Fund Denmark through the DABAI project.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-10925-7_7) contains supplementary material, which is available to authorized users.

M. Berlingerio et al. (Eds.): ECML PKDD 2018, LNAI 11051, pp. 105–121, 2019.
https://doi.org/10.1007/978-3-030-10925-7_7

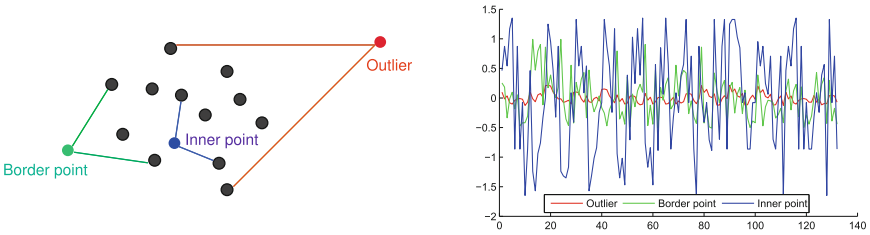

Fig. 1. Variance of angles of different types of points. Outliers have small variances whereas inliers have large variances.

"an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". Detecting such outlier patterns is a fundamental and well studied data mining task due to its several application domains, such as fraud detection in finance, author verification for forensic investigation, and detecting anomalous patterns for medical diagnosis.

One of the earliest methods to detect outliers is based the concept of depth in statistics [11,18] due to the natural correlation between depth-based measure and outlierness measure. The depth-based methods organize the data points in many layers, with the expectation that the "deepest" layers with large depth values contain points close to the center of the point cloud and outliers are likely to appear in the shallow layers with small depth values. However, since most outlier detection applications often arise in high-dimensional domains and most of depth-based methods do not scale up with data dimensionality [9], depth-based approaches suffer from a computational bottleneck for searching for high-dimensional outliers.

Since then, several outlierness measures based on the notion of proximity have been proposed to detect high-dimensional outliers. Due to the phenomenon "curse of dimensionality", proximity-based approaches in the literature which are implicitly or explicitly based on the concept of proximity in Euclidean distance metric between points in full-dimensional space do not work efficiently and effectively. Traditional solutions to detect distance-based global outliers [4,13,19] and density-based local outliers [5,16] often suffer from the high computational cost due to their core operation, near(est) neighbor search in high dimensions. Moreover, the high-dimensional data is often very sparse and therefore the measures like distances or nearest neighbors may not be qualitatively meaningful [1,3].

In order to alleviate the effects of the "curse of dimensionality", Kriegel et al. [14] proposed a novel outlier ranking approach based on the broadness of angle spectrum of data. The approach named *Angle-based Outlier Detection (ABOD)* evaluates the degree of outlierness on the *variance of the angles (VOA)* between a point and all other pairs of points in the data set. The intuition of ABOD, as shown in Fig. 1, is that the smaller the angle variance of the point has, the more likely it is an outlier. Since angles are more stable than distances, the ABOD approach does not substantially deteriorate in high-dimensional data. It is worth noting that the proposed outlierness measure in [14], called *ABOF*, does

not deal directly with the intuition of variance of angles. Indeed, ABOF assesses the weighted variance of weighted cosine of angles where the both weights are the corresponding distances between the assessed point and other pairs of points. The variant notion ABOF with weight of distances is more robust than the original intuition VOA in low-dimensional space since it allows distance affects the outlierness measure.

Despite many advantages of alleviating the effects of the "curse of dimensionality" and being a *parameter-free* measure, there are two intrinsic drawbacks with the ABOD approaches.

- There is no theoretical foundation connecting to the ABOD observation so it is difficult to understand and explain the outlierness behaviors detected by ABOD.
- The cubic time complexity taken to compute angle-based measures is very significant and problematic for many applications with large-scale data sets.

To avoid the cubic time complexity, Kriegel et al. [14] also proposed a heuristic approximation variant of ABOF, called *approxABOF*, for efficient computations. Instead of computing ABOF over all other pairs in the point set, approxABOF computes ABOF value over all pairs in the k-nearest neighbor (kNN) set. Hence, approxABOF requires a quadratic time complexity used in sequential search for kNN. Moreover, there is no analysis on the approximation error between approxABOF and ABOF and hence the reliability of detecting outliers using approxABOF is not guaranteed.

Recently, Pham and Pagh [17] investigated the variance of angles (VOA) as an ABOD outlierness measure and proposed an efficient algorithm for the ABOD approach. They proved that VOA is well preserved under random projections and introduced *FastVOA*, a near-linear time algorithm for estimating VOA for the point set. Despite many advantages of the fast running time and the quality of approximation guarantee, FastVOA might introduce large approximation errors. The large approximation errors result in detection performance degradation when the VOA gap between outliers and inliers is rather small. Furthermore, large approximation error of estimation can be problematic when combining VOA with other outlier factors to build outlier detection ensembles [2,24].

In this work, we investigate both mentioned drawbacks of the ABOD method. We examine the first drawback via a well-established concept of *data depth* in statistics [18,22]. In particular, we consider the *L1-depth* notion [21,23], which intuitively measures how much additional probability mass needed for moving a point in a set to the multivariate median of its point set. We study the notion of L1-depth in details and provide a strong conceptual relationship between L1-depth and the ABOD observation. Deriving from this relationship, we propose to use L1-depth as a variant of angle-based outlier factor since it requires a *quadratic* time computation as proximity-based outlier factors. Empirically, ABOD using L1-depth is superior to using VOA and ABOF, i.e. the computational cost is much smaller and the outlier detection accuracy is much higher.

