# Dimensionality-Aware Outlier Detection*

Alastair Anderberg † James Bailey ‡ Ricardo J. G. B. Campello §
Michael E. Houle ¶‡ Henrique O. Marques § Miloš Radovanović ‖ Arthur Zimek §

**Abstract**

We present a nonparametric method for outlier detection that takes full account of local variations in intrinsic dimensionality within the dataset. Using the theory of Local Intrinsic Dimensionality (LID), our 'dimensionality-aware' outlier detection method, DAO, is derived as an estimator of an asymptotic local expected density ratio involving the query point and a close neighbor drawn at random. The dimensionality-aware behavior of DAO is due to its use of local estimation of LID values in a theoretically-justified way. Through comprehensive experimentation on more than 800 synthetic and real datasets, we show that DAO significantly outperforms three popular and important benchmark outlier detection methods: Local Outlier Factor (LOF), Simplified LOF, and $k$NN.

**Keywords:** outlier detection, intrinsic dimensionality

## 1 Introduction

Outlier detection, one of the most fundamental tasks in data mining, aims to identify observations that deviate from the general distribution of the data. Such observations often deserve special attention as they may reveal phenomena of extreme importance, such as network intrusions [1], sensor failures [39], or disease [2].

The study of outliers has its origins in the field of statistics. There exist dozens of parametric statistical tests that can be applied to detect outliers [9, 18]. Although these tests have shown good performance when the underlying theoretical assumptions are met, in real-world applications these assumptions usually do not hold [35]. This well-known limitation of the parametric approach has triggered research on unsupervised, non-parametric methods for outlier detection, as far back as the seminal work of Knorr and Ng in 1997 [27]. Non-parametric approaches make no explicit assumptions on the nature of the underlying data distribution, but can estimate some of its local characteristics, such as probability density at a point of interest. Although nonparametric methods are usually more suitable for real-world applications due to their flexibility, generally speaking they lack the theoretical justification that parametric approaches have enjoyed [47].

The estimates computed by non-parametric methods for outlier detection usually rely on the distances from the test point to its nearest neighbors. As such, these methods are subject to the well-known 'curse of dimensionality' phenomenon, by which the quality of distance information diminishes as the dimensionality of the data increases [48], leading to such observable effects as the concentration of distance values about their mean [10]. Contrary to what is commonly assumed, however, most of the challenges associated with high-dimensional data analysis do not depend directly on the representational data dimension (number of attributes); rather, they are better explained by the notion of 'intrinsic dimensionality' (ID), which can be understood intuitively as the number of features required to explain the distributional characteristics observed within the data, or the dimension of the surface (or manifold or subspace) achieving the best fit to the data. In practice, for many models of ID, the estimated number of explanatory features or surface dimensions is often much smaller than the dimension of the embedding data space.

Typically, the intrinsic dimension is not uniform across the whole dataset: applying a model of ID to subregions of the data (such as the neighborhood of a query point) almost always produces different results for each. The added complexity associated with variation of local ID has the potential to increase the difficulty of data analysis, thereby resulting in a performance loss for many similarity-based data mining and indexing methods [3, 7, 21, 26, 48]. For this reason, there has been considerable recent attention to 'local' models of ID as an alternative to the classic 'global' models that seek to characterize the complexity of the entire dataset.

*The full version of the paper can be accessed at `https://arxiv.org/abs/2401.05453`

†The University of Newcastle, Callaghan, NSW, Australia. `anderberg.alastair@gmail.com`

‡The University of Melbourne, Parkville, VIC, Australia. `baileyj@unimelb.edu.au`

§University of Southern Denmark, Odense, Denmark. `{campello, oli, zimek}@imada.sdu.dk`

¶New Jersey Institute of Technology, Newark, NJ, USA. `michael.houle@njit.edu`

‖University of Novi Sad, Serbia. `radacha@dmi.uns.ac.rs`

In this paper, we focus on a theoretical model of local complexity, the Local Intrinsic Dimensionality (LID) [19, 20], which was originally motivated by the need to characterize the interrelationships between probability and distance within neighborhoods. Unlike other measures of ID which are formulated as heuristics for direct use on discrete datasets, LID is a theoretical quantity for which effective estimators have been developed [4, 5, 32]. LID has had many recent successes in data analysis, both theoretical and practical, in areas such as search and indexing [7, 13], AI and data mining [8, 41], and deep learning [3, 33]. There is empirical evidence to suggest that outlierness is correlated in practice with high local intrinsic dimensionality [22].

The main contribution of our paper is the first known nonparametric method for *dimensionality-aware outlier detection* (DAO), one whose formulation we derive as an estimator of an *asymptotic local expected density ratio* (ALDR), using the theory of LID. The dimensionality-aware behavior of DAO is due to its use of local estimation of LID values in a theoretically-justified way. Our proposed model will be seen to resemble the classic LOF outlier detection criterion [11], and (even more closely) its very popular simplified variant, SLOF [43]. However, like all known nonparametric outlier detection criteria, LOF and SLOF both differ from our proposed model in that they rely solely on distance-based criteria for density estimation, without taking local dimensionality explicitly into account.

Through our theoretical model we gain an understanding of the susceptibility of traditional outlier detection methods to variation in local ID within the dataset. As a second main contribution, we verify this understanding through a comprehensive empirical study (involving a total of more than 800 datasets) of the performance of DAO versus three of the most popular and effective nonparametric outlier detection methods known to date: LOF, SLOF, and $k$NN [38]. In particular, we present visualizations of outlier detection performance for 393 real datasets, that empirically confirm the tendency of DAO to outperform its dimensionality-unaware competitors, particularly when the variation of LID values within a dataset is high (as indicated by measures of high dispersion or low autocorrelation).

