Outlier Detection: A Novel Depth Approach

Yixin Chen*, Xin Dang†, Hanxiang Peng‡

November 29, 2008

Abstract

Statistical depth functions provide from the "deepest" point a "center-outward ordering" of multi-dimensional data. In this sense, depth functions can measure the "extremeness" or "outlyingness" of a data point with respect to a given data set. Hence they can detect outliers – observations that appear extreme relative to the rest of the observations. Of the various statistical depths, the spatial depth is especially appealing because of its computational efficiency and mathematical tractability. This chapter presents a survey of a novel statistical depth, the *kernelized spatial depth* (KSD), and a novel *outlier detection algorithm* based on the KSD. A significant portion of the chapter is built upon the material from the articles we have written, in particular [11].

1 Introduction

In a variety of applications, e.g., network security [14, 20, 32], visual surveillance [24], remote sensing [5], medical diagnostics [26], and image processing [18], it is of great importance to identify observations that are "inconsistent" with the "normal" data. The research problem underlying these applications is commonly referred to as *outlier detection* [6]. Outlier detection has been investigated extensively over the last several decades by researchers from statistics, data mining, and machine learning communities. Next we review some of the related work. For a more comprehensive survey of this subject, the reader is referred to Barnett and Lewis [6], Hawkins [23], and Markou and Singh [33, 34].

*Yixin Chen is with the Department of Computer and Information Science, The University of Mississippi, University, MS 38677, USA. E-mail: ychen@cs.olemiss.edu.

†Xin Dang is with the Department of Mathematics, The University of Mississippi, University, MS 38677, USA. E-mail: xdang@olemiss.edu.

‡Hanxiang Peng is with the Department of Mathematical Science, Purdue School of Science, IUPUI, Indianapolis, IN 46202, USA. E-mail: hpeng@math.iupui.edu.

1.1 Statistical Methods

Outlier detection is recognized to be important in statistical analysis. If a classical statistical method is blindly applied to data containing outliers, the results can be misleading at best. Outliers themselves are actually informative and often of interest in many practical situations. They may provide additional insights to the model under consideration. In the statistical literature, there are two commonly used approaches for multi-dimensional outlier detection – distance-based methods and projection pursuit.

Distance-based methods aim to detect outliers by computing a distance measure of a particular point to the centroid of a data. The point with such "outlyingness" measure above a threshold is claimed as an outlier. The most commonly used measure is perhaps the Mahalanobis distance (MD) which is defined as

$$MD(x, F_n) = \sqrt{(x - \hat{\mu})^T \hat{\Sigma}^{-1} (x - \hat{\mu})},$$

where F_n is the empirical distribution calculated from the data and $\hat{\mu}$, $\hat{\Sigma}$ are location, scatter estimators, respectively. The Mahalanobis distance is a more meaningful measure of distance than the Euclidean distance. It incorporates both the variability of each marginal direction and the correlation structure of a data, and possesses the affine invariant property.

In the outlier detection context, the location and scale estimators must be resistant to outliers, and can be obtained either by robust methods or by forward search technique. Possessing many good properties and being available a fast algorithm developed by Rousseeuw and van Driessen [46], the minimum covariance determinant (MCD) estimator is perhaps the most commonly used scale estimator. Other commonly used robust methods include M estimators [35] and S estimators [45, 48]. The sequentially forward search [4, 19, 7] starts with a small subset of observations presumed to be outlier-free, to which it iteratively adds points that have a small MD based on $\hat{\mu}$ and $\hat{\Sigma}$ of the current subset. Outlier region based on the MD is constrained to have elliptical contours, which do not follow the shape of the distribution unless the underlying model is elliptically distributed.

The projection pursuit (PP) approach intents to reduce a multivariate detection problem to a set of univariate problems through looking at projections of the data onto some (univariate) directions. The techniques for one-dimensional outlier identification are extensive [23, 6]. The reference [6] includes 47 tests for normal data, 23 tests for data having a gamma distribution, and 17 tests for data having other distributions. The key of the projection pursuit approach is to find the "interesting" directions.

Gnanadesikan and Kettenring [17] proposed to obtain the principal components (PCs) of the data and search for outliers in the first few PCs. Rao [41] and Hawkins [22] argued that the last few PCs are likely to be more useful than the first few in detecting outliers that are not apparent from the original variables. Caussinus and Ruiz [13] introduced a metric weight and examined the first few generalized PCs for outliers. Peña and Prieto [38] suggested the directions based on values of the kurtosis coefficients of the projected data. The PP provides the correct solution when the outliers are located close to the considered directions. Projection onto a low-dimensional space may provide a graphical visualization for the behavior of outliers. It, however, may fail to identify outliers in the general case.

Extensive attempts have been made to adopt the above two approaches for outlier detection in structured data, for example, in regression analysis, time series, and directional data. We refer keen readers to [47] for a comprehensive survey.

1.2 Machine Learning Methods

From a machine learning perspective, outlier detection can be categorized into *a missing label problem* and *a one-class learning problem*, depending on the way in which the normal samples are defined in a training data set. In a missing label problem, the data of interest consist of a mixture of normal samples and outliers, in which the labels are missing. The goal is to identify outliers from the data and, in some applications, to predict outliers from unseen data. In a one-class learning problem, normal samples are given as the training data. An outlier detector is built upon the normal samples to detect observations that deviate markedly from the normal samples, i.e., outliers. This is closely related to the standard supervised learning problem except that all the training samples have the same *normal* label.

1.2.1 Outlier Detection as a Missing Label Problem

Because only unlabeled samples are available in a missing label problem, prior assumptions are needed in order to define and identify outliers. Frakt et al. [15], proposed an anomaly detection framework for tomographic data where an image is modeled as a superposition of background signal and anomaly signal. Background signal is a zero mean, wide-sense stationary, Gaussian random field with a known covariance. Anomaly signal is assumed to be zero everywhere except over a square patch, with prior knowledge of minimal and maximal possible size, where it is constant. As a result,

anomaly detection is equivalent to determining whether or not an image region is identically zero, which is formulated as a multiscale hypothesis testing problem. Reed and Yu [43] developed an anomaly detection algorithm for detecting targets of an unknown spectral distribution against a background with an unknown spectral covariance. The background is modeled as a Gaussian distribution with zero mean and an unknown covariance matrix. The target is described by a Gaussian distribution with the mean equal to the known signature of the target and the covariance matrix identical to that of the background. Kwon and Nasrabadi [29] introduced a nonlinear version of Reed and Yu's algorithm using feature mappings induced by positive definite kernels. Kollios et al. [28] observed that the density of a data set contains sufficient information to design sampling techniques for clustering and outlier detection. In particular, when outliers mainly appear in regions of low density, a random sampling method that is biased towards sparse regions can recognize outliers with high probability.