To overcome the drawback of quadratic computational time, we introduce a simple but efficient sampling method named *SamDepth* for estimating

L1-depth measure. The empirical experiments on many real-world high-dimensional data sets demonstrate that SamDepth often runs much faster, provide smaller approximation errors and therefore more accurate outlier rankings than other ABOD competitors. Especially, SamDepth with $\sqrt{n}$ samples where n is the data size achieves very competitive accuracy and runs several orders of magnitude faster than other proximity-based and ABOD methods on several large-scale data sets.

2 Notation and Background

Given a point set $\mathbb{S} \subseteq \mathcal{R}^d$ of size n and a point $\boldsymbol{p} \in \mathbb{S}$, we denote by $P = \mathbb{S} \setminus \{\boldsymbol{p}\}$ since most of outlier factors of $\boldsymbol{p}$ are evaluated on the set P. We also denote by $(\boldsymbol{a}, \boldsymbol{b})$ a pair of any two *different* points in P. As we will elaborate later, ABOD algorithms compute outlier factors of $\boldsymbol{p}$ by values dependent on $\boldsymbol{p}$ and *each* pair $(\boldsymbol{a}, \boldsymbol{b})$. Hence, we will use the notation $\sum_{\boldsymbol{a},\boldsymbol{b}}$ for short to represent for the summation on $\boldsymbol{a}, \boldsymbol{b} \in P$.

For a given pair $(\boldsymbol{a}, \boldsymbol{b})$, we denote by Θ_{apb} the angle between the difference vectors $\boldsymbol{p} - \boldsymbol{a}$ and $\boldsymbol{p} - \boldsymbol{b}$. Since we will show a conceptual relationship between L1-depth measure and the variance of angles (VOA), an ABOD outlier factor, we will describe VOA and discuss about the time complexity of a naïve algorithm to compute exactly this measure.

Definition 1. *The variance of angle of a point $\boldsymbol{p}$ is computed via the first moment MOA1 and the second moment MOA2 of the angle Θ_{apb} between $\boldsymbol{p}$ and each pair $(\boldsymbol{a}, \boldsymbol{b})$. That is*

$$VOA(\boldsymbol{p}) = \boldsymbol{Var}[\Theta_{apb}] = MOA2(\boldsymbol{p}) - (MOA1(\boldsymbol{p}))^2$$

where $MOA2(\boldsymbol{p})$ and $MOA1(\boldsymbol{p})$ are defined as follows:

$$MOA2(\boldsymbol{p}) = \frac{\sum_{\boldsymbol{a},\boldsymbol{b}} \Theta_{apb}^2}{(n-1)(n-2)}; MOA1(\boldsymbol{p}) = \frac{\sum_{\boldsymbol{a},\boldsymbol{b}} \Theta_{apb}}{(n-1)(n-2)}.$$

Note that the VOA definition is identical to the definition in [17] since we take into account both Θ_{apb} and Θ_{bpa}. A naïve algorithm computing VOA for n points takes $O(n^3)$ time since computing $VOA(\boldsymbol{p})$ for each $\boldsymbol{p}$ takes $O(n^2)$ time. Note that VOA values are often very small and this challenges approximation methods to have good approximation errors in order to preserve the ABOD ranking.

3 L1-Depth and Its Conceptual Relationship with the ABOD Intuition

This section will study the L1-depth concept in statistics and provide a conceptual relationship between L1-depth concept and variance of angles of the ABOD intuition. We also discuss some benefits derived from this relationship.

3.1 L1-Depth as an ABOD Measure

Vardi and Zhang [23] studied the multivariate L1-median point (i.e. the point that minimizes the weighted sum of the Euclidean distances to all points in a high-dimensional cloud). Associating to the multivariate L1-median concept, they also proposed a simple close-form formula for the data depth called L1-depth function. The L1-depth function shares the same spirit with other proposed data depth [18], that is deeper points with larger depth are relatively closer to the center of the cloud (i.e. the L1-median).

Of the various depth notions, L1-depth is computationally efficient, i.e. $O(n)$ for each point. It has been used in clustering and classification tasks for microarray gene expression data [12] and novelty detection in taxonomic applications [7]. The definition of L1-depth (L1D) is as follows:

Definition 2 ([7, Eq. 3], [23, Eq. 4.3]). *L1-depth*

$$L1D(\boldsymbol{p}) = 1 - \frac{1}{n-1} \left\| \sum_{\boldsymbol{a} \in P} \frac{\boldsymbol{p} - \boldsymbol{a}}{\|\boldsymbol{p} - \boldsymbol{a}\|} \right\|$$

It is clear that L1-depth shares the same spirit as VOA and ABOF on dealing with the angle spectrum and being a parameter-free measure but has more efficient computation, i.e. $O(n)$ time for each point. In particular, the intuition of L1-depth concept is very similar to the ABOD idea since it assesses the broadness of *directions* of distance vectors, and the smaller $L1D(\boldsymbol{p})$ is, the more likely $\boldsymbol{p}$ is an outlier. Considering again Fig. 1, for inliers within the cluster, their L1D values will be close to 1. However, the L1D value of the outlier tends to be close to 0 since most the other points locate in some particular direction.

The following lemma shows that L1-depth can be derived from the sum of cosine of angles between a point $\boldsymbol{p}$ and all other pairs of points. This lemma also sheds the light on the conceptual relationship between L1-depth and variance of angles in the ABOD methods.