The remainder of this paper is organized as follows. In Section 2 we discuss related work. In Section 3 we provide background on the LOF and SLOF outlier detection methods, and the theory of local intrinsic dimensionality. In Section 4 we derive and theoretically justify our proposed dimensionality-aware outlierness model, DAO. We present our experimental setup in Section 5, and discuss the results in Section 6. Finally, we present concluding remarks in Section 7.

## 2 Related Work

Non-parametric approaches for outlier detection [12] either explicitly or implicitly aim to assess the density in the vicinity of a query point, such that points with the lowest densities are reported as the strongest outlier candidates. The assessment of density can be direct, or based on distances, or on the ratio of one density with respect to another. The distance-based DB-outlier method [27] estimates density in the vicinity of a query by counting the number of data points contained in a neighborhood with predefined radius. Conversely, the $k$-nearest-neighbor algorithm ($k$NN) [38] measures the radius needed so as to capture a fixed number of points, $k$. Local outlier detection methods based on density ratios, such as LOF [11], identify outliers to be those points having local densities that are small relative to those of their nearest neighbors.

Many variations of the aforementioned outlier models have been proposed over the past decades. Some rely on nonstandard notions of neighborhood, such as COF (connectivity-based outlier factor) [44], INFLO (Influenced Outlierness) [24], and others based on reverse nearest neighbors [37]. Others estimate the local density in different ways, such as LDF (Local Density Factor) [31], LOCI (Local Outlier Integral) [36], and KDEOS (Kernel Density Estimation Outlier Score) [42]. Yet other variations derive an outlier score from a comparison of a local model for the query point to local models of other data points; these include the local distance-based outlier detection (LDOF) approach [46], probabilistic modeling of local outlier scores (LoOP) [28], or meta-modeling of outlierness [29]. A somewhat different approach is angle-based outlier detection (ABOD) [30], which bases the degree of outlierness of a query point on the variance of the angles formed by it and other pairs of points.

Despite the many variants of the fundamental techniques of non-parametric outlier detection that have appeared over the past quarter century, two classic methods in particular, LOF and $k$NN, have repeatedly been confirmed as top performers or recommended baselines in larger comparative studies involving local anomaly detection [12, 16, 17]. None of these methods, however, take into account the possibility of variation in local intrinsic dimensionality within the dataset.

## 3 Background

**3.1 Local Outlier Factor.** The term 'local outlier' refers to an observation that is sufficiently different from observations in its vicinity. Typical density-based outlier detection methods consider a point $\mathbf{q}$ as a local outlier if a given neighborhood of $\mathbf{q}$ is less dense than neighborhoods centered at $\mathbf{q}$'s own neighbors, according

to some criterion. Following this principle, Local Outlier Factor (LOF) [11] contrasts the local density at $\mathbf{q}$ with the local densities at the members of its $k$-nearest neighbor set, $\mathrm{NN}_k(\mathbf{q})$:

$$\mathrm{LOF}_k(\mathbf{q}) \triangleq \frac{1}{k} \sum_{\mathbf{o} \in \mathrm{NN}_k(\mathbf{q})} \frac{\mathrm{lrd}_k(\mathbf{o})}{\mathrm{lrd}_k(\mathbf{q})},$$

where the local reachability density (lrd) at point $\mathbf{p}$ is defined in terms of the inverse of an average of so-called 'reachability distances' taken from the $k$-nearest neighbors of $\mathbf{p}$:

$$\mathrm{lrd}_k(\mathbf{p}) \triangleq \left( \frac{\sum_{\mathbf{s} \in \mathrm{NN}_k(\mathbf{p})} \mathrm{reach\_dist}_k(\mathbf{p} \leftarrow \mathbf{s})}{k} \right)^{-1}.$$

Such a distance is defined as the maximum of the neighbor's own $k$-NN distance, $k\_\mathrm{dist}(\mathbf{s})$, and its distance to $\mathbf{p}$, $d(\mathbf{p}, \mathbf{s})$:

$$\mathrm{reach\_dist}_k(\mathbf{p} \leftarrow \mathbf{s}) = \max\{k\_\mathrm{dist}(\mathbf{s}), d(\mathbf{p}, \mathbf{s})\}.$$

The LOF reachability distance can be regarded as using the distance between $\mathbf{p}$ and its neighbor $\mathbf{s}$ by default, except when $\mathbf{s}$ is closer to $\mathbf{p}$ than it is to its own $k$-th nearest neighbor.

Although the local reachability density of LOF aggregates the contribution of many neighbors to produce a smoother and more stable estimate, it requires multiple levels of neighborhood computation (for each neighbor $\mathbf{o}$ of $\mathbf{q}$, the $k$-NN distance of each of the neighbors of $\mathbf{o}$). The Simplified LOF (SLOF) variant [43] avoids one level of neighborhood computation by using the inverse $k$-NN distance in place of the local reachability density:

$$\mathrm{SLOF}_k(\mathbf{q}) \triangleq \frac{1}{k} \sum_{\mathbf{o} \in \mathrm{NN}_k(\mathbf{q})} \frac{\mathrm{slrd}_k(\mathbf{o})}{\mathrm{slrd}_k(\mathbf{q})},$$

where

$$\mathrm{slrd}_k(\mathbf{p}) \triangleq \frac{1}{k\_\mathrm{dist}(\mathbf{p})}.$$

The density of the neighborhood $\mathrm{NN}_k(\mathbf{q})$ can be regarded as the ratio between the mass (the number of points $k$) and the volume of the ball with radius $k\_\mathrm{dist}(\mathbf{q})$. In the Euclidean setting, this ratio is proportional to $k/(k\_\mathrm{dist}(\mathbf{q}))^m$, where $m$ is the dimension of the space. $\mathrm{slrd}_k(\mathbf{q})$ can thus be interpreted as a proportional density estimate that treats $k$ as a constant, and ignores the dimension of the ambient space, $m$.