All the aforementioned algorithms have one characteristic, the key component of the method, in common: the estimation of probability density functions. There are several algorithms in the literature that are developed based upon the geometric aspects of a data set rather than upon distributional assumptions, in particular, the distance-based algorithms. Knorr and Ng [27] introduced the notion of distance-based outliers, the $DB(p, d)$-*outlier*. A data point $\mathbf{x}$ in a given data set is a $DB(p, d)$-outlier if at least p fraction of the data points in the data set lies more than d distance away from $\mathbf{x}$. The parameters p and d are to be specified by a user. Ramaswamy et al. [40] extended the notion of distance-based outliers by ranking each point on the basis of its distance to its k-th nearest neighbor and declaring the top n points as outliers. Under the notions in [27] and [40], outliers are defined based on a global view of the data set. Breunig et al. [8] proposed the local outlier factor (LOF) that takes into consideration the local structure of the data set. The LOF of a data point is computed using the distances between the point and its "close" neighbors. Hence LOF describes how isolated a data point is with respect to its surrounding neighbors. Tang et al. [55] defined the connectivity-based outlier factor that compares favorably with LOF at low density regions. Along the line of Breunig et al. [8], Sun and Chawla [54] introduced a measure for spatial local outliers, which takes into account both spatial autocorrelation and spatially non-uniform variance of the data. Angiulli et al. [3] designed a distance-based method to find outliers from a given data set and to predict if an unseen data point is an outlier based on a carefully selected subset of the given data. Aggarwal and Yu [2] investigated the influence of high dimensionality on distance-based

outlier detection algorithms. It is observed that most of the above distance-based approaches become less meaningful for sparse high dimensional data. Therefore, projection methods are tested for outlier detection. Lazarevic and Kumar [31] proposed a feature bagging approach to handle high dimensionality. The method combines outputs of multiple outlier detectors, each of which is built on a randomly selected subset of features.

1.2.2 Outlier Detection as a One-Class Learning Problem

When normal observations are given as a training data set, outlier detection can be formulated as finding observations that significantly deviate from the training data. A statistically natural tool for quantifying the deviation is the probability density of the normal observations. Roberts and Tarassenko [44] approximated the distribution of the training data by a Gaussian mixture model. For every observation, an outlier score is defined as the maximum of the likelihood that the observation is generated by each Gaussian component. An observation is identified as an outlier if the score is less than a threshold. Schweizer and Moura [51] modeled normal data, background clutter in hyperspectral images, as a 3-dimensional Gauss-Markov random field. Several methods are developed to estimate the random field parameters. Miller and Browning [36] proposed a mixture model for a set of labeled and unlabeled samples. The mixture model includes two types of mixture components: predefined components and nonpredefined components. The former generate data from known classes and assume class labels are missing at random. The latter only generate unlabeled data, corresponding to the outliers in the unlabeled samples. Parra et al. [37] proposed a class of volume conserving maps (i.e., those with unit determinant of Jacobian matrix) that transforms an arbitrary distribution into a Gaussian. Given a decision threshold, novelty detection is based on the corresponding contour of the estimated Gaussian density, i.e., novelty lies outside the hypersphere defined by the contour.

Instead of estimating the probability density of the normal observations, Schölkopf et al. [50] introduced a technique to capture the support of the probability density, i.e., a region in the input space where most of the normal observations reside in. Hence outliers lie outside the boundary of the support region. The problem is formulated as finding the smallest hypersphere to enclose most of the training samples in a kernel induced feature space, which can be converted to a quadratic program. Because of its similarity to support vector machines (SVM) from an optimization viewpoint, the method is called 1-class SVM. Along the line of 1-class SVM, Campbell and Bennett [12]

estimated the support region of a density using hyperplanes in a kernel induced feature space. The "optimal" hyperplane is defined as one that puts all normal observations on the same side of the hyperplane (the support region) and as close to the hyperplane as possible. Such a hyperplane is the solution of a linear program. Rätsch et al. [42] developed a boosting algorithm for one-class classification based on connections between boosting and SVMs. Banerjee et al. [5] applied 1-class SVM for anomaly detection in hyperspectral images and demonstrated improved performance compared with the method described in [43].

There is an abundance of prior work that applies standard supervised learning techniques to tackle outlier detection [1, 21, 53]. These methods generate a labeled data set by assigning one label to the given normal examples and the other label to a set of artificially generated outliers. Han and Cho [21] use artificially generated intrusive sequences to train an evolutionary neural network for intrusion detection. Abe et al. [1] propose a selective sampling method that chooses a small portion of artificial outliers in each training iteration. In general, the performance of these algorithms depends on the choice of the distribution of the artificial examples and the employed sampling plan. Steinwart et al. [53] provide an interesting justification for the above heuristic by converting outlier detection to a problem of finding level sets of data generating density.

1.3 An Overview of the Chapter

In this Chapter, we present a survey of a novel outlier detection framework based on the notion of *statistical depths* [11]. Outlier detection methods that are based on statistical depths have been studied in statistics and computational geometry [39, 49]. These methods provide a center-outward ordering of observations. Outliers are expected to appear more likely in outer layers with small depth values than in inner layers with large depth values. Depth-based methods are completely data-driven and avoid strong distributional assumption. Moreover, they provide intuitive visualization of the data set via depth contours for a low dimensional input space. However, most of the current depth-based methods do not scale up with the dimensionality of the input space.

Of the various depths the *spatial depth* is especially appealing because of its computational efficiency and mathematical tractability [52]. Its computational complexity is of magnitude $O(\ell^2)$, independent of dimension d. Because each observation from a data set contributes equally to the value of depth function, spatial depth takes a global view of the data set. Conse-

quently the outliers can be called as "global" outliers. Nevertheless, many data sets from real-world applications exhibit more delicate structures that entail identification of outliers relative to their neighborhood, i.e., "local" outliers. We survey an outlier detection framework that avoids the above limitation of spatial depth.

The remainder of the chapter is organized as follows. Section 2 motivates the spatial depth-based outlier detection via the connection between the spatial depth and the L_1 median. Section 3 introduces the kernelized spatial depth. Section 4 presents several upper bounds on the false alarm probability of the proposed kernelized spatial depth-based outlier detectors for a one-class learning problem and a missing label problem. Experimental results are reported in Section 5. We conclude in Section 6.

2 The Spatial Depth Function and Outlier Detection

As Barnett and Lewis described [6], *"what characterizes the 'outlier' is its impact on the observer (not only will it appear extreme but it will seem, to some extent, surprisingly extreme)"*. An intuitive way of measuring the extremeness is to examine the relative location of an observation with respect to the rest of the population. An observation that is far away from the center of the distribution is more likely to be an outlier than observations that are closer to the center. This suggests a simple outlier detection approach based on the distance between an observation and the center of a distribution.

2.1 The Spatial Depth

Although both the sample mean and median of a data set are natural estimates for the center of a distribution, the median is insensitive to extreme observations while the mean is highly sensitive. A single contaminating point to a data set can send the sample mean, in the worst case, to infinity, whereas in order to have the same effect on the median, at least 50% of the data points must be moved to infinity. Let $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ be observations from a univariate distribution F and $\mathbf{x}_{(1)} \leq \ldots \leq \mathbf{x}_{(\ell)}$ be the sorted observations in an ascending order. The sample median is $\mathbf{x}_{((\ell+1)/2)}$ when ℓ is odd. When ℓ is even, any number in the interval $[\mathbf{x}_{(\ell/2)}, \mathbf{x}_{((\ell+1)/2)}]$ can be defined to be the sample median. A convenient choice is the average $\frac{\mathbf{x}_{(\ell/2)}+\mathbf{x}_{((\ell+1)/2)}}{2}$. Next, we present an equivalent definition that can be naturally generalized to a higher dimensional setting.