Lemma 3.

$$(1 - L1D(\boldsymbol{p}))^2 = \frac{1}{n-1} + \frac{1}{(n-1)^2} \sum_{\boldsymbol{a},\boldsymbol{b}} \cos \Theta_{apb}$$

Proof. Using the extension

$$\left\| \sum_i \boldsymbol{x}_i \right\|^2 = \sum_i \|\boldsymbol{x}_i\|^2 + \sum_{i \neq j} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle,$$

we have:

$$
\begin{aligned}
(1 - L1D(\boldsymbol{p}))^2 &= \frac{1}{(n-1)^2} \left\| \sum_{\boldsymbol{a} \in P} \frac{\boldsymbol{p} - \boldsymbol{a}}{\|\boldsymbol{p} - \boldsymbol{a}\|} \right\|^2 \\
&= \frac{1}{(n-1)^2} \left(\sum_{\boldsymbol{a} \in P} \left\| \frac{\boldsymbol{p} - \boldsymbol{a}}{\|\boldsymbol{p} - \boldsymbol{a}\|} \right\|^2 + \sum_{\boldsymbol{a},\boldsymbol{b}} \left\langle \frac{\boldsymbol{p} - \boldsymbol{a}}{\|\boldsymbol{p} - \boldsymbol{a}\|}, \frac{\boldsymbol{p} - \boldsymbol{b}}{\|\boldsymbol{p} - \boldsymbol{b}\|} \right\rangle \right) \\
&= \frac{1}{(n-1)^2} \left((n-1) + \sum_{\boldsymbol{a},\boldsymbol{b}} \cos \Theta_{apb} \right) = \frac{1}{n-1} + \frac{1}{(n-1)^2} \sum_{\boldsymbol{a},\boldsymbol{b}} \cos \Theta_{apb}
\end{aligned}
$$

□

Intuitively, since $\cos(x)$ is a strictly monotonically decreasing function in the range $[0, \pi]$, L1D will be correlated to the first moment of angles MOA1. Hence, the outlier ranking produced by L1D is highly positively correlated to MOA1's one. Mathematically, we can exploit the Taylor series approximation $\cos(x) \approx 1 - x^2/2$ on Lemma 3 to show the relationship between L1D and the second moment MOA2 as follows.

$$
\begin{aligned}
(1 - L1D(\boldsymbol{p}))^2 &= \frac{1}{(n-1)^2} \left((n-1) + \sum_{\boldsymbol{a},\boldsymbol{b}} \cos \Theta_{apb} \right) \\
&\approx \frac{1}{(n-1)^2} \left((n-1) + \sum_{\boldsymbol{a},\boldsymbol{b}} \left(1 - \frac{\Theta_{apb}^2}{2} \right) \right) \\
&= \frac{1}{(n-1)^2} \left((n-1)^2 - \sum_{\boldsymbol{a},\boldsymbol{b}} \frac{\Theta_{apb}^2}{2} \right) \\
&= 1 - \frac{1}{(n-1)^2} \sum_{\boldsymbol{a},\boldsymbol{b}} \frac{\Theta_{apb}^2}{2} \\
&= 1 - \frac{(n-1)(n-2) MOA2(p)}{2(n-1)^2} \\
&= 1 - \frac{n-2}{2(n-1)} MOA2(p). \qquad (1)
\end{aligned}
$$

When the Taylor series approximation $\cos(x) \approx 1 - x^2/2$ provides a small error, the outlier factor L1D is highly proportional to MOA2. Therefore, we can use both VOA, the *central* second moment and MOA2, the second moment of angles as angle-based outlier factors, and the smaller $MOA2(\boldsymbol{p})$ or $VOA(\boldsymbol{p})$ is, the more likely $\boldsymbol{p}$ is an outlier.

3.2 An Empirical Study and Benefits from the Conceptual Relationship

In order to confirm our theoretical finding, we compute the exact VOA, MOA1, MOA2, and L1D values on a synthetic data set generated by a Gaussian mixture

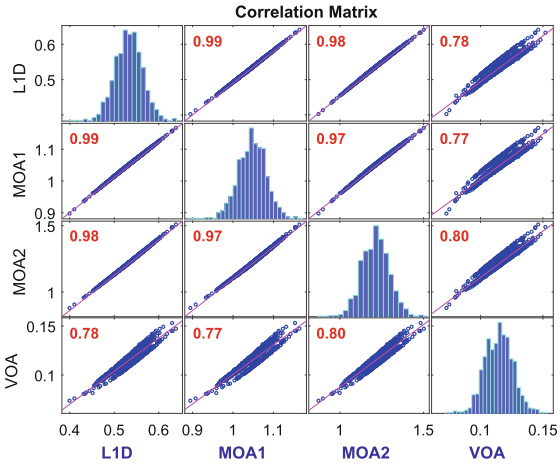

Fig. 2. Kendall's rank correlation between L1D, MOA1, MOA2, VOA measures on the synthetic data set.

as provided in the original ABOD paper [14]. The 100-dimensional synthetic data set contains 1000 inliers generated by independent Gaussian distributions and 10 outliers generated by a uniform distribution. Figure 2 shows very high Kendall's rank correlation coefficients between L1D, MOA1, MOA2 and VOA on the data set. We note that the Spearman's rank correlation coefficients among them are even higher and not reported here.

It is very clear that the outlier ranking based on L1D is almost identical to that of MOA1, MOA2 and very highly correlated to that of VOA. This means that instead of using VOA, the *central* second moment of angle distribution, we can use L1D, an approximation of the second moment with less computational resource for outlier ranking in high-dimensional data sets.

Next we discuss about the two benefits from this observed relationship, including an algorithmic benefit for approximating VOA and an application benefit on using ABOD measures for other data analytics tasks.

Algorithmic Benefit: Note that the time complexity of computing $L1D(\boldsymbol{p})$ is $O(n)$. Hence we can use $L1D(\boldsymbol{p})$ to derive an approximation of the second moment $MOA2(\boldsymbol{p})$ (see Eq. 1), which can replace the main computational resource in FastVOA [17]. Also note that FastVOA approximates the first moment $MOA1(\boldsymbol{p})$ in $O(n \log n)$ time. Combining these two approximations, we can estimate VOA of all n points with high accuracy in quadratic time without utilizing the AMS Sketches. In small data sets, this combination runs faster and provides better outlier detection performance than FastVOA [17].