**3.2 Local Intrinsic Dimensionality.** The Local Intrinsic Dimensionality (LID) model [20] can be regarded as a continuous extension of the expansion dimension due to Karger and Ruhl [26], which derives a measure of dimensionality from the relationship between volume and radius in an expanding ball centered at a point of interest in a Euclidean data domain. Given two measurements of radii ($r_1$ and $r_2$) and volume ($V_1$ and $V_2$), the dimension $m$ can be obtained from the ratios of the measurements:

$$\frac{V_2}{V_1} = \left(\frac{r_2}{r_1}\right)^m \implies m = \frac{\ln(V_2/V_1)}{\ln(r_2/r_1)}.$$

Early expansion models are discrete, in that they estimate volume by the number of data points captured by the ball. The LID model, by contrast, allows data to be viewed as samples drawn from an underlying distribution, with the volume of a ball represented by the probability measure associated with its interior. For balls centered at a common reference point, the probability measure can be expressed as a function $F(r)$ of the radius $r$, and as such $F$ can be viewed as the cumulative distribution function (CDF) of the distribution of distances to samples drawn from the underlying global distribution. However, it should be noted that the LID model has been developed to characterize the complexity of growth functions in general: the variable $r$ need not be a Euclidean distance, and the function $F$ need not satisfy the conditions of a CDF.

DEFINITION 3.1. ([20]) *Let $F$ be a real-valued function that is non-zero over some open interval containing $r \in \mathbb{R}$, $r \neq 0$. The* intrinsic dimensionality of $F$ at $r$ *is defined as follows, whenever the limit exists:*

$$\mathrm{IntrDim}_F(r) \triangleq \lim_{\epsilon \to 0} \frac{\ln\left(F((1+\epsilon)r)/F(r)\right)}{\ln(1+\epsilon)}.$$

When $F$ is 'smooth' (continuously differentiable) in the vicinity of $r$, its intrinsic dimensionality has a closed-form expression:

THEOREM 3.1. ([20]) *Let $F$ be a real-valued function that is non-zero over some open interval containing $r \in \mathbb{R}$, $r \neq 0$. If $F$ is continuously differentiable at $r$, then*

$$\mathrm{ID}_F(r) \triangleq \frac{r \cdot F'(r)}{F(r)} = \mathrm{IntrDim}_F(r).$$

In characterizing the local intrinsic dimensionality at a query location, we are interested in the limit of $\mathrm{ID}_F(r)$ as the distance $r$ tends to 0, which we denote by

$$\mathrm{ID}_F^* \triangleq \lim_{r \to 0} \mathrm{ID}_F(r).$$

Henceforth, when we refer to the local intrinsic dimensionality of a function $F$, or of a reference location whose induced distance distribution has $F$ as its CDF, we will take 'LID' to mean the quantity $\mathrm{ID}_F^*$.

## 4 The Dimensionality-Aware Outlier Model

As discussed previously, local outliers can in general be found by comparing the density of the neighborhood of a point to the densities of the neighborhoods of that point's neighbors. Here, we emulate the design choices of traditional (discrete) density-based outlier models using distributional concepts, to produce a theoretical dimensionality-aware model that treats the dataset as samples drawn from some unknown underlying distribution that is assumed to be continuous everywhere except (perhaps) at the query location. After establishing our model, we develop a practical estimator of outlierness suitable for use on discrete datasets.

**4.1 Asymptotic Local Density Ratio.** For any distribution over an isometric representation space, the volume $V(\epsilon)$ of any ball of radius $\epsilon$ is the same throughout the domain. For a suitably small choice of $\epsilon$, we consider the density at a point $\mathbf{p}$ to be the probability measure $F_{\mathbf{p}}(\epsilon)$ associated with its $\epsilon$-neighborhood ball $B_{\mathbf{p}}(\epsilon)$, divided by the volume of the ball, $V(\epsilon)$. Note that in any such density ratio between a test point $\mathbf{q}$ and its neighbor $\mathbf{o}$, the volumes cancel out to produce the simple ratio ${}^{F_{\mathbf{o}}(\epsilon)}/_{F_{\mathbf{q}}(\epsilon)}$. Rather than aggregating density ratios for a fixed number of discrete neighbors, we instead reason in terms of the expectation of density ratios involving a random sample $\mathbf{o}$ drawn from $B_{\mathbf{q}}(\epsilon)$:

$$\mathop{\mathbb{E}}_{\mathbf{o}\in B_{\mathbf{q}}(\epsilon)}\left[\frac{F_{\mathbf{o}}(\epsilon)}{F_{\mathbf{q}}(\epsilon)}\right].$$

In practice, models for outlier detection are faced with the problem of deciding the neighborhood radius $\epsilon$, or neighborhood cardinality $k$. Here, we resolve this issue by examining the tendency of our density-based criterion as the ball radius tends to zero, thereby obtaining a ratio of infinitesimals. Accordingly, we define the asymptotic local expected density ratio (ALDR) of a query point $\mathbf{q}$ to be:

$$\mathrm{ALDR}(\mathbf{q}) \triangleq \lim_{\epsilon\to 0^+}\mathop{\mathbb{E}}_{\mathbf{o}\in B_{\mathbf{q}}(\epsilon)}\left[\frac{F_{\mathbf{o}}(\epsilon)}{F_{\mathbf{q}}(\epsilon)}\right].$$

Intuitively, an ALDR score of 1 is associated with inlierness: it indicates that the probability measure function $F_{\mathbf{q}}$ in the vicinity of the test point $\mathbf{q}$ agrees perfectly with that of its neighbors, in that their (expected) local probability measures $F_{\mathbf{o}}$ converge to $F_{\mathbf{q}}$ as $\mathbf{o}$ tends to $\mathbf{q}$.