Let $s : \mathbb{R} \to \{-1, 0, 1\}$ be the sign function, i.e.,

$$s(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}}{|\mathbf{x}|}, & \mathbf{x} \neq 0, \\ 0, & \mathbf{x} = 0. \end{cases}$$

For $\mathbf{x} \in \mathbb{R}$, the difference between the numbers of observations on the left and right of $\mathbf{x}$ is $\left|\sum_{i=1}^{\ell} s(\mathbf{x}_i - \mathbf{x})\right|$. There are an equal number of observations on both sides of the sample median, so that the sample median is

$$\text{any } \mathbf{x} \in \mathbb{R} \text{ that satisfies } \left|\sum_{i=1}^{\ell} s(\mathbf{x}_i - \mathbf{x})\right| = 0. \tag{1}$$

Replacing the absolute value $|\cdot|$ with the 2-norm (Euclidean norm) $\|\cdot\|$, the sign function is readily generalized to multidimensional data: *the spatial sign function* [58] or *the unit vector* [9], which is a map $S : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$S(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}}{\|\mathbf{x}\|}, & \mathbf{x} \neq \mathbf{0}, \\ \mathbf{0}, & \mathbf{x} = \mathbf{0} \end{cases}$$

where $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}}$ and $\mathbf{0}$ is the zero vector in $\mathbb{R}^n$. With the spatial sign function, the *multidimensional sample median* for multidimensional data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\} \subset \mathbb{R}^n$ is a straightforward analogy of the univariate sample median (1), i.e., it is

$$\text{any } \mathbf{x} \in \mathbb{R}^n \text{ that satisfies } \left\|\sum_{i=1}^{\ell} S(\mathbf{x}_i - \mathbf{x})\right\| = 0. \tag{2}$$

The median defined in (2) is named as the *spatial median* [58] or the L_1 *median* [57, 56].

The concept of spatial depth was formally introduced by Serfling [52] based on the notion of the spatial quantiles proposed by Chaudhuri [10], while a similar concept, the L_1 depth, was described by Vardi and Zhang [56]. For a multivariate cumulative distribution function (cdf) F on $\mathbb{R}^n$, the spatial depth of a point $\mathbf{x} \in \mathbb{R}^n$ with respect to the distribution F is defined as

$$D(\mathbf{x}, F) = 1 - \left\| \int S(\mathbf{y} - \mathbf{x}) dF(\mathbf{y}) \right\|.$$

For an unknown cdf F, the spatial depth is unknown and can be approximated by the *sample spatial depth*:

$$D(\mathbf{x}, \mathcal{X}) = 1 - \frac{1}{|\mathcal{X} \cup \{\mathbf{x}\}| - 1} \left\| \sum_{\mathbf{y} \in \mathcal{X}} S(\mathbf{y} - \mathbf{x}) \right\| \tag{3}$$

where $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\}$ and $|\mathcal{X} \cup \{\mathbf{x}\}|$ denotes the cardinality of the union $\mathcal{X} \cup \{\mathbf{x}\}$. Note that both $D(\mathbf{x}, F)$ and its sample version have a range $[0, 1]$.

Observing (2) and (3), it is easy to see that the depth value at the spatial median is one. In other words, the spatial median is a set of data points that have the "deepest" depth value one. Indeed, the spatial depth provides from the "deepest" point a "center-outward" ordering of a multidimensional data. The depth attains the maximum value at the deepest point and decreases to zero as the point $\mathbf{x}$ moves away from the deepest to infinity. Thus the depth value of a point gives us a measure of the "extremeness" or "outlyingness" of a data point, which can be used for *outlier detection*. From now on all depths are referred to the sample depth.

2.2 Outlier Detection Using the Spatial Depth

Figure 1 shows a contour plot of the spatial depth $D(\mathbf{x}, \mathcal{X})$ based on 100 random observations (marked with $\circ$'s) generated from a 2-dimensional Gaussian distribution with mean zero and a covariance matrix whose diagonal and off-diagonal entries are 2.5 and -1.5, respectively. On each contour the depth function is constant with the indicated value. The depth values decrease outward from the "center" (i.e., the spatial median) of the cloud. This suggests that a point with a low depth value is more likely to be an outlier than a point with a high depth value. For example, the point on the upper right corner on Figure 1 (marked with $*$) has a very low depth value of 0.0539. It is isolated and far away from the rest of the data points. This example motivates a simple outlier detection algorithm: *Identify a data point as an outlier if its depth value is less than a threshold.*

In order to make this algorithm a practical method, the following two issues need to be addressed:

1. The spatial depth function captures the structure of a data cloud.

2. A method to decide a threshold value.

We postpone the discussion on the second issue to Section 4 where we present a framework to determine the threshold. The first issue is related to the shape of depth contours. The depth contours of a spatial depth function tend to be circular [25], especially at low depth values (e.g., the outer contour in Figure 1). For a spherical symmetric distribution, such contours fit nicely to the shape of the data cloud. It is therefore reasonable to consider a data point as an outlier if its depth value is low because a lower depth implies a

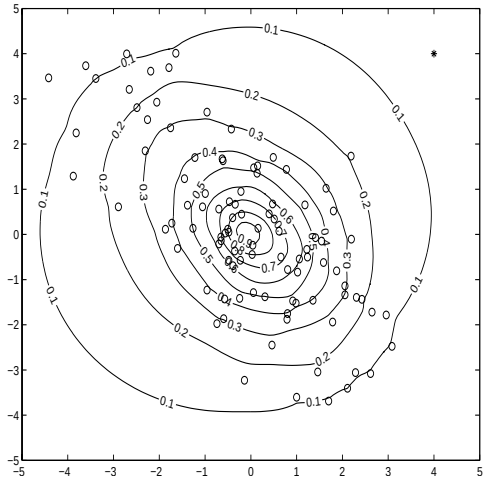


Figure 1: A contour plot of the sample spatial depth based on 100 random observations (represented by ∘'s) from a 2-dimensional Gaussian distribution. The depth values are indicated on the contours. A possible outlier is the observation (marked with ∗) on the upper left corner which has a very low depth value 0.0539.

larger distance from the "center" of the data cloud, which is defined by the spatial median. However, in general, the relationship between a depth value and outlyingness in a data may not be as straightforward as is depicted in Figure 1. For example, Figure 2 shows the contours of the spatial depth function based on 100 random observations generated from a ring shaped distribution. From the shape of the distribution, it is reasonable to view the point (marked with ∗) in the center as an outlier. However, the depth value at the location ∗ is 0.9544. A threshold larger than 0.9544 would classify all of the 100 observations as outliers.

The above example demonstrates that the spatial depth function may not capture the structure of a data cloud in the sense that a point isolated from the rest of the population may have a large depth value. This is due to the fact that the value of the depth function at a point depends only upon the resultant vector of the unit vectors, each of which represents the direction from the point to an observation. This definition, on one hand, downplays the significance of distance hence reduces the impact of those extreme observations whose extremity is measured in (Euclidean) distance, so that it gains *resistance against these extreme observations*. On the other hand, the acquirement of the *robustness* of the depth function trades off

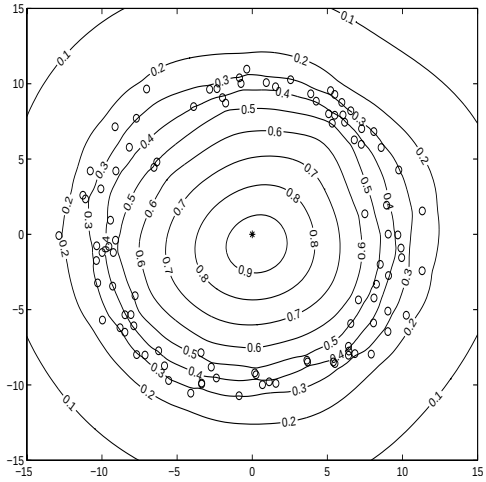


Figure 2: Contour plots of the sample spatial depths based on 100 random observations (denoted by ∘'s) of a ring shaped distribution. The depth values are indicated on the contours. The observation (denoted by $*$) at the center of the plot represents a possible outlier. The depth values for the $*$ observation is 0.9544.

some distance measurement, resulting in certain loss of the measurement of the *similarity* of the data points expressed in the Euclidean distance. The distance of a point from the data cloud plays an important role in revealing the structure of the data cloud. In [11], we proposed a method to tackle this limitation of spatial depth by incorporating into the depth function a distance metric (or a similarity measure) induced by a *positive definite kernel function*.

