Outlier Detection with the Kernelized Spatial Depth Function

Yixin Chen, Xin Dang, Hanxiang Peng, Henry L. Bart, Jr.

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 and Hanxiang Peng are with Department of Mathematics, The University of Mississippi, University, MS 38677, USA. E-mail: {xdang,mmpeng}@olemiss.edu.

Henry L. Bart, Jr. is with Tulane University Museum of Natural History, Belle Chasse, LA 70037, USA, and the Department of Ecology and Evolutionary Biology, Tulane University, New Orleans, LA 70118, USA. E-mail: hank@museum.tulane.edu.

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. In this article, we propose a novel statistical depth, the *kernelized spatial depth* (KSD), which generalizes the spatial depth via *positive definite kernels*. By choosing a proper kernel, the KSD can capture the local structure of a data set while the spatial depth fails. We demonstrate this by the half-moon data and the ring-shaped data. Based on the KSD, we propose a novel *outlier detection algorithm*, by which an observation with a depth value less than a threshold is declared as an outlier. The proposed algorithm is simple in structure: the threshold is the only one parameter for a given kernel. It applies to a one-class learning setting, in which "normal" observations are given as the training data, as well as to a missing label scenario where the training set consists of a mixture of normal observations and outliers with unknown labels. We give upper bounds on the false alarm probability of a depth-based detector. These upper bounds can be used to determine the threshold. We perform extensive experiments on synthetic data and data sets from real applications. The proposed outlier detector is compared with existing methods. The KSD outlier detector demonstrates competitive performance.

Index Terms

Outlier detection, novelty detection, anomaly detection, statistical depth function, spatial depth, kernel method, unsupervised learning.

I. INTRODUCTION

In a variety of applications, e.g., network security [18], [26], [42], [65], [71], visual surveillance [29], [66], remote sensing [6], [10], [36], medical diagnostics [33], [20], image processing [24], zoology and anthropology [76], and revisionary systematics [14], 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* (or *novelty detection*, or *anomaly detection*, or *fault detection*) [7].

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 an 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 samples 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.

Outlier detection has been investigated extensively over the last several decades by researchers from statistics, data mining, and machine learning communities. Next we review work most related to this article. For a more comprehensive survey of this subject, the reader is referred to Barnett and Lewis [7], Hawkins [28], and Markou and Singh [43], [44].

A. 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. [20], 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. Carlotto [10] presented a method to detect man-made objects (anomalies) in images. For the scenario under consideration where the occurrence of man-made objects is rare compared with that of the background clutters, it is assumed that the pixel values of a man-made object deviate significantly from those of the background, which is modeled by a mixture of Gaussian distributions. Reed and Yu [51] 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 [36] introduced a nonlinear version of Reed and Yu's algorithm using feature mappings induced by positive definite kernels. Kollios et al. [35] 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 [3], [4], [8], [34], [50], [68], [70]. Knorr and Ng [34] 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. [50] 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 [34] and [50], 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. [70] 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 [68] 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. [4] 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 [3] 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 [38] 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.

Outlier detection method based on Mahalanobis distance (MD) has been extensively studied in the statistics literature [56], [5], [54]. MD is affine invariant. It is robust if robust estimates of location and scatter matrix are used. A fast algorithm provided by Rousseeuw and Van Driessen [55] makes robust version MD-based methods feasible for large sample size data. As a missing label problem, outlier detection has also been studied as byproducts of robust statistical methods [11], [17], [19], [69]. Danuser and Stricker [17] presented a framework for generalized least squares fitting of multiple parametric models. For each fitted model, the data that support other models are viewed as outliers. Fidler et al. [19] proposed a classification algorithm, which is not sensitive to outliers, using a projection method developed on the basis of the robust dimensionality reduction technique described in [40]. Takeuchi and Yamanishi [69] explored outliers and change points detection in time series using an auto regression model. Castaño and Kunoth [11] applied a robust regression to the wavelet representation of 1- and 2-dimensional

data to estimate outliers.