Application Benefit: Since L1-depth is a variant of ABOD measures and since depth notions have been used in clustering and classification [10,12], we can use other ABOD measures, including VOA and ABOF on these settings. For example in classification tasks, instead of using kNN relationship, we can use ABOD

measures to assign the label to the test data. The basic classification rule is that a test data will be assigned into a class that maximizes its ABOD measure. Therefore, it is necessary to reduce the cost of computing ABOD measures to avoid the computational bottleneck for these data analytic tasks.

4 Sampling Algorithms for L1-Depth

Since computing exactly L1D for each point takes $O(dn)$ time in a data set of size n with d dimensions, we propose efficient sampling algorithms to approximate L1D for speeding up the outlier detection process. We also show theoretical analysis to guarantee the reliability of our sampling algorithms.

As can be seen from Lemma 3, we can approximate L1D from an accurate estimate of the mean of cosine of angles $\mu = \frac{\sum_{\boldsymbol{a},\boldsymbol{b}} \cos\Theta_{apb}}{(n-1)(n-2)}$. Our standard sampling method called *BasicSam* is that, given a point $\boldsymbol{p}$, we randomly sample a pair $(\boldsymbol{a}, \boldsymbol{b})$ in P and define a random variable

$$X = \begin{cases} \cos\Theta_{apb} & \text{if } \boldsymbol{a} \text{ and } \boldsymbol{b} \text{ are chosen;} \\ 0 & \text{otherwise.} \end{cases}$$

Then, we have $\mathbf{E}[X] = \mu$. Using Hoeffding's inequality with $t = O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ independent random sample *pairs*, we can guarantee an absolute approximation error at most ϵ with probability at least $1 - \delta$. We note that for t random pairs BasicSam takes $O(dt)$ time since it needs to compute $2t$ difference vectors $\boldsymbol{p} - \boldsymbol{a}$ and $\boldsymbol{p} - \boldsymbol{b}$.

As can be seen from Definition 2, computing directly $L1D(\boldsymbol{p})$ takes $O(dn)$ time. We now exploit this property to avoid sampling random pairs and propose *SamDepth*, a more efficient sampling method for L1D, as shown in Algorithm 1.

Algorithm 1. SamDepth (n)

Input: A data set $\mathbb{S}$ of size n and a point $\boldsymbol{p} \in \mathbb{S}$
Output: An estimate of $L1D(\boldsymbol{p})$

1 Sample without replacement a subset $S \subset \mathbb{S} \backslash \{\boldsymbol{p}\}$ of $t = \sqrt{n}$ points ;
2 Compute the norm $m = \left\| \sum_{\boldsymbol{a} \in S} \frac{\boldsymbol{p}-\boldsymbol{a}}{\|\boldsymbol{p}-\boldsymbol{a}\|} \right\|^2$;
3 Output $1 - \sqrt{\frac{1}{n-1} + \frac{n-2}{n-1}\left(\frac{m}{t(t-1)} - \frac{1}{t-1}\right)}$ as an estimate of $L1D(\boldsymbol{p})$;

Instead of sampling a random pair, we sample without replacement a subset $S \subset P$ of t points, and define a random variable $Z = \sum_{\boldsymbol{a},\boldsymbol{b}} Z_{ab}$ where

$$Z_{ab} = \begin{cases} \cos\Theta_{apb}/t(t-1) & \text{if } \boldsymbol{a} \text{ and } \boldsymbol{b} \text{ are in } S; \\ 0 & \text{otherwise.} \end{cases}$$

Since $\mathbf{Pr}[\boldsymbol{a}, \boldsymbol{b} \in S] = \mathbf{Pr}[\boldsymbol{a} \in S | \boldsymbol{b} \in S]\,\mathbf{Pr}[\boldsymbol{b} \in S] = \frac{t(t-1)}{(n-1)(n-2)}$, we have $\mathbf{E}[Z] = \mu$. Note that we can evaluate Z in $O(dt)$ time due to Lemma 3 as follows

$$Z = \frac{\sum_{\boldsymbol{a},\boldsymbol{b} \in S} \cos \Theta_{apb}}{t(t-1)} = \frac{1}{t(t-1)} \left\| \sum_{\boldsymbol{a} \in S} \frac{\boldsymbol{p} - \boldsymbol{a}}{\|\boldsymbol{p} - \boldsymbol{a}\|} \right\|^2 - \frac{1}{t-1}.$$

Hence, we can estimate

$$L1D(\boldsymbol{p}) \approx 1 - \sqrt{\frac{1}{n-1} + \frac{n-2}{n-1} \left(\frac{1}{t(t-1)} \left\| \sum_{\boldsymbol{a} \in S} \frac{\boldsymbol{p} - \boldsymbol{a}}{\|\boldsymbol{p} - \boldsymbol{a}\|} \right\|^2 - \frac{1}{t-1} \right)}.$$

Theoretical Analysis: For notational simplicity, let σ^2 be the variance of cosine of angles. Hence we have $\frac{\sum_{\boldsymbol{a},\boldsymbol{b}} \cos^2 \Theta_{apb}}{(n-1)(n-2)} = \sigma^2 + \mu^2$. For simplicity, we also assume that $\cos \Theta_{apb} \geq 0$ for *any* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{p} \in \mathbb{S}$. The following theorem shows the upper bound of variance of estimator provided by SamDepth.