3 The Kernelized Spatial Depth

In various applications of machine learning and pattern analysis, carefully recoding the data can make "patterns" standing out. Positive definite kernels provide a computationally efficient way to recode the data. A positive definite kernel, $\kappa : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, implicitly defines an embedding map

$$\phi : \mathbf{x} \in \mathbb{R}^n \longmapsto \phi(\mathbf{x}) \in \mathbb{F}$$

via an inner product in the feature space $\mathbb{F}$,

$$\kappa(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

For certain stationary kernels [16], e.g. the Gaussian kernel $\kappa(\mathbf{x},\mathbf{y}) = \exp\left(\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2\right)$, $\kappa(\mathbf{x},\mathbf{y})$ can be interpreted as a *similarity* between $\mathbf{x}$ and $\mathbf{y}$, hence it encodes a similarity measure.

The basic idea of the *kernelized spatial depth* is to evaluate the spatial depth in a feature space induced by a positive definite kernel. Noticing that

$$\|\mathbf{x}-\mathbf{y}\|^2 = \langle \mathbf{x},\mathbf{x}\rangle + \langle \mathbf{y},\mathbf{y}\rangle - 2\langle \mathbf{x},\mathbf{y}\rangle = \mathbf{x}^T\mathbf{x} + \mathbf{y}^T\mathbf{y} - 2\mathbf{x}^T\mathbf{y},$$

with simple algebra, one rewrites the norm in (3) as

$$\left\| \sum_{\mathbf{y}\in\mathcal{X}} S(\mathbf{y}-\mathbf{x}) \right\|^2 = \sum_{\mathbf{y},\mathbf{z}\in\mathcal{X}} \frac{\mathbf{x}^T\mathbf{x} + \mathbf{y}^T\mathbf{z} - \mathbf{x}^T\mathbf{y} - \mathbf{x}^T\mathbf{z}}{(\mathbf{x}^T\mathbf{x} + \mathbf{y}^T\mathbf{y} - 2\mathbf{x}^T\mathbf{y})^{1/2}(\mathbf{x}^T\mathbf{x} + \mathbf{z}^T\mathbf{z} - 2\mathbf{x}^T\mathbf{z})^{1/2}}.$$

Replacing the inner products with the values of kernel κ, we obtain the *(sample) kernelized spatial depth (KSD) function*

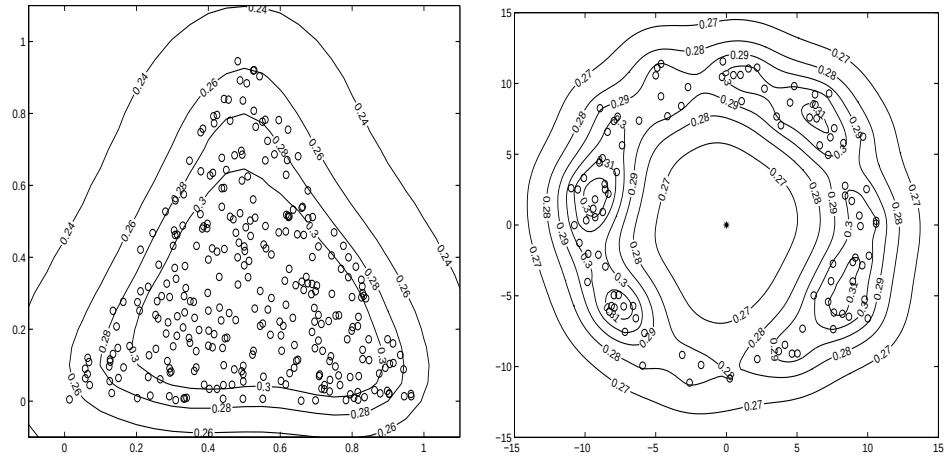
$$D_\kappa(\mathbf{x},\mathcal{X}) = 1 - \frac{1}{|\mathcal{X}\cup\{\mathbf{x}\}|-1}\sqrt{\sum_{\mathbf{y},\mathbf{z}\in\mathcal{X}} \frac{\kappa(\mathbf{x},\mathbf{x}) + \kappa(\mathbf{y},\mathbf{z}) - \kappa(\mathbf{x},\mathbf{y}) - \kappa(\mathbf{x},\mathbf{z})}{\delta_\kappa(\mathbf{x},\mathbf{y})\delta_\kappa(\mathbf{x},\mathbf{z})}} \tag{4}$$

where $\delta_\kappa(\mathbf{x},\mathbf{y}) = \sqrt{\kappa(\mathbf{x},\mathbf{x}) + \kappa(\mathbf{y},\mathbf{y}) - 2\kappa(\mathbf{x},\mathbf{y})}$. Analogous to the spatial sign function at $\mathbf{0}$, we define

$$\frac{\kappa(\mathbf{x},\mathbf{x}) + \kappa(\mathbf{y},\mathbf{z}) - \kappa(\mathbf{x},\mathbf{y}) - \kappa(\mathbf{x},\mathbf{z})}{\delta_\kappa(\mathbf{x},\mathbf{y})\delta_\kappa(\mathbf{x},\mathbf{z})} = 0$$

for $\mathbf{x}=\mathbf{y}$ or $\mathbf{x}=\mathbf{z}$.

The KSD (4) is defined for any positive definite kernels. Here we shall be particularly interested in *stationary kernels* (e.g., the Gaussian kernel), because of their close relationship with similarity measures. Figure 3 shows the two contour plots of the KSD based on 100 random observations generated from the two distributions presented in Figure 2, the half-moon distribution (Figure 3.a) and the ring-shaped distribution (Figure 3.b). The Gaussian kernel with $\sigma = 3$ is used to kernelize the spatial depth. Interestingly, unlike the spatial depth, we observe that the kernelized spatial depth captures the shapes of the two data sets. Specifically, the contours of KSD follow closely the respective shapes of the data clouds. Moreover, the depth values are small for the possible outliers. The depth values at the locations $(*)$, which can be viewed as outliers, are 0.2495 for the half-moon data and 0.2651 for the ring-shaped data. Consequently a threshold of 0.25 (or 0.27) can separate the outliers from the rest of the half-moon data (or ring data). The remaining question is how we determine the threshold. This is addressed in the following section.



(a) Triangular data (b) Ring data

Figure 3: The contour plots of the KSD functions based on 100 random observations (marked with ∘'s) from (a) a triangular distribution and (b) a ring-shaped distribution. The depth values are marked on the contours. The depth is kernelized with the Gaussian kernel $\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(\|\mathbf{x} - \mathbf{y}\|^2/\sigma^2\right)$ with $\sigma = 3$. The observation (marked with $*$) at the center of each plot represents a possible outlier. The depth values for the observation $*$ in (b) is 0.2651.

4 Bounds on the False Alarm Probability

The idea of selecting a threshold is rather simple, i.e., choose a value which controls the *false alarm probability (FAP)* under a given significance level. FAP is the probability that normal observations are classified as outliers. In the following, we first derive probabilistic bounds on FAP formulated as a one-class learning problem. We then extend the results to a missing label problem. The proofs of the results in this section can be found in [11].