A limit value different than 1 indicates a discontinuity of the neighborhood probability measure at $\mathbf{q}$ relative to its neighbors. Limit values greater than 1 (including the case where ALDR diverges to infinity) are associated with outlierness in the usual sense of sparseness: they indicate that the test point has a local probability measure too small to be consistent with those of its neighbors in the domain. Limit values less than 1 can also be regarded as anomalous, in that they can be interpreted as an abnormally large concentration of probability measure at the individual point $\mathbf{q}$. In both cases, the degree of discontinuity can be regarded as increasing as the ratio diverges from 1. In this paper, however, we will be concerned with identifying cases for which the ALDR score exceeds 1 (sparse outlierness).

**4.2 Dimensionality-Aware Reformulation of ALDR.** For the purposes of deriving an estimator of ALDR, we introduce a minor reformulation. Instead of taking the radius of the ball $B_{\mathbf{q}}$ to be the same as the neighborhood radius within which probability measure is assessed around $\mathbf{q}$ and $\mathbf{o}$, we decouple the rates by which these radii tend to zero. In the reformulation, the inner limit controls the neighborhood radius, and the outer limit controls the ball radius.

$$\mathrm{ALDR}'(\mathbf{q}) \triangleq \lim_{\epsilon\to 0^+}\mathop{\mathbb{E}}_{\mathbf{o}\in B_{\mathbf{q}}(\epsilon)}\left[\lim_{\gamma\to 0^+}\frac{F_{\mathbf{o}}(\gamma)}{F_{\mathbf{q}}(\gamma)}\right].$$

With this decoupling, we show that we can convert the ratio of neighborhood probabilities ${}^{F_{\mathbf{o}}(\gamma)}/_{F_{\mathbf{o}}(\gamma)}$ to one that involves only distances and LID values. Given a probability value $p \in [0,1]$ and any point $\mathbf{o}$ in the domain, let $\delta_{\mathbf{o}}(p)$ be the infimum of the distance values $r$ for which $F_{\mathbf{o}}(r) = p$. This definition ensures that if $F_{\mathbf{o}}$ is continuously differentiable at $\delta_{\mathbf{o}}(p)$, then $F_{\mathbf{o}}(\delta_{\mathbf{o}}(p)) = p$, and the distance function $\delta_{\mathbf{o}}$ is also continuously differentiable at $p$.

THEOREM 4.1. *Let $\mathbf{q}$ be a query point. If there exists a constant $c > 0$ such that for all $\mathbf{o} \in B_{\mathbf{q}}(c) \setminus \{\mathbf{q}\}$,*

- *$F_{\mathbf{o}}$ is continuously differentiable over the range $[0, c]$,*
- *$\mathrm{ID}^*_{F_{\mathbf{o}}}$ exists and is positive, and*
- *the limit of ${}^{\delta_{\mathbf{q}}(p)}/_{\delta_{\mathbf{o}}(p)}$ as $p \to 0^+$ either exists or diverges to $+\infty$,*

*then*

$$\mathit{ALDR}'(\mathbf{q}) = \lim_{\epsilon\to 0^+}\mathop{\mathbb{E}}_{\mathbf{o}\in B_{\mathbf{q}}(\epsilon)}\left[\lim_{p\to 0^+}\left(\frac{\delta_{\mathbf{q}}(p)}{\delta_{\mathbf{o}}(p)}\right)^{\mathrm{ID}^*_{F_{\mathbf{o}}}}\right].$$

For the details of the proof, see the full version of the paper.

**4.3 The Dimensionality-Aware Outlierness Criterion.** We now make use of the formulation in the statement of Theorem 4.1 to produce a practical

estimate of ALDR′ on finite datasets. For this, we consider the value of ALDR′ for small (but positive) choices of the limit parameters $\epsilon$ and $p$.

Following the convention of LOF, SLOF, and other traditional outlier detection algorithms, the ball radius can be set to the familiar $k$-nearest neighbor distance, $\epsilon = k\_\mathrm{dist}(\mathbf{q})$. Similarly, if $n$ is the size of the dataset, choosing $p = k/n$ would set $\delta_{\mathbf{q}}(p)$ and $\delta_{\mathbf{o}}(p)$ to the distances at which their associated neighborhoods would be expected to contain $k$ samples out of $n$; these distances can be approximated by the $k$-NN distances $k\_\mathrm{dist}(\mathbf{q})$ and $k\_\mathrm{dist}(\mathbf{o})$, respectively. Note that by fixing $k$ to some reasonably small value, we have the desirable effect that the probability $p$ tends to zero as the dataset size $n$ increases.

Given these choices for $\epsilon$ and $p$, the expectation in the formulation of ALDR′ can be estimated by taking the average over the $k$ nearest neighbors of $\mathbf{q}$.

Using these approximation choices, we now state our proposed dimensionality-aware outlierness criterion, DAO:

$$\mathrm{DAO}_k(\mathbf{q}) \triangleq \frac{1}{k} \sum_{\mathbf{o} \in \mathrm{NN}_k(\mathbf{q})} \left( \frac{k\_\mathrm{dist}(\mathbf{q})}{k\_\mathrm{dist}(\mathbf{o})} \right)^{\widehat{\mathrm{ID}^*_{F_\mathbf{o}}}}, \tag{4.1}$$

where the neighborhood size $k$ is a hyperparameter, and the LID estimator $\widehat{\mathrm{ID}^*_{F_\mathbf{o}}}$ is left as an implementation choice.