Theorem 4.

$$\boldsymbol{Var}[Z] \leq \left(\sigma^2 + \mu^2\right) \left(\frac{1}{t-1} - \frac{1}{n-2} \right)$$

Proof. Due to limited space, we just sketch the proof. In order to bound $\mathbf{Var}[Z] = \mathbf{E}[Z^2] - \mu^2$ where $Z = \sum_{\boldsymbol{a},\boldsymbol{b}} Z_{ab}$, we decompose it into three terms corresponding to $\sum_{\boldsymbol{a},\boldsymbol{b}} \cos^2 \Theta_{apb}$, $\sum_{\boldsymbol{a} \neq \boldsymbol{a}', \boldsymbol{b} \neq \boldsymbol{b}'} \cos \Theta_{apb} \cos \Theta_{a'pb'}$, and $\sum_{\boldsymbol{a},\boldsymbol{b} \neq \boldsymbol{b}'} \cos \Theta_{apb} \cos \Theta_{apb'}$. Since $\sum_{\boldsymbol{a},\boldsymbol{b} \neq \boldsymbol{b}'} \cos \Theta_{apb} \cos \Theta_{apb'} \leq (n-3) \sum_{\boldsymbol{a},\boldsymbol{b}} \cos^2 \Theta_{apb}$ by Cauchy–Schwarz inequality and the contribution of the second term is negative, we can bound $\mathbf{Var}[Z]$ using only the first term, which leads to the result. □

Discussion: It is worth noting that SamDepth with $t = n - 1$ provides an exact $L1D(\boldsymbol{p})$ while the basic sampling method can only give an estimate. When t is large, SamDepth gives sufficiently small variance of estimator, and hence results in negligible loss on outlier detection using L1D. For the general case when any $\cos \Theta_{apb}$ might be negative, we will consider random variables $Y_{ab} = \frac{1}{t(t-1)} \frac{1 + \cos \Theta_{apb}}{2}$ instead. Applying Theorem 4, we can also bound the variance of estimator provided by SamDepth. Due to limited space, we leave the detail of the proof in the supplementary material[1].

Parameter Setting and Reproducibility: In order to make SamDepth completely parameter-free and efficient, we simply set $t = \sqrt{n}$. Hence SamDepth computes an unbiased estimate of $L1D(\boldsymbol{p})$ in $O(d\sqrt{n})$ time, which is significantly faster than $O(dn)$ time required by standard proximity-based outlier detectors. For reproducibility, we have released a C++ source code of SamDepth[2]. Our

[1] https://www.dropbox.com/s/yzbam4heruglj4i/Supplementary.pdf.
[2] https://github.com/NinhPham/Outlier.

empirical experiments on 14 real-world high-dimensional data sets demonstrate that SamDepth with $\sqrt{n}$ samples often achieves very competitive accuracy and runs several orders of magnitude faster than other proximity-based and ABOD competitors.

5 Experiment

We implemented SamDepth and other competitors in C++ and conducted experiments on a 3.40 GHz core i7 Windows platform with 32 GB of RAM. We compared the performance of SamDepth with ABOD detectors using L1D, VOA and ABOF and other proximity-based detectors on real-world high-dimensional data sets. We used the area under the ROC curve (AUC) to evaluate the accuracy of our *unsupervised* outlier detection methods since they deliver outlier rankings. For measuring efficiency, we computed the total running time in seconds for each detector. All results are over 5 runs of the algorithms.

Table 1. Data set properties: short names, number of points n, dimensionality d, and number of outliers o.

	Mam	Shuttle	Cover	Cardio	KDD	Spam	Opt	Mnist	Musk	Arr	Speech	Isolet	Mfeat	Ads
n	11183	49097	286048	2126	60839	4207	5216	7603	3062	452	3686	945	440	1966
d	6	9	10	21	41	57	64	100	166	274	400	617	649	1555
o	260	3511	2747	471	246	1679	150	700	97	66	60	45	40	368

5.1 Experiment Setup

Due to the cubic time complexity of VOA and ABOF, we will compute these exact values in some small data sets for comparison. For large-scale data sets, we used FastVOA [17] and FastABOF [14] to approximate VOA and ABOF, respectively. Below is the list of all implemented algorithms used in our experiment.

- **L1D**: L1D (exact), BasicSam (basic sampling), SamDepth.
- **VOA**: VOA (exact), FastVOA [17].
- **ABOF**: ABOF (exact), FastABOF ($k = \lceil 0.1 \cdot n \rceil$) [14].
- **Proximity-based factors**: kNN [19], kNNW [4] with fixed parameter $k = 10$ and LOF [5] with fixed parameter $k = 40$.

For FastVOA, we used 100 random projections and fixed the size of AMS Sketches $s_1 = 3200, s_2 = 5$ in all experiments. We used $k = \lceil 0.1 \cdot n \rceil$ for FastABOF as suggested in [14] since if k is small, approxABOF is simply the local outlier factor of ABOF and cannot reflect well the original ABOF idea. For consistency, we used $k = 10$ for kNN and kNNW as used in [19] and [4]. For LOF, we used $k = 40$ as suggested in [5]. We note that finding the best parameter k for proximity-based methods is an extremely time-consuming task, especially for large-scale data sets. The brute force procedure to find the best parameter k requires $O(\kappa n \log n)$ time for evaluating κ AUC scores where κ is

the largest possible value of k. With the KDDCup99 data set of size $n = 60839$, this process will need approximately 100 h to finish for $\kappa = 1000$ on our machine.

It is worth noting that in the realistic cases where we do not actually know the outlier labels, incorrect settings in parameter-laden measures may cause the algorithms to fail in finding the true anomaly patterns. Parameter-free outlier factors including L1D, VOA and ABOF would limit our ability to impose our prejudices or presumptions on the problem, and "let the data speak for themselves".

5.2 Data Sets

We conducted experiments on real-world high-dimensional data sets, including widely used data sets in literature and semantically meaningful data sets with interpretation for outliers from popular resources, as shown in Table 1.