4.1 The One-Class Learning Problem

Outlier detection formulated as a one-class learning problem can be described as follows. We have observations $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\} \subset \mathbb{R}^n$ from an unknown cdf, F_{good}. Based on the observations $\mathcal{X}$, a given datum $\mathbf{x}$ is classified as a *normal observation* or an *outlier* according to whether or not it is generated from F_{good}. Let $g : \mathbb{R}^n \to [0, 1]$ be an outlier detector where $g(\mathbf{x}) = 1$ indicates that $\mathbf{x}$ is an outlier. The FAP of an outlier detector g, $P_{FA}(g)$, is the probability that an observation generated from F_{good} is classified by the detector g as an outlier, i.e.

$$P_{FA}(g) = \int_{\mathbf{x} \in \mathcal{R}_o} dF_{good}(\mathbf{x})$$

where $\mathcal{R}_o = \{\mathbf{x} \in \mathbb{R}^n : g(\mathbf{x}) = 1\}$ is the collection of all observations that are classified as outliers. The FAP can be estimated by the *false alarm rate*, $\hat{P}_{FA}(g)$, which is computed by

$$\hat{P}_{FA}(g) = \frac{|\{\mathbf{x} \in \mathcal{X} : g(\mathbf{x}) = 1|}{|\mathcal{X}|}.$$

Consider a KSD-based outlier detector depicted in Figure 4 where $t \in [0, 1]$ is a threshold and b determines the rate of transition of output from 1 to 0. For a given data set $\mathcal{X}$ and kernel κ and $b \in [0, 1]$, we define an outlier detector $g_\kappa(\mathbf{x}, \mathcal{X})$ by

$$g_\kappa(\mathbf{x}, \mathcal{X}) = \begin{cases} 1, & \text{if } D_\kappa(\mathbf{x}, \mathcal{X}) \le t, \\ \frac{t+b-D_\kappa(\mathbf{x}, \mathcal{X})}{b}, & \text{if } t < D_\kappa(\mathbf{x}, \mathcal{X}) \le t + b, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

An observation $\mathbf{x}$ is classified as an outlier according to $g_\kappa(\mathbf{x}, \mathcal{X}) = 1$. Denote $\mathbb{E}_F$ the expectation calculated under cdf F. We have the following theorem for the bound of the FAP.

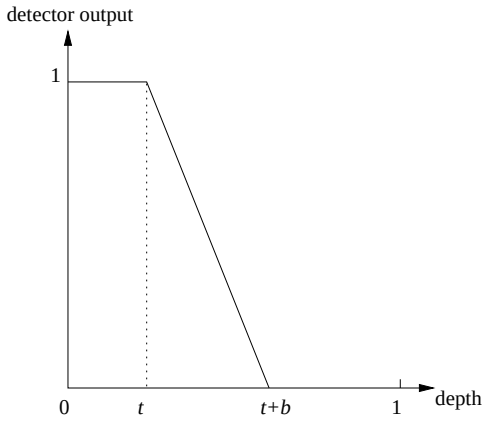


Figure 4: A depth-based outlier detector. An output value of 1 indicates an outlier, i.e., an observation with depth smaller than t is classified as an outlier.

Theorem 1 *Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\} \subset \mathbb{R}^n$ be an independent and identically distributed(i.i.d.) sample from cdf F. Let $g_\kappa(\mathbf{x}, \mathcal{X})$ be an outlier detector defined in (5). Fix $\delta \in (0,1)$. For a new random observation $\mathbf{x}$ from F, the following inequality holds with probability at least $1-\delta$:*

$$\mathbb{E}_F\left[g_\kappa(\mathbf{x}, \mathcal{X})\right] \leq \frac{1}{\ell}\sum_{i=1}^{\ell} g_\kappa(\mathbf{x}_i, \mathcal{X}) + \frac{2}{\ell b} + \left(1 + \sqrt{1 + \left(1 + \frac{4}{b}\right)^2}\right)\sqrt{\frac{\ln(2/\delta)}{2\ell}}\,. \tag{6}$$

It is worthwhile to note that there are two sources of randomness in the above inequality: the random sample $\mathcal{X}$ and the random observation $\mathbf{x}$. For a specific $\mathcal{X}$, the above bound is either true or false, i.e., it is not random. For a random sample $\mathcal{X}$, the probability that the bound is true is at least $1-\delta$. For a one-class learning problem, we can let $F = F_{good}$. It is not difficult to show that $P_{FA}(g_\kappa) \leq \mathbb{E}_F\left[g_\kappa(\mathbf{x}, \mathcal{X})\right]$ and $\hat{P}_{FA}(g_\kappa) \leq \frac{1}{\ell}\sum_{i=1}^{\ell} g_\kappa(\mathbf{x}_i, \mathcal{X})$ where the equalities hold when $b = 0$. This suggests that (6) provides us an upper bound on the FAP.

Theorem 1 suggests that we can control the FAP by adjusting the t parameter of the detector. Although t does not appear explicitly in (6), it affects the value of $\frac{1}{\ell}\sum_{i=1}^{\ell} g_\kappa(\mathbf{x}_i, \mathcal{X})$, which is an upper bound on the false alarm rate, the sample version of FAP. Note that the detector is constructed and evaluated using the same set of observations $\mathcal{X}$. A bound as such is usually called a *training set bound* [30]. Next we show a *test set bound* where

the detector is built upon a collection of observations, called *a training data set*, and evaluated on a different collection of observations called *a test set*.

Theorem 2 *Let* $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\ell_{train}}\} \subset \mathbb{R}^n$ *and* $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{\ell_{test}}\} \subset \mathbb{R}^n$ *be i.i.d. samples from a distribution* F *on* $\mathbb{R}^n$*. Let* $g_\kappa(\mathbf{x}, \mathcal{X})$ *be an outlier detector defined in (5). Fix* $\delta \in (0, 1)$*. For a new random observation* $\mathbf{x}$ *from cdf* F*, the following bound holds with probability at least* $1 - \delta$*:*

$$\mathbb{E}_F\left[g_\kappa(\mathbf{x}, \mathcal{X})\right] \leq \frac{1}{\ell_{test}} \sum_{i=1}^{\ell_{test}} g_\kappa(\mathbf{y}_i, \mathcal{X}) + \sqrt{\frac{\ln 1/\delta}{2\ell_{test}}}\,. \tag{7}$$

It is not difficult to validate that $\frac{1}{\ell_{test}} \sum_{i=1}^{\ell_{test}} g_\kappa(\mathbf{y}_i, \mathcal{X})$ monotonically decreases when b approaches 0. Hence for a fixed threshold t, the test set bound is the tightest at $b = 0$ (recall that $E_F[g_\kappa(\mathbf{x}, \mathcal{X})] = P_{FA}(g_\kappa)$ at $b = 0$). In this scenario, the FAP is bounded by the false alarm rate, evaluated on the test set, plus a term that shrinks in a rate proportional to the square root of the size of the test set. This suggests that we can always set $b = 0$ if we apply the above test set bound to select an outlier detector. For a given desired FAP, we should choose the threshold to be the maximum value of t such that the right-hand side of (7) does not exceed the desired FAP.

The training set bound in (6) is usually looser than the above test set bound because of the $1/b$ factor. Moreover, unlike the test set bound, we cannot set b be 0 for the obvious reason. Hence we have to do a search on both b and t to choose an "optimal" outlier detector, the one with the largest t that gives an upper bound on the FAP no greater than the desired level. As a result, the test set bound is usually preferred when the number of observations is large so that it is possible to have enough observations in both the training set and test set. One the other hand, we argue that the training set bound is more useful for small sample size, under which both bounds will be loose. Therefore, it is more desirable to build the outlier detector upon all available observations instead of sacrificing a portion of the precious observations on the test set. In this scenario, the relative, rather than the absolute, value of the bounds can be used to select the t parameter of an outlier detector.