Although DAO is theoretically justified as an estimator of ALDR′ by Theorem 4.1, it also can serve as an estimator of ALDR. Setting $\epsilon = \delta_{\mathbf{q}}(k/n)$, we obtain

$$\begin{aligned} \mathrm{ALDR}(\mathbf{q}) &\approx \frac{1}{k} \sum_{\mathbf{o} \in \mathrm{NN}_k(\mathbf{q})} \frac{F_\mathbf{o}(\delta_\mathbf{q}(k/n))}{F_\mathbf{q}(\delta_\mathbf{q}(k/n))} \\ &= \frac{1}{k} \sum_{\mathbf{o} \in \mathrm{NN}_k(\mathbf{q})} \frac{F_\mathbf{o}(\delta_\mathbf{q}(k/n))}{F_\mathbf{o}(\delta_\mathbf{o}(k/n))}, \end{aligned}$$

each term of which can be approximated using the limit equality stated in Theorem 4.1:

$$\mathrm{ALDR}(\mathbf{q}) \approx \frac{1}{k} \sum_{\mathbf{o} \in \mathrm{NN}_k(\mathbf{q})} \left( \frac{\delta_\mathbf{q}(k/n)}{\delta_\mathbf{o}(k/n)} \right)^{\mathrm{ID}^*_{F_\mathbf{o}}} \approx \mathrm{DAO}_k(\mathbf{q}).$$

We conclude this section by noting that DAO is nearly identical to SLOF, with the exception that the $k$-NN distance ratio of DAO has exponent equal to the LID of the neighbor $\mathbf{o}$. In essence, SLOF makes the implicit (but theoretically unjustified) assumption that the underlying local intrinsic dimensionalities are equal to 1 at every neighbor. We also note that the dimensionality-aware DAO criterion has a computational cost similar to that of SLOF whenever a linear-time LID estimator is employed (such as MLE [4, 32], reusing the $k$-NN queries also required to compute the outlier scores).

Table 1: Summary of 393 real datasets, with ranges showing numbers of features, dataset sizes, and proportions of outliers.

| Repository | Features | Size | Outliers | Datasets |
|---|---|---|---|---|
| Campos *et al.* [12] | [5, 259] | [50, 49534] | [3%, 36%] | 15 |
| Marques *et al.* [34] | [10, 649] | [100, 910] | [1%, 10%] | 3 |
| Rayana [40] | [6, 274] | [129, 7848] | [2%, 36%] | 11 |
| Goldstein & Uchida [16] | [27, 400] | [367, 49534] | [2%, 3%] | 3 |
| Emmott *et al.* [14] | [7, 128] | [992, 515129] | [9%, 50%] | 11 |
| Kandanaarachchi *et al.* [25] | [2, 649] | [72, 9083] | [1%, 3%] | 350 |
| Overall | [2, 649] | [50, 515129] | [1%, 50%] | 393 |

## 5 Evaluation

We compare DAO against its dimensionality-unaware counterpart SLOF [43], as well as LOF [11] and $k$NN [38], the two models with best overall performance from the extensive comparative study in [12].

In our experimentation, we employ four different estimators of local intrinsic dimensionality: the classical maximum-likelihood estimator (MLE) [4, 32], tight local estimation using pairwise distances (TLE) [5], two-nearest-neighbor point estimation (TwoNN) [15], and estimation derived from changes in parametric probability density after data perturbation (LIDL) [45].

For experiments on synthetic data, we generated 480 datasets consisting of two clusters ($c_1$ and $c_2$) embedded in $\mathbb{R}^{32}$, with each cluster containing 800 data points drawn from a standard Gaussian distribution ($\mu = \mathbf{0}$, $\Sigma = \mathbf{I}$). Cluster $c_1$ was generated within a subspace of dimension 8, and cluster $c_2$ within subspaces of dimensionality varying between 2 and 32. Extreme points with respect to the clusters were labeled as outliers. We also make use of 393 real-world datasets drawn from 6 different repositories for outlier detection, as summarized in Table 1. Further details can be found in the full version of the paper.

## 6 Experimental Results

The following is a summary of our experimental results. For more details, please see the full version of the paper.

**6.1 Comparative Evaluation on Synthetic Datasets.** We begin our analysis with the synthetic dataset collection, focusing on the relative performance between DAO and its 3 dimensionality-unaware competitors. From Figure 1, one can see that when both clusters share the same intrinsic dimensionality (8), DAO and its dimensionality-unaware competitors perform equally well. However, as the difference in the dimensionality of the cluster manifolds increases, the performances of SLOF, LOF, and $k$NN degrade noticeably. The experiments also show that of the various LID-aware variants considered, $\mathrm{DAO}_{\mathrm{MLE}}$ had

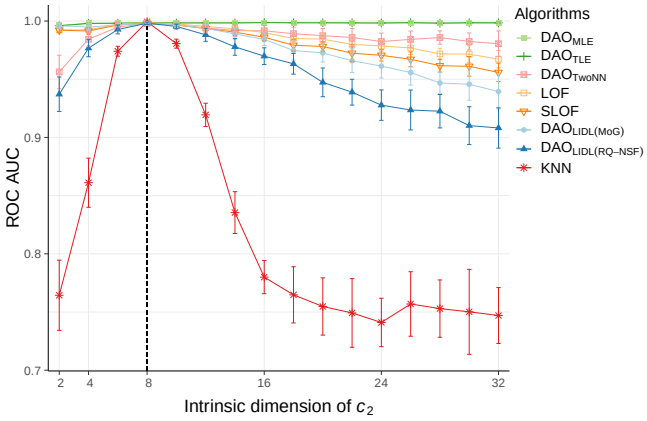



Figure 1: ROC AUC values for outlier detection performance over 480 synthetic datasets containing 2 clusters. One of the clusters ($c_1$) has intrinsic dimension fixed at 8. The intrinsic dimension of the other cluster ($c_2$) varies across the datasets ($x$-axis). The dashed vertical line indicates the reference set with both clusters sharing the same intrinsic dimension (8). The results shown are averages over 30 datasets with the same characteristics. Bars indicate standard deviation.

consistently superior performance as the dimensionality of cluster $c_2$ was varied.

Table 2 shows linear regression models fitted to predict the difference in ROC AUC between $\text{DAO}_{\text{MLE}}$ and each dimension-unaware method, as a function of the difference in the intrinsic dimension of the two data clusters manifolds. Among these methods, the greatest degradation of performance is that of $k$NN. From the slope of the linear regression, one can see that on average, the ROC AUC performance of $k$NN as compared to DAO decreased by almost 0.01 for each unit increment in the difference between the dimensions.