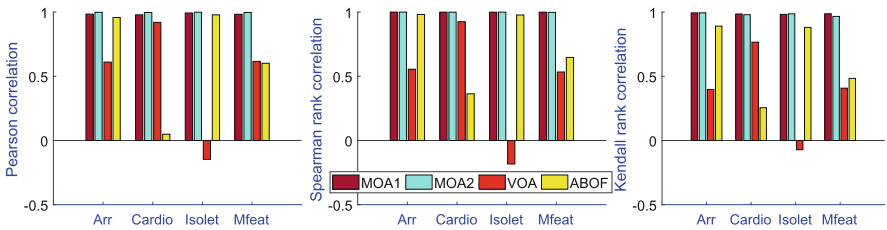

Fig. 3. Correlation coefficients between L1D and MOA1, MOA2, VOA and ABOF on 4 data sets: Arr, Cardio, Isolet and Mfeat.

- [20][3]: Shuttle, Optdigits (Opt for short), Mnist, Musk, Arrhythmia (Arr for short), Speech, Mammography (Mam for short), ForestCover (Cover for short).
- [6][4]: Cardiotocography (Cardio for short, 22% of outliers, not normalized, duplicates), KDDCup99 (KDD for short, normalized, duplicates, idf weighted), SpamBase (Spam for short, 40% of outliers, not normalized, without duplicates), InternetAds (Ads for short, 19%, normalized, w.o. duplicates)
- [15][5]: Isolet (classes C, D, E as inliers and random points in class Y as outliers) and Multiple Features (Mfeat for short, classes 6 and 9 as inliers and random points in class 0 as outliers).

5.3 Relationship Between L1D and ABOD Measures

This subsection conducted experiments on evaluating the correlation between the proposed outlier factor L1D with other ABOD measures. We computed the exact values of L1D, MOA1, MOA2, VOA and ABOF and evaluated the statistical relationships using correlation coefficients, including Pearson correlation

[3] http://odds.cs.stonybrook.edu/.
[4] http://www.dbs.ifi.lmu.de/research/outlier-evaluation/DAMI/.
[5] https://archive.ics.uci.edu/ml/datasets.html.

coefficients, Spearman's rank correlation coefficients and Kendall's rank correlation coefficients. We computed these coefficients over 4 small data sets, including Arr, Cardio, Isolet and Mfeat, and demonstrates the results in Fig. 3.

It is clear that the outlier rankings based on L1D are almost identical to that of MOA1 and MOA2. L1D is also highly correlated to ABOF, except the low-dimensional data set Cardio. This is due to the fact that the distance's effect is significant in ABOF for low-dimensional data.

The L1D's rankings are also highly correlated to the VOA's ones, except the Isolet data set. In fact, on Isolet, the inlier classes C, D, and E have very similar MOA1 values, whereas the average MOA1 of the outlier class Y significantly deviates from the rest. Considering variance as a central second moment, we know that the VOA's ranking of outliers will change significantly compared to the MOA1's ranking, which leads to the situation where there is a negative correlation between L1D (or MOA1, MOA2) and VOA. Besides, on Isolet, the L1D's ranking is also highly correlated to the ABOF's one. This is due to the effect of distances in ABOF measure, which can be observed in the next experiment where the kNN detector shows the best performance on Isolet.

We also note that since L1D's ranking is almost identical to MOA1, we can use MOA1 as an ABOD outlier factor. The near-linear time approximation algorithm to estimate MOA1 for all points proposed in [17] can be used to speed up the outlier detection process.

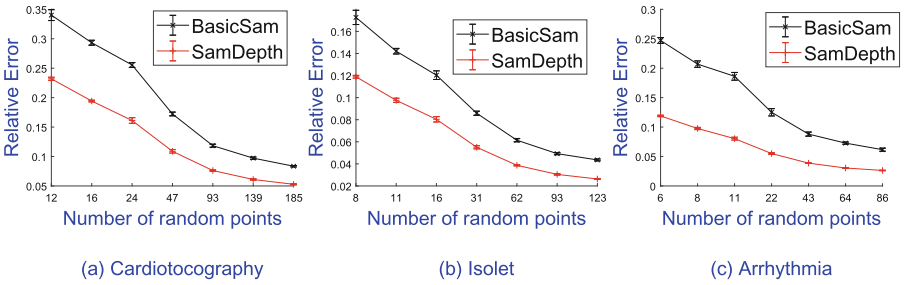

Fig. 4. Relative approximation errors provided by SamDepth and BasicSam on the Cardio, Isolet and Arr data sets when increasing the sample size.

5.4 Relative Approximation Errors

This subsection presents experiments to measure *relative* approximation errors of SamDepth and BasicSam on previous data sets, including Cardio, Isolet and Arr. For the sake of comparison, we used t random *points* for SamDepth and $\lceil t/2 \rceil$ random *pairs* for BasicSam due to the fact that BasicSam needs two random points for each random pair.

Figure 4 displays the *average* relative approximation errors provided by SamDepth and BasicSam when varying the number of sample *points* t in range $\lceil \frac{\sqrt{n}}{i} \rceil$ where $i = \{\frac{1}{4}, \frac{1}{3}, \cdots, 3, 4\}$. It is clear that the average relative errors and its variances of both sampling methods decrease dramatically when increasing the

sample size. Particularly, SamDepth with $\sqrt{n}$ samples provides average relative errors less than $\epsilon = 0.1$ on the three data sets. Since the errors of SamDepth are significantly smaller than that of BasicSam, SamDepth will achieve higher accuracy than BasicSam on detecting outliers using L1-depth.

5.5 Outlier Detection Performance

In this subsection, we compare the outlier detection performance of sampling methods using the AUC value (i.e. the area under the ROC curve). The AUC value for an ideal outlier ranking is 1 when all outliers are top-ranked points. The AUC value of a "less than perfect" outlier detection algorithm is typically less than 1.

We again used 3 data sets, Arr and Isolet as high-dimensional data sets and Cardio as a low-dimensional data set to measure AUC values of L1D measure provided by the exact and sampling solutions. We studied the performance of sampling methods where we varied the sample size as described in the previous subsections. Figure 5 reveals the AUC values of BasicSam and SamDepth compared to the exact solution (L1D) on the data sets.