B. 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 [53] 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 [60] 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 [46] 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. [47] 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. [59] 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) [73] from an optimization viewpoint, the method is called 1-class SVM. Along the line of 1-class SVM, Campbell and Bennett [9] 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. [49] developed a boosting algorithm for one-class classification based on connections between boosting and SVMs. Banerjee et al. [6] applied 1-class SVM for anomaly detection in hyperspectral images and demonstrated improved performance compared with the method described in [51].

There is an abundance of prior work that applies standard supervised learning techniques to tackle outlier detection [1], [27], [45], [67]. 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. In [45], a neural network-based novelty detector is trained based on normal observations and artificial novel examples generated by a uniform distribution. Han and Cho [27] 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. [67] provide an interesting justification for the above heuristic by converting outlier detection to a problem of finding level sets of data generating density.

C. An Overview of the Proposed Approach

In this paper, we propose a novel outlier detection framework based on the notion of *statistical depths*. Outlier detection methods that are based on statistical depths have been studied in statistics and computational geometry [48], [58], [16]. 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. For example, finding peeling and depth contours, in practice, require the computation of d-dimensional convex hulls [48], [58], for which the computational complexity is of magnitude $O(\ell^{d/2})$, where ℓ is the sample size and d is the dimension of an input space. The computational complexity for halfspace depth [72] and simplicial depth [41] is $O(\ell^{d-1}\log\ell)$ [57]; for projection depth [78], it is $O([\binom{2(d-1)}{d-1}/d]^2\ell^3)$ [22].

Of the various depths the *spatial depth* is especially appealing because of its computational efficiency and mathematical tractability [61]. Its computational complexity is of magnitude $O(\ell^2)$, independent of dimension d. Spatial depth has been applied in clustering and classification problems [31], [23]. 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. Consequently 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 develop an outlier detection framework that avoids the above limitation of spatial depth. The contributions of this paper are as follows.

- A new statistical depth function. We introduce a new depth function, *kernelized spatial depth* (KSD), which defines the spatial depth in a feature space induced by a positive definite kernel. By choosing a proper kernel, e.g., Gaussian kernel, the contours of a kernelized spatial depth function conform with the structure of the data set. Consequently the kernelized spatial depth can provide a local perspective of the data set.
- A simple outlier detection algorithm. The kernelized spatial depth of any observation can be evaluated directly from the data set with computational complexity $O(\ell^2)$. Observations with depth values less than certain threshold are declared as outliers. For a given kernel, the threshold on the depth value is the only parameter of the algorithm. We provide upper bounds on the false alarm probability of the detector, i.e., the probability of misclassifying a normal observation as an outlier. These upper bounds can be used to determine the threshold.
- Broad adaptability. The proposed framework applies to a one-class learning problem as well as to a missing label problem provided that an upper bound on the ratio of normal observations to outliers is given. Our extensive experimental results on artificial data and real applications demonstrate competitive performance of the proposed framework.

D. Outline of the Paper

The remainder of the paper is organized as follows. Section II motivates spatial depth-based outlier detection via the connection between spatial depth and L_1 median. Section III introduces kernelized spatial depth. Section IV 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. Section V provides an algorithmic view of the approach. We compare the proposed approach with density based outlier detection methods in Section VI. In Section VII, we explain the extensive experimental studies conducted and demonstrate the results. We conclude and discuss possible future work in Section VIII.

II. Medians, Spatial Depth, and Outlier Detection

As Barnett and Lewis described [7], *"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.

A. Medians

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 $[\mathbf{x}_{(\ell/2)}, \mathbf{x}_{((\ell+2)/2)}]$ can be defined to be the sample median. A convenient choice is the average $\frac{\mathbf{x}_{(\ell/2)}+\mathbf{x}_{((\ell+2)/2)}}{2}$. 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* [77] or *the unit vector* [12], 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 version (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* [77] or the L_1 *median* [75], [74]. We refer keen readers to [64] for a comprehensive review of a variety of *multidimensional medians*. Next we give another equivalent definition of the spatial median that motivates the depth-based outlier detection.