These experimental outcomes on synthetic data confirm the theoretical analysis in Section 4, in that the performance of SLOF is seen to degrade relative to its dimensionality-aware variant DAO, as the differences in the dimensions of the cluster subspaces increase. The degradation of LOF is slightly less rapid than that of its close variant SLOF. The performance drop for $k$NN is much more drastic, possibly due to its use of absolute distance thresholds as the outlier criterion. Unlike unitless 'local' methods based on density ratios such as LOF and SLOF, 'global' distance-based methods such as $k$NN favor the identification of points from higher-dimensional local distributions as outliers, due to the concentration effect associated with the so-called 'curse of dimensionality'. This tendency becomes more pronounced as the relative difference between the underlying dimensionalities increases.

Table 2: Simple linear regression to predict the difference in ROC AUC between $\text{DAO}_{\text{MLE}}$ and its dimensionality-unaware competitors on synthetic datasets. The explanatory variable is the absolute difference between the intrinsic dimensions of the two cluster manifolds. For each, we show the slope $m$, the $p$-value, and the Pearson correlation $\rho$.

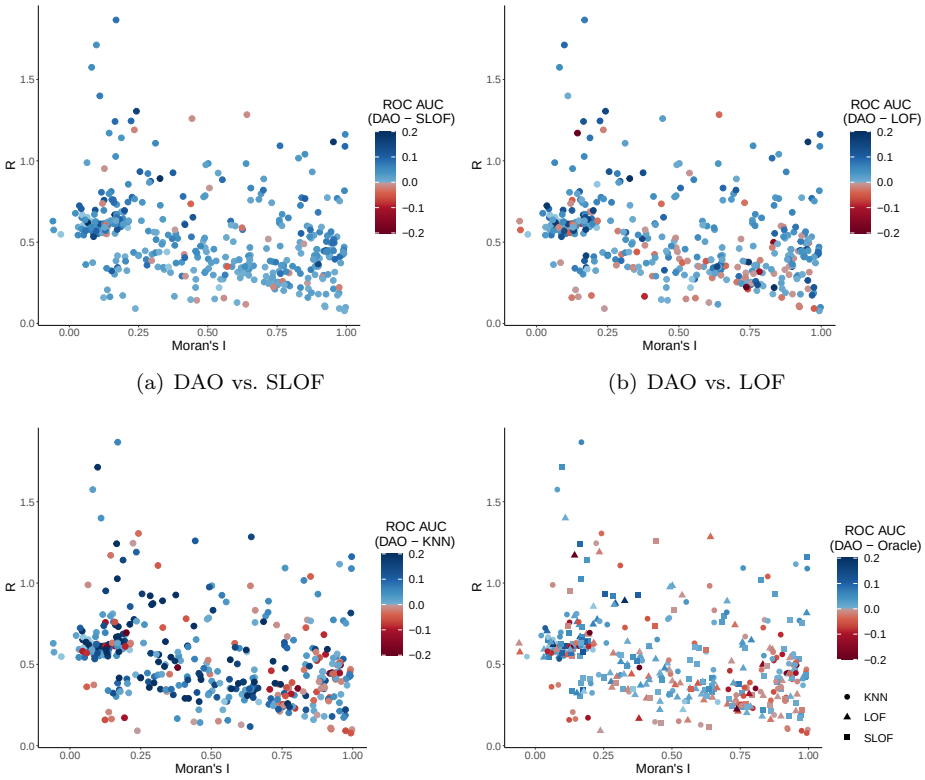
| ROC AUC | Regression on the absolute difference between the IDs of the manifolds | | |
|---|---|---|---|
| | $m$ | $p$ | $\rho$ |
| DAO : $k$NN | 0.0099 | 1e-4 | 0.806 |
| DAO : SLOF | 0.0018 | 8e-14 | 0.991 |
| DAO : LOF | 0.0013 | 3e-14 | 0.992 |

### 6.2 Comparative Evaluation on Real Datasets.

In Figure 2, for each of the 393 real datasets, we visualize the differences in ROC AUC performance between $\text{DAO}_{\text{MLE}}$ and the dimensionality-unaware outlier detection methods. Each colored dot in the scatterplot represents a single dataset, where blue indicates the outperformance of DAO relative to its competitor, and red indicates underperformance. The $y$-axis indicates the dispersion $R$ (mean absolute difference) of log-LID values computed at the data samples [23], and the $x$-axis shows their Moran's I autocorrelation [6].

In Figure 2(a), we compare the performance of DAO against SLOF. As discussed previously, SLOF can be seen as a dimensionally-unaware variant of DAO, which implicitly assumes that the local intrinsic dimensionality of the test point always equals 1. From the clear predominance of blue dots, one can see that ignoring the intrinsic dimension leads to a performance loss in most cases. When fitting linear regression to predict the difference in ROC AUC between DAO and SLOF (Table 3), the dispersion $R$ and the gain of performance of DAO relative to SLOF are seen to have a direct relationship, as indicated by the positive regression slope and Pearson correlation. On the other hand, an inverse relationship exists between the Moran's I autocorrelation and the performance of DAO relative to SLOF, as seen from the negative regression slope and Pearson correlation. In other words, as the correlation decreases between the intrinsic dimension at a query location and those of its neighbors' locations, DAO tends to outperform SLOF by a greater margin.