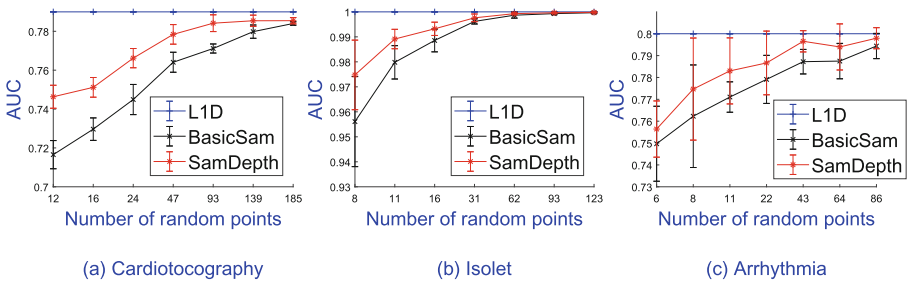

Fig. 5. Comparison of AUC values of SamDepth, BasicSam and L1D when varying number of samples on 3 data sets.

It is clear that the AUC values of sampling methods converge to the AUC value of the exact solution and the deviations of sampling methods are significantly reduced when increasing the sample size. SamDepth provides superior performance compared to BasicSam regarding both AUC values and its deviation on the 3 data sets. Since SamDepth outperforms BasicSam on detecting outliers, we will use SamDepth with L1D to compare the outlier detection performance with other proximity-based and ABOD-based outlier factors.

In order to quantify ABOD measures, we compare L1D, VOA, and ABOF with proximity-based measures, including kNN, kNNW and LOF. For L1D, we study the exact method and SamDepth with sample size $\sqrt{n}$. Since VOA and ABOF require $O(n^3)$ time for exact values, we computed exact values for small data sets and used approximation methods, including FastVOA and FastABOF, for large data sets. Table 2 shows the AUC values and Table 3 depicts the running

time in seconds on all used data set, except the 4 large-scale data sets Mam, Shuttle, Cover and KDD.

Table 2. Comparison of AUC values of several outlier detectors on high-dimensional data sets. The top-2 AUCs are in boldface on each data set.

	kNN	kNNW	LOF	L1D	SamDepth	VOA	FastVOA	ABOF	FastABOF
Cardio	0.62	0.61	0.63	**0.79**	**0.78**	**0.78**	0.77	0.57	0.55
SpamBase	**0.71**	**0.68**	0.51	0.49	0.49	–	0.44	–	0.54
OptDigits	0.41	0.40	0.54	**0.56**	0.55	–	**0.62**	–	0.47
Mnist	0.82	0.80	0.72	**0.84**	0.82	–	0.57	–	**0.86**
Musk	0.64	0.24	0.41	**0.91**	**0.89**	0.79	0.79	0.1	0.06
Arrhythmia	**0.81**	**0.80**	**0.81**	**0.80**	0.79	0.68	0.56	**0.81**	0.79
Speech	0.48	**0.52**	0.49	0.47	0.47	0.40	0.50	0.47	**0.51**
Isolet	**0.96**	0.94	0.26	**1**	**1**	0.44	0.32	**1**	0.83
Mfeat	0.41	0.41	0.37	**0.95**	**0.92**	0.90	0.79	0.49	0.45
InternetAds	**0.70**	**0.72**	0.67	0.69	0.68	0.44	0.57	0.68	0.69
Avg AUC	0.66	0.61	0.54	**0.75**	**0.74**	–	0.59	–	0.58

Table 3. Comparison of running time (in seconds) of several outlier detectors on high-dimensional data sets. The smallest running time values are in boldface on each data set.

	kNN	kNNW	LOF	L1D	SamDepth	FastVOA	FastABOF
Cardio	0.3	0.3	0.4	0.4	**0.03**	17.6	3.1
SpamBase	1.6	1.6	2.1	3.4	**0.2**	36.1	47.3
OptDigits	2.9	2.7	3.6	5.3	**0.2**	45.6	102.2
Mnist	8.2	8.3	10.0	17.4	**0.6**	66.0	460.0
Musk	2.1	2.1	2.4	4.7	**0.2**	26.9	51.2
Arrhythmia	0.06	0.08	0.09	0.16	**0.02**	3.6	0.3
Speech	6.8	6.8	7.0	15.9	**0.8**	31.5	196.1
Isolet	0.6	0.6	0.7	1.6	**0.1**	8.2	5.8
Mfeat	0.14	0.14	0.17	0.34	**0.05**	3.6	0.7
InternetAds	6.9	7.1	7.3	17.0	**1.0**	16.7	125.6
Avg Time	3.0	3.0	3.4	6.6	**0.3**	25.6	99.2

In general, L1D provides superior performance compared to other outlier factors regarding the AUC with the highest average value of 75%. In particular, L1D significantly outperforms VOA and ABOF for all 7 small data sets regarding both accuracy and efficiency. Its AUC values are in the top-2 of 7 over 10 used data sets. While AUC scores of SamDepth, kNN and kNNW are in the top-2 of 4 data sets, FastVOA and FastABOF show slightly less detection performance due to the approximation errors. Among the proximity-based factors, kNN shows the

Table 4. Comparison of AUC values and running time in seconds of representative outlier detectors on 4 large-scale data sets.

Methods	AUC				Time (s)			
	Mam	Shuttle	Cover	KDDCup99	Mam	Shuttle	Cover	KDDCup99
kNN	**0.85**	0.76	**0.85**	0.85	8	138	4837	387
SamDepth	0.84	**0.99**	**0.85**	**0.99**	**0.2**	**3**	**265**	**10**
FastVOA	0.79	0.71	0.74	**0.99**	123	625	15413	1637
FastABOF	0.62	0.66	0.81	0.57	30	330	12956	3557

superior performance and LOF shows the inferior performance in average. Hence we used KNN as a representative algorithm for the proximity-based methods on large-scale experiments.