B. The Spatial Depth

The concept of spatial depth was formally introduced by Serfling [61] based on the notion of spatial quantiles proposed by Chaudhuri [13], while a similar concept, L_1 depth, was first described by Vardi and Zhang [74]. 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 1. In other words, the spatial median is a set of data points that have the "deepest" depth 1. Indeed, the spatial depth provides from the "deepest" point a "center-outward" ordering of multidimensional data. The depth attains the maximum value 1 at the deepest point and decreases to zero as a point moves away from the deepest to the infinity. Thus it 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 refer to the sample depth.

C. Outlier Detection Using 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

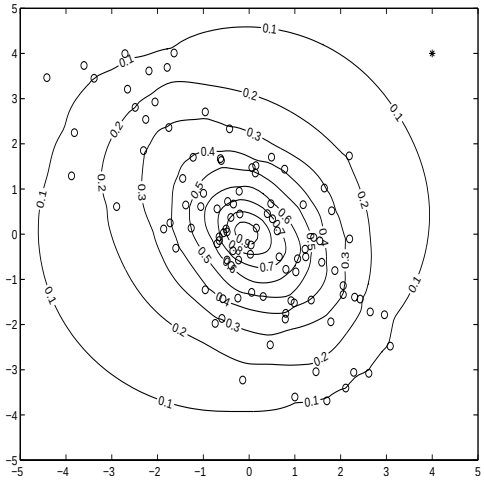


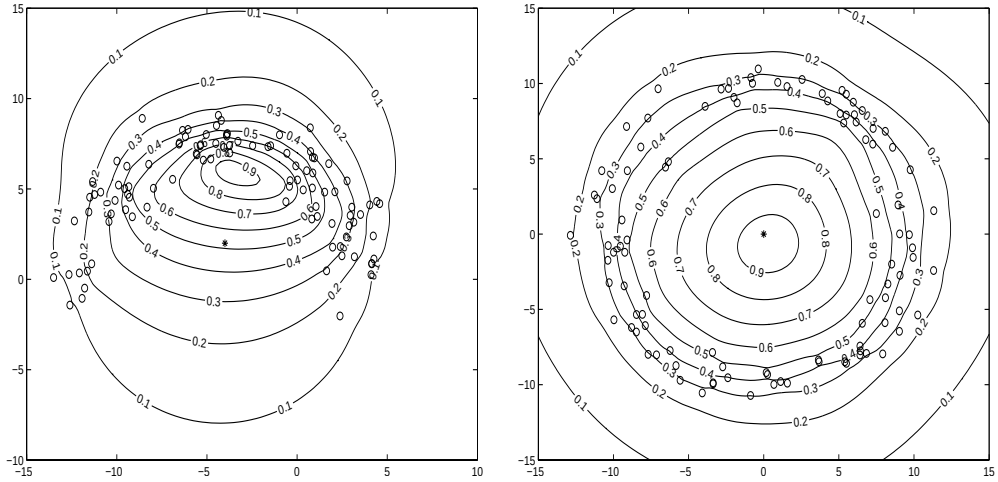
Fig. 1. A contour plot of the sample spatial depth based on 100 random observations (represented by ○'s) from a 2-dimensional Gaussian distribution. The depth values are indicated on the contours. A possible outlier is the observation (marked with ∗) on the upper left corner which has a very low depth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

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 a practical method, the following two issues need to be addressed:

1) How can we decide the threshold?
2) Can the spatial depth function capture the structure of the data cloud?

We postpone the discussion on the first question to Section IV where we present a framework to determine the threshold. The second question is related to the shape of depth contours. The depth contours of a spatial depth function tend to be circular [30], 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 view a data point as an outlier if its depth is low because a lower depth implies a larger distance from the "center" of the data cloud, which is defined by the spatial median. However, in general, the relationship between the depth and the outlyingness in a data cloud 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 half-moon shaped distribution (Figure 2.a) and a ring shaped distribution (Figure 2.b). From the shapes of the two distributions, it is reasonable to view the points (marked with $*$'s) in the center of both figures as outliers. However, the depth at the location of the $*$'s is 0.5155 for the half-moon data and 0.9544 for the ring data. A threshold larger than 0.5155 would classify more than 70% of the half-moon observations as outliers. For the ring data, all of the 100 observations have depth smaller than that of the "outlier" at the