Overall, the results and major trends are similar when comparing DAO against LOF in Figure 2(b), against $k$NN in Figure 2(c), and even (to a lesser extent) against an oracle that uses the best-performing competitor for each individual dataset in Figure 2(d). Their respective regression analyses, shown in Table 3,




(a) DAO vs. SLOF (b) DAO vs. LOF

(c) DAO vs. $k$NN (d) DAO vs. Oracle

Figure 2: Differences in ROC AUC performance between $\text{DAO}_{\text{MLE}}$ and the dimensionality-unaware methods over 393 real datasets. Blue dots indicate datasets where DAO outperforms its competitor, whereas red dots indicate the opposite. The 'Oracle' method indicates the best-performing competitor for each individual dataset. Color intensity is proportional to the ROC AUC difference. On the $x$- and $y$-axis we show the Moran's I autocorrelation and dispersion $R$ (mean absolute difference) of log-LID estimates, respectively.

Table 3: Simple linear regression to predict the difference in ROC AUC between $\text{DAO}_{\text{MLE}}$ and its dimensionality-unaware competitors on real datasets. The explanatory variables are the dispersion $R$ and the Moran's I autocorrelation, both with respect to log-LID values. For each, we show the slope $m$, the $p$-value, and the Pearson correlation $\rho$.

| ROC AUC | $R$ (MAD) | | | Moran's I | | |
|---|---|---|---|---|---|---|
| | $m$ | $p$ | $\rho$ | $m$ | $p$ | $\rho$ |
| DAO : $k$NN | 0.059 | 6e-3 | 0.14 | -0.075 | 1e-5 | -0.21 |
| DAO : SLOF | 0.051 | 4e-12 | 0.34 | -0.021 | 5e-4 | -0.17 |
| DAO : LOF | 0.046 | 5e-6 | 0.23 | -0.016 | 5e-2 | -0.1 |

lead essentially to the same conclusions as for SLOF. It is worth noting that $k$NN exhibits the largest (absolute) regression coefficients, which is consistent with the results from the synthetic experiments.

Our experimentation reveals that dimensionality-aware outlier detection is of greatest advantage when the dataset has a complex LID profile, as indicated by a high dispersion ($R$ value) and/or a low autocorrelation (Moran's I value). The four scatterplots of Figure 2 all show that the dimensionality-unaware methods are more competitive when there is less contrast in the LID values across the dataset — that is, when the dispersion is low or the autocorrelation is high. Note that no outlier detection method can be expected to have perfect performance, as there are multiple factors that can favor any given model over any other [12]. For example, among the outlier models studied in this paper, $k$NN is known to be favored when the dataset contains many distance-based outliers.

We also summarize the overall results in a critical distance diagram (Figure 3), which shows the average ranks of the outlier detection methods with respect to ROC AUC, taken across the 393 real datasets.

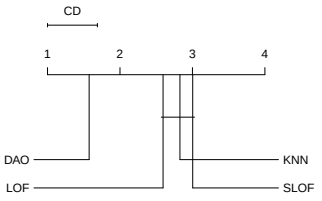



Figure 3: Critical difference diagram (significance level $\alpha$ = 1e-16) of average ranks of the methods on 393 real datasets: $\text{DAO}_{\text{MLE}}$ vs. baseline competitors.

The width of the upper bar (CD) indicates the critical distance of the well-known Friedman-Nemenyi statistical test at significance level $\alpha$ = 1e-16. The large gap between DAO and LOF serves as quantitative evidence that DAO outperformed its dimensionality-unaware competitors by a significant margin.

## 7 Conclusion

In our derivation of DAO via the theoretical LID model, and its subsequent empirical validation, we have made the case for a dimensionality-aware treatment of the problem of outlier detection. The theoretical and empirical evidence presented in this paper establishes that conventional, dimensionality-unaware approaches are susceptible to the variations and correlations in intrinsic dimensionality observed in most real datasets, and that the theory of local intrinsic dimensionality allows for a more principled treatment of outlierness.

Our analyses have shed some light on the fact that the quality of dimensionality-aware local outlier detection depends crucially on the properties of the estimator of LID. Estimators that learn by optimizing an objective function that favors inliers (such as LIDL), or those that perform smoothing (such as TLE), should be either avoided or used with caution. As our experiment results suggest, the use of an unsuitable estimator of LID may introduce errors that may outweight the benefits of dimensionality-aware techniques. It is still an open question as to which estimators of LIDs lead to the best outlier detection performance in practice. However, in our experimentation involving synthetic data, and the success of $\text{DAO}_{\text{MLE}}$ against top-performing nonparametric outlier methods (LOF, SLOF and $k$NN) on hundreds of real datasets, we have seen the emergence of the MLE estimator of LID as a sensible option for practical outlier detection tasks.

**Acknowledgement.** This study received partial funding under the Reliable Outlier Detection project from the Independent Research Fund Denmark.

## References

[1] Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Trans. Emerg. Telecommun. Technol.*, 32(1), 2021.

[2] Z. Alaverdyan, J. Jung, R. Bouet, and C. Lartizien. Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: Application to epilepsy lesion screening. *Medical Image Anal.*, 60, 2020.

[3] L. Amsaleg, J. Bailey, A. Barbe, S. M. Erfani, T. Furon, M. E. Houle, M. Radovanovic, and X. V. Nguyen. High intrinsic dimensionality facilitates adversarial attack: Theoretical evidence. *IEEE Trans. Inf. Forensics Secur.*, 16:854–865, 2021.

[4] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K. Kawarabayashi, and M. Nett. Extreme-value-theoretic estimation of local intrinsic dimensionality. *Data Min. Knowl. Discov.*, 32(6):1768–1805, 2018.

[5] L. Amsaleg, O. Chelly, M. E. Houle, K. Kawarabayashi, M. Radovanović, and W. Treeratanajaru. Intrinsic dimensionality estimation within tight localities: A theoretical and experimental analysis. *arXiv*, (2209.14475), 2022.

[6] L. Anselin. Local indicators of spatial association–LISA. *Geograph. Anal.*, 27(2):93–115, 1995.

[7] M. Aumüller and M. Ceccarello. The role of local dimensionality measures in benchmarking nearest neighbor search. *Inf. Syst.*, 101:101807, 2021.