4.2 The Missing Label Problem

For a missing label problem, all observations are unlabeled, or, put it equivalently, they come from a mixture of F_{good} and $F_{outlier}$, i.e., $F = (1-\alpha)F_{good} + \alpha F_{outlier}$ for some $\alpha \in [0, 1]$. Consequently, the above training set and test

set bounds cannot be directly applied to select detectors because $P_{FA}(g_\kappa)$ could be greater than $\mathbb{E}_F[g_\kappa(\mathbf{x}, \mathcal{X})]$ – an upper bound on $\mathbb{E}_F[g_\kappa(\mathbf{x}, \mathcal{X})]$ does not imply an upper bound on the FAP.

Fortunately, the results of Theorem 1 and Theorem 2 can be extended to the missing label problem under a mild assumption, namely, the prior probability α for outliers does not exceed a given number $r \in [0, 1]$. In other words, $\alpha \leq r$ means that the probability of a randomly chosen observation being an outlier is not greater than r. Since outliers are typically rare in almost all applications that outliers are sought, quantifying the rareness via an upper bound on α is actually not a restrictive but a defining presumption.

Theorem 3 *Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\} \subset \mathbb{R}^n$ be i.i.d. samples from a mixture distribution*

$$F = (1 - \alpha)F_{good} + \alpha F_{outlier}, \quad \alpha \in [0, 1],$$

on $\mathbb{R}^n$. Let $g_\kappa(\mathbf{x}, \mathcal{X})$ be an outlier detector defined in (5). Suppose that $\alpha \leq r$ for some $r \in [0, 1]$. Then

$$\mathbb{E}_{F_{good}}[g_\kappa(\mathbf{x}, \mathcal{X})] \leq \frac{1}{1 - r}\mathbb{E}_F[g_\kappa(\mathbf{x}, \mathcal{X})] \; . \tag{8}$$

A proof of Theorem 3 is given in the Appendix.

Based on (8), the bounds on FAP for the one-class learning problem can be extended to the missing label problem: the training set bound (6) is of the form

$$P_{FA}(g_\kappa) \leq \frac{1}{1 - r}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} g_\kappa(\mathbf{x}_i, \mathcal{X}) + \frac{2}{\ell b} + \left(1 + \sqrt{1 + \left(1 + \frac{4}{b}\right)^2}\right)\sqrt{\frac{\ln(2/\delta)}{2\ell}}\right] \; ,$$

and the test set bound (7) is of the form

$$P_{FA}(g_\kappa) \leq \frac{1}{1 - r}\left[\frac{1}{\ell_{test}}\sum_{i=1}^{\ell_{test}} g_\kappa(\mathbf{y}_i, \mathcal{X}) + \sqrt{\frac{\ln 1/\delta}{2\ell_{test}}}\right] \; . \tag{9}$$

If r is small, $1/(1 - r) \approx 1$. This suggests that the bounds for the missing label problem are only slightly larger than those for the one-class learning problem for small r.

5 Experimental Results

In the first experiment, we test kernelized spatial depth outlier detection on a synthetic data set. Next we compare the performance of the proposed

method with that of three well-established outlier detection algorithms, the LOF [27], the feature bagging [31], and the active learning [1]. In all the experiments, the KSD is computed using the Gaussian kernel with the σ parameter being determined from the following procedure.

Algorithm 1 Deciding σ for Gaussian Kernel

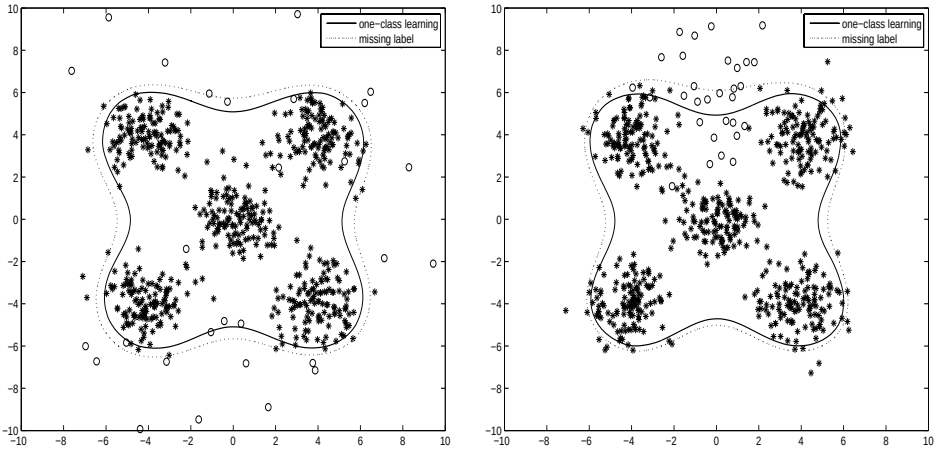
```
1 FOR (every observation x_i in X)
2    d_i = min_{j=1,...,ℓ,j≠i} ||x_i − x_j||
3 END
4 OUTPUT (σ = median(d_1, d_2, ..., d_ℓ))
```

5.1 Synthetic Data

For the synthetic data, we assume that $F_{outlier}$ is uniform over the region $[-9, 9] \times [-9, 9]$, and F_{good} is a mixture of five 2-dimensional Gaussian distributions (with equal weights): $N_1 \sim N([0, 0]^T, I)$, $N_2 \sim N([4, 4]^T, I)$, $N_3 \sim N([-4, 4]^T, I)$, $N_4 \sim N([-4, -4]^T, I)$, and $N_5 \sim N([4, -4]^T, I)$, where $N(\boldsymbol{\mu}, \Sigma)$ denotes Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

We first simulate an outlier detector in a one-class learning scenario. A training set of 500 i.i.d. observations and a validation set of 500 i.i.d. observations are generated from F_{good}. The KSD function is constructed based on the 500 training observations. In order to control the FAP under 0.1, we select the threshold t of the detector based on the test set bound (7) with $\delta = 0.05$ such that t is chosen to be the maximum value subject to the condition that the right-hand side of (7) does not exceed 0.1. Note that the validation set here is the test set in (7). It turns out that $t = 0.29160165$ is the desired threshold, i.e., all observations with depth vale less than 0.29160165 are identified as outliers. This threshold will, with probability at least 0.95, keep the FAP less than 0.1. We apply the detector to a test set of 525 i.i.d. observations, among which 500 are generated from F_{good} and the remaining 25 from $F_{outlier}$. Figure 5.a shows all 525 test observations superimposed with the contour of the KSD at value t. The $*$'s and $\circ$'s represent observations from F_{good} and $F_{outlier}$, respectively. The regions enclosed by the contour have KSD values greater than t. For the test set, the false alarm rate of our detector is 0.052, which is below the required 0.1 false alarm probability. At the same time, the detector identifies 17 out of the 25 outliers. Hence the detection rate is 0.68.

Next, we simulate the missing label scenario. Each of the training and validation set contains 500 i.i.d. observations generated from $F = (1 -$



(a) One-class learning scenario (b) Missing label scenario

Figure 5: Decision boundaries of the proposed outlier detectors in (a) one-class learning scenario and (b) missing label scenario, based on 525 i.i.d. test observations in which 500 (marked with $*$'s) were generated from F_{good} and 25 (marked with $\circ$'s) from $F_{outlier}$. Here F_{good} is a mixture of 5 Gaussian distributions and $F_{outlier}$ is uniform over $[-9, 9] \times [-9, 9]$. The boundaries are chosen such that the upper bound on the false alarm probability is 0.1. Observations falling outside the boundaries are classified as outliers. The false alarm rate and the detection rate on the test set are: (a) 0.052 and 0.68; (b) 0.026 and 0.56.