Regarding both effectiveness and efficiency, SamDepth illustrates substantial advantages with the second highest average AUC 74% but runs up to several orders of magnitude faster than other methods. In average, SamDepth runs approximately 10× faster than proximity-based methods, 22× faster than exact L1D, 85× faster than FastVOA, and 330× faster than FastABOF.

We conclude the empirical evaluation by depicting the performance of detectors on 4 large-scale data sets, including Mam, Shuttle, Cover and KDD. For each type of outlier factors, we used its representative algorithm, including kNN, SamDepth, FastVOA and FastABOF. Since the data set's size is very large, FastABOF with $k = \lceil 0.1 \cdot n \rceil$ would not finish after 10 h. Hence we set $k = \lceil \sqrt{n} \rceil$ for FastABOF. Table 4 reveals the AUC values and running time in seconds on 4 large-scale data sets. Again, SamDepth provides superior performance compared to the other methods. It almost obtains the highest AUC values and runs several orders of magnitude faster than other competitors.

6 Conclusions

The paper investigates the *parameter-free* angle-based outlier detection (ABOD) in high-dimensional data. Exploiting the conceptual relationship between the ABOD intuition and the L1-depth notion (L1D), we propose to use L1D as a robust variant of ABOD measures, which only requires a quadratic computational time. Empirical experiments on many real-world high-dimensional data sets show that L1D is superior to other ABOD measures, such as ABOF and VOA, and very competitive to other proximity-based measures, including kNN, kNNW and LOF on detecting high-dimensional outliers regarding ROC AUC scores.

In order to avoid the high computational complexity of L1D measures, we propose SamDepth, a simple but efficient sampling algorithm which often runs faster and achieves very comparable outlier detection performance compared to the exact method. Especially, SamDepth with $\sqrt{n}$ samples shows the superior performance compared to widely used detectors regarding both effectiveness and efficiency on many real-world high-dimensional data sets.

Acknowledgments. We would like to thank Rasmus Pagh for useful discussion and comments in the early stage of this work. We thank members of the DABAI project and anonymous reviewers for their constructive comments and suggestions.

References

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, pp. 420–434. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44503-X_27
2. Aggarwal, C.C., Sathe, S.: Outlier Ensembles: An Introduction. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54765-7
3. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. In: Proceedings of SIGMOD 2001, pp. 37–46 (2001)
4. Angiulli, F., Pizzuti, C.: Outlier mining in large high-dimensional data sets. IEEE Trans. Knowl. Data Eng. **17**(2), 203–215 (2005)
5. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. In: Proceedings of SIGMOD 2000, pp. 93–104 (2000)
6. Campos, G.O., et al.: On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Min. Knowl. Discov. **30**(4), 891–927 (2016)
7. Chen, Y., Bart Jr., H.L., Dang, X., Peng, H.: Depth-based novelty detection and its application to taxonomic research. In: Proceedings of ICDM 2007, pp. 113–122 (2007)
8. Hawkins, D.: Identification of Outliers. Chapman and Hall, London (1980)
9. Hugg, J., Rafalin, E., Seyboth, K., Souvaine, K.: An experimental study of old and new depth measures. In: Proceedings of ALENEX 2006, pp. 51–64 (2006)
10. Jeong, M., Cai, Y., Sullivan, C.J., Wang, S.: Data depth based clustering analysis. In: Proceedings of SIGSPATIAL 2016, pp. 29:1–29:10 (2016)
11. Johnson, T., Kwok, I., Ng, R.T.: Fast computation of 2-dimensional depth contours. In: Proceedings of KDD 1998, pp. 224–228 (1998)
12. Jörnsten, R.: Clustering and classification based on the L1 data depth. J. Multivar. Anal. **90**(1), 67–89 (2004)
13. Knorr, E.M., Ng, R.T.: Algorithms for mining distance-based outliers in large datasets. In: Proceedings of VLDB 1998, pp. 392–403 (1998)
14. Kriegel, H.-P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In: Proceedings of KDD 2008, pp. 444–452 (2008)
15. Lichman, M.: UCI machine learning repository (2013)
16. Papadimitriou, S., Kitagawa, H., Gibbons, P.B., Faloutsos, C.: LOCI: fast outlier detection using the local correlation integral. In: Proceedings of ICDE 2003, pp. 315–326 (2003)
17. Pham, N., Pagh, R.: A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data. In: Proceedings of KDD 2012, pp. 877–885 (2012)
18. Preparata, F.P., Shamos, M.: Computational Geometry: An Introduction. Springer, New York (1985). https://doi.org/10.1007/978-1-4612-1098-6
19. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. In: Proceedings of SIGMOD 2000, pp. 427–438 (2000)
20. Rayana, S.: ODDS library (2016)

21. Serfling, R.: A depth function and a scale curve based on spatial quantiles. In: Dodge, Y. (ed.) Statistical Data Analysis Based on the L_1-Norm and Related Methods. SIT, pp. 25–38. Birkhäuser Basel, Basel (2002). https://doi.org/10.1007/978-3-0348-8201-9_3
22. Tukey, J.W.: Mathematics and picturing data. In: Proceedings of the International Congress of Mathematicians Vancouver, pp. 523–531 (1974)
23. Vardi, Y., Zhang, C.-H.: The multivariate L1-median and associated data depth. Proc. Natl. Acad. Sci. U. S. A. **97**(4), 1423–1426 (2000)
24. Zimek, A., Campello, R.J.G.B., Sander, J.: Ensembles for unsupervised outlier detection: challenges and research questions a position paper. SIGKDD Explor. **15**(1), 11–22 (2013)