[8] J. Bailey, M. E. Houle, and X. Ma. Local intrinsic dimensionality, entropy and statistical divergences. *Entropy*, 24(9):1220, 2022.

[9] V. Barnett. The study of outliers: Purpose and model. *Appl. Stat.*, 27(3):242–250, 1978.

[10] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *Proc. ICDT*, pages 217–235, 1999.

[11] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proc. SIGMOD*, pages 93–104, 2000.

[12] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Min. Knowl. Disc.*, 30:891–927, 2016.

[13] G. Casanova, E. Englmeier, M. E. Houle, P. Kröger, M. Nett, E. Schubert, and A. Zimek. Dimensional testing for reverse $k$-nearest neighbor search. *PVLDB*, 10(7):769–780, 2017.

[14] A. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. Wong. A meta-analysis of the anomaly detection problem. *arXiv*, (1503.01158), 2016.

[15] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a

minimal neighborhood information. *Scientific Reports*, 7(12140), 2017.
[16] M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE*, 11(4), 2016.
[17] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao. ADBench: Anomaly detection benchmark. In *NeurIPS*, 2022.
[18] D. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.
[19] M. E. Houle. Dimensionality, discriminability, density and distance distributions. In *Proc. ICDM Workshops*, pages 468–473, 2013.
[20] M. E. Houle. Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications. In *Proc. SISAP*, pages 64–79, 2017.
[21] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *Proc. SSDBM*, pages 482–500, 2010.
[22] M. E. Houle, E. Schubert, and A. Zimek. On the correlation between local intrinsic dimensionality and outlierness. In *Proc. SISAP*, pages 177–191, 2018.
[23] H. Huang, R. J. G. B. Campello, S. M. Erfani, X. Ma, M. E. Houle, and J. Bailey. LDReg: Local dimensionality regularized self-supervised learning. In *Proc. ICLR*, pages 1–26, 2024.
[24] W. Jin, A. K. H. Tung, J. Han, and W. Wang. Ranking outliers using symmetric neighborhood relationship. In *Proc. PAKDD*, pages 577–593, 2006.
[25] S. Kandanaarachchi, M. A. Muñoz, R. J. Hyndman, and K. Smith-Miles. On normalization and algorithm selection for unsupervised outlier detection. *Data Min. Knowl. Discov.*, 34(2):309–354, 2020.
[26] D. R. Karger and M. Ruhl. Finding nearest neighbors in growth-restricted metrics. In *Proc. STOC*, 2002.
[27] E. M. Knorr and R. T. Ng. A unified notion of outliers: Properties and computation. In *Proc. KDD*, 1997.
[28] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. LoOP: local outlier probabilities. In *Proc. CIKM*, pages 1649–1652, 2009.
[29] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Interpreting and unifying outlier scores. In *Proc. SDM*, pages 13–24, 2011.
[30] H.-P. Kriegel, M. Schubert, and A. Zimek. Angle-based outlier detection in high-dimensional data. In *Proc. KDD*, pages 444–452, 2008.
[31] L. J. Latecki, A. Lazarevic, and D. Pokrajac. Outlier detection with kernel density functions. In *Proc. MLDM*, pages 61–75, 2007.
[32] E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Proc. NIPS*, pages 777–784, 2004.
[33] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. M. Erfani, S. Xia, S. N. R. Wijewickrema, and J. Bailey. Dimensionality-driven learning with noisy labels. In *Proc. ICML*, pages 3361–3370, 2018.
[34] H. O. Marques, R. J. G. B. Campello, J. Sander, and A. Zimek. Internal evaluation of unsupervised outlier detection. *ACM Trans. Knowl. Discov. Data*, 14(4):47:1–47:42, 2020.
[35] H. O. Marques, L. Swersky, J. Sander, R. J. G. B. Campello, and A. Zimek. On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection, and ensembles. *Data Min. Knowl. Discov.*, 2023.
[36] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. LOCI: Fast outlier detection using the local correlation integral. In *Proc. ICDE*, 2003.
[37] M. Radovanović, A. Nanopoulos, and M. Ivanović. Reverse nearest neighbors in unsupervised distance-based outlier detection. *IEEE TKDE*, 2014.
[38] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proc. SIGMOD*, pages 427–438, 2000.
[39] D. T. Ramotsoela, A. M. Abu-Mahfouz, and G. P. Hancke. A survey of anomaly detection in industrial wireless sensor networks with critical water system infrastructure as a case study. *Sensors*, 18(8):2491, 2018.
[40] S. Rayana. ODDS library. `http://odds.cs.stonybrook.edu`, 2016.
[41] S. Romano, O. Chelly, V. Nguyen, J. Bailey, and M. E. Houle. Measuring dependency via intrinsic dimensionality. In *Proc. ICPR*, pages 1207–1212, 2016.
[42] E. Schubert, A. Zimek, and H.-P. Kriegel. Generalized outlier detection with flexible kernel density estimates. In *Proc. SDM*, pages 542–550, 2014.
[43] E. Schubert, A. Zimek, and H.-P. Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Min. Knowl. Disc.*, 28(1):190–237, 2014.
[44] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung. Enhancing effectiveness of outlier detections for low density patterns. In *Proc. PAKDD*, pages 535–548, 2002.
[45] P. Tempczyk, R. Michaluk, L. Garncarek, P. Spurek, J. Tabor, and A. Golinski. LIDL: local intrinsic dimension estimation using approximate likelihood. In *Proc. ICML*, pages 21205–21231, 2022.
[46] K. Zhang, M. Hutter, and H. Jin. A new local distance-based outlier detection approach for scattered real-world data. In *Proc. PAKDD*, pages 813–822, 2009.
[47] A. Zimek and P. Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. *WIREs Data Mining Knowl. Discov.*, 8(6), 2018.
[48] A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat. Anal. Data Min.*, 5(5):363–387, 2012.