$\alpha)F_{good} + \alpha F_{outlier}$ where $\alpha = 0.05$. The kernelized spatial depth function is built upon the training set. The threshold t of the detector is determined based on the validation set and the inequality (9) where $\delta = 0.05$; the target false alarm rate is less than 0.1; and $\alpha \leq r = 0.05$. It turns out that $t = 0.29164708$ is the desired threshold. We apply the detector to the same test set as in the above one-class learning scenario. Figure 5.b shows all 525 observations and the contour of KSD at the selected threshold. The false alarm rate of this detector is 0.026 which is much below 0.1, the required level of false alarm probability. The detector identifies 14 out of the 25 outliers. The detection rate is therefore 0.56. Compared with the one-class learning setting, the detection rate is lower in the missing label case. This is because we need to be more conservative in selecting the threshold, which leads to a smaller false alarm rate and a smaller detection rate.

Table 1: Performance comparison of KSD, LOF, feature bagging (FB), and active learning (AL) outlier detection methods. The area under the ROC curve (AUC) for each method and each data set is shown. KSD1 and KSD2 refer to the one-class learning and missing label scenarios, respectively. A larger AUC value (closer to 1) indicates better performance.

Data Set	Ann-Thyroid 1	Ann-Thyroid 2	Shuttle	KDD-Cup'99
Outlier Class	Class 1	Class 2	Class 2, 3, 5–7	U2R
|Data Set|	3428	3428	14500	60839
KSD1	0.9725 ± 0.008	0.8074 ± 0.007	0.9969 ± 0.001	0.9403 ± 0.006
KSD2	0.9381 ± 0.010	0.7287 ± 0.010	0.7754 ± 0.032	0.8884 ± 0.044
LOF	0.869	0.761	0.825	0.61 ± 0.1
FB	0.869	0.769	0.839	0.74 ± 0.1
AL	0.97 ± 0.01	0.89 ± 0.11	0.999 ± 0.0006	0.935 ± 0.04

5.2 Comparison with Other Approaches

We compare the performance of the proposed approach with three existing outlier detection algorithms: the well-known LOF method [27], the recent feature bagging method [31], and the most recent active learning outlier detection method [1]. The data sets we used for the comparison include two versions of Ann-Thyroid, the Shuttle data, and the KDD-Cup 1999 intrusion detection data. Ann-Thyroid and Shuttle data sets are available from the UCI Machine Learning Repository. The KDD-Cup 1999 data set is available at the UCI KDD Archive. To be consistent with the experimental set-up in [31] and [1], one of the rare classes is chosen as the outlier class in our experiment. The outlier classes are listed in Table 1. In [31], the smallest intrusion class, U2R, was chosen as the outlier class. We found that the outlier class in [31] actually contains several other types of attacks including ftp_write, imap, multihop, nmap, phf, pod, and teardrop. The number of outliers is 246.

Each data set is randomly divided into a training set and a test set. Approximately half of the observations in Thyroid data sets are selected as training data. For the Shuttle and KDD-Cup 1999 data sets, the training set contains 500 randomly chosen observations and the test set has the remaining $14,000$ and $60,339$ observations. In the one-class learning scenario, the outliers in the training set are excluded from the construction of the KSD function, while in the missing label scenario, the KSD function is built on all observations in the training set. As in [31] and [1], we use the area under the ROC curve (AUC) as the performance metric. The average AUC over 10 random splits are reported for the proposed approach in Table 1

along with the standard deviation. The AUC values of the LOF, the feature bagging, and the active learning methods are obtained from [31] and [1]. The standard deviations are included when they are available.

As expected, the performance of the proposed approach degrades when the outliers are included in the construction of the KSD function, i.e., in the missing label scenario. Both LOF and feature bagging were evaluated under the one-class learning scenario where detectors were built from normal observations. From Table 1, it is clear that the KSD based outlier detection (one-class learning) consistently outperforms the LOF and the feature bagging methods on all four data sets. The performance of the proposed approach is comparable with that of the active learning outlier detection on all four data sets. The mean AUC of the proposed approach is slightly lower than that of the active learning on Ann-Thyroid 2. However, the variance is significantly smaller (by one order of magnitude). Hence the difference of mean ACU on Ann-Thyroid2 is not statistically significant. The active learning outlier detection transforms outlier detection to a binary classification problem using artificially generated observations that play the role of potential outliers. As pointed out by the authors of [1], the choice of the distribution of synthetic observations is domain dependent. In contrast, no prior knowledge on the distribution of outliers is required by the KSD outlier detection.

6 Conclusions

In this chapter, we presented a statistical depth function, the kernelized spatial depth (KSD), and an outlier detection method using the KSD function. The KSD is a generalization of the spatial depth [52, 10, 56]. It defines a depth function in a feature space induced by a positive definite kernel. The KSD of any observation can be evaluated using a given set of samples. The depth value is always within the interval $[0, 1]$, and decreases as a data point moves away from the center, the spatial median, of the data cloud. This motivates a simple outlier detection algorithm that identifies an observation as an outlier if its KSD value is smaller than a threshold. We derived the probabilistic inequalities for the false alarm probability of an outlier detector. These inequalities can be applied to determine the threshold of an outlier detector, i.e., the threshold is chosen to control the upper bound on the false alarm probability under a given level. We evaluated the proposed outlier detection algorithm over synthetic data sets and real life data sets. In comparison with other methods, the KSD based outlier detection demonstrates

competitive performance on all data sets tested.

Acknowledgments

Yixin Chen is supported by the University of Mississippi. Xin Dang and Hanxiang Peng are supported by the US National Science Foundation under Grant DMS No. 0707074.

References

[1] N. Abe, B. Zadrozny, and J. Langford, "Outlier Detection by Active Learning," *Proc. 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 504–509, 2006.

[2] C. C. Aggarwal and P. S. Yu, "Outlier Detection for High Dimensional Data," *Proc. 2001 ACM SIGMOD Int'l Conf. on Management of Data*, pp. 37–46, 2001.

[3] F. Angiulli, S. Basta, C. Pizzuti, "Distance-Based Detection and Prediction of Outliers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 2, pp. 145–160, 2006.

[4] A. C. Atkinson, "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, vol. 89, pp. 1329–1339, 1994.

[5] A. Banerjee, P. Burlina, and C. Diehl, "A Support Vector Method for Anomaly Detection in Hyperspectral Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 8, pp. 2282–2291, 2006.

[6] V. Barnett and T. Lewis, *Outliers in Statistical Data*, John Wiley and Sons, 1994.

[7] N. Billor, A. Hadi and P. Velleman, "BACON: Blocked Adaptive Computationally-Efficient Outlier Nominators," *Computational Statistics & Data Analysis*, vol. 34, pp. 279–298, 2000.

[8] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," *Proc. 2000 ACM SIGMOD Int'l Conf. on Management of Data*, pp. 93–104, 2000.

[9] P. Chaudhuri, "Multivariate Location Estimation Using Extension of R-Estimates Through U-Statistics Type Approach," The Annals of Statistics, vol. 20, no. 2, pp. 897–916, 1992.

[10] P. Chaudhuri, "On a Geometric Notion of Quantiles for Multivariate Data," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 862–872, 1996.

[11] Y. Chen, X. Dang, H. Peng, and H. L. Bart, "Outlier Detection with the Kernelized Spatial Depth Function," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, 18 pages, 2009.

[12] C. Campbell and K. P. Bennett, "A Linear Programming Approach to Novelty Detection," *Advances in Neural Information Processing Systems 13*, pp. 395–401, 2001.

[13] H. Caussinus and A. Ruiz-Gazen, "Projection Pursuit and Generalized Principal Component Analysis," In *New Directions in Statistical Data Analysis and Robusteness*, pp. 35–46, Basel: Birkhäuser Varlag, 1993.

[14] E. Eskin, "Anomaly Detection over Noisy Data Using Learned Probability Distributions," *Proc. 17th Int'l Conf. on Machine Learning*, pp. 255–262, 2000.

[15] A. B. Frakt, W. C. Karl, and A. S. Willsky, "A Multiscale Hypothesis Testing Approach to Anomaly Detection and Localization from Noisy Tomographic Data," *IEEE Transactions on Image Processing*, vol. 7, no. 6, pp. 825–837, 1998.

[16] M. G. Genton, "Classes of Kernels for Machine Learning: A Statistics Perspective," *Journal of Machine Learning Research*, vol. 2, pp. 299–312, 2001.

[17] R. Gnanadesikan and J. R. Kettenring, "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," *Biometrics*, vol. 28, pp. 81–124, 1972.

[18] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd edition, Addison-Wesley, 1992.

[19] A. S. Hadi, "Identification Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society, Ser. B*, vol. 54, pp. 761–771, 1972.

[20] H. Hajji, "Statistical Analysis of Network Traffic for Adaptive Faults Detection," *IEEE Transactions on Neural Networks*, vol. 16, no. 5, pp. 1053–1063, 2005.

[21] S.-J. Han and S.-B. Cho, "Evolutionary Neural Networks for Anomaly Detection Based on the Behavior of a Program," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 36, no. 3, pp. 559–570, 2006.

[22] D. W. Hawkins, "The Detection of Errors in Multivariate Data Using Principal Components," *Journal of the American Statistical Association*, **69**, pp. 340-344, 1974.

[23] D. M. Hawkins, *Identification of Outliers*, Chapman and Hall, 1980.

[24] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan and S. Maybank, "A System for Learning Statistical Motion Patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1450–1464, 2006.

[25] J. Hugg, E. Rafalin, K. Seyboth, and D. Souvaine, "An Experimental Study of Old and New Depth Measures," *Workshop on Algorithm Engineering and Experiments (ALENEX06)*, pp. 51–64, 2006.

[26] E. Keogh, J. Lin, A. W. Fu, and H. Van Herle, "Finding Unusual Medical Time-Series Subsequences: Algorithms and Applications," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 3, pp. 429–439, 2006.

[27] E. M. Knorr and R. T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Datasets," *Proc. 24th Int'l Conf. on Very Large Data Bases*, pp. 392–403, 1998.

[28] G. Kollios, D. Gunopulos, N. Koudas, and S. Berchtold, "Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Data Sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 5, pp. 1170–1187, 2003.

[29] H. Kwon and N. M. Nasrabadi, "Kernel RX-Algorithm: A Nonlinear Anomaly Detector for Hyperspectral Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 2, pp. 388–397, 2005.

[30] J. Langford, "Tutorial on Practical Prediction Theory for Classification," *Journal of Machine Learning Research*, vol. 6, pp. 273–306, 2005.

[31] A. Lazarevic and V. Kumar, "Feature Bagging for Outlier Detection," *Proc. 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 157–166, 2005.

[32] C. Manikopoulos and S. Papavassiliou, "Network Intrusion and Fault Detection: A Statistical Anomaly Approach," *IEEE Communications Magazine*, vol. 40, no. 10, pp. 76–83, 2002.

[33] M. Markou and S. Singh, "Novelty Detection: a Review–Part 1: Statistical Approaches," *Signal Processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[34] M. Markou and S. Singh, "Novelty Detection: a Review–Part 2: Neural Network based Approaches," *Signal Processing*, vol. 83, no. 12, pp. 2499–2521, 2003.

[35] R. A. Maronna, "Robust M-estimators of Multivariate Location and Scatter," *Annals of Statistics*, vol. 4, pp. 51–67, 1976.

[36] D. J. Miller and J. Browning, "A Mixture Model and EM-Based Algorithm for Class Discovery, Robust Classification, and Outlier Rejection in Mixed Labeled/Unlabeled Data Sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1468–1483, 2003.

[37] L. Parra, G. Deco, and S. Miesbach, "Statistical Independence and Novelty Detection with Information Preserving Non-Linear Maps," *Neural Computation*, vol. 8, no. 2, pp. 260–269, 1996.

[38] D. Peña and F. Prieto, "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," *Technometrics*, vol. 43, pp. 286–300, 2001.

[39] F. Preparata and M. Shamos, *Computational Geometry: An Introduction*, Springer-Verlag, 1988.

[40] S. Ramaswamy, R. Rastogi, and S. Kyuseok, "Efficient Algorithms for Mining Outliers from Large Data Sets," *Proc. of 2000 ACM SIGMOD Int'l Conf. on Management of Data*, pp. 427–438, 2000.

[41] C. R. Rao, "The Use and Interpretation of Principal Component Analysis in Applied Research," *Sankhya A*, vol. 26, pp. 329–358, 1964.

[42] G. Rätsch, S. Mika, B. Schölkopf, and K.-R. Müller, "Constructing Boosting Algorithms from SVMs: An Application to One-Class Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1184–1199, 2002.

[43] I. S. Reed and X. Yu, "Adaptive Multiple-Band CFAR Detection of an Optical Pattern with Unknown Spectral Distribution," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 10, pp. 1760–1770, 1990.

[44] S. Roberts and L. Tarassenko, "A Probabilistic Resource Allocating Network for Novelty Detection," *Neural Computation*, vol. 6, no. 2, pp. 270–284, 1994.

[45] D. M. Rocke and D. L. Woodruff, "Idenfication of Outliers in Multivariate Data," *Journal of the American Statistical Association*, vol. 91, pp. 1047–1061, 1996.

[46] P. J. Rousseeuw and K. van Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, vol. 41, pp. 212–223, 1999.

[47] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, New York: Wiley, 2003.

[48] D. Ruppert, " Computiing S-estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics*, vol. 1, pp. 253–270, 1992.

[49] I. Ruts and P. Rousseeuw, "Computing Depth Contours of Bivariate Point Clouds," *Computational Statistics and Data Analysis*, vol. 23, no. 1, pp. 153–168, 1996.

[50] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[51] S. M. Schweizer and J. M. F. Moura, "Hyperspectral Imagery: Clutter Adaptation in Anomaly Detection," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1855–1871, 2000.

[52] R. Serfling, "A Depth Function and a Scale Curve Based on Spatial Quantiles," In *Statistical Data Analysis Based on the L1-Norm and Related Methods* (Y. Dodge, ed.), pp. 25–38, 2002.

[53] I. Steinwart, D. Hush, and C. Scovel, "A Classification Framework for Anomaly Detection," *Journal of Machine Learning Research*, vol. 6, pp. 211–232, 2005.

[54] P. Sun and S. Chawla, "On Local Spatial Outliers," *Proc. 4th IEEE Int'l Conf. on Data Mining*, pp. 209–216, 2004.

[55] J. Tang, Z. Chen and A. W.-C. Fu, and D. Cheung, "A Robust Outlier Detection Scheme in Large Data Sets," *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, LNCS 2336, pp. 535–548, 2002.

[56] Y. Vardi and C.-H. Zhang, "The Multivariate L_1-Median and Associated Data Depth," *Proceedings of the National Academy of Sciences*, vol. 97, no. 4, pp. 1423–1436, 2000.

[57] A. Weber, *Theory of the Location of Industries* (translated by C. J. Friedrich from Weber's 1909 book), Chicago: The University of Chicago Press, 1929.

[58] W. Zhou and R. Serfling, "Multivariate Spatial U-quantiles: a Bahadur-Kiefer Representation, a Theil-Sen Estimator for Multiple Regression, and a Robust Dispersion Estimator," *Journal of Statistical Planning and Inference*, vol. 138, pp. 1660–1678, 2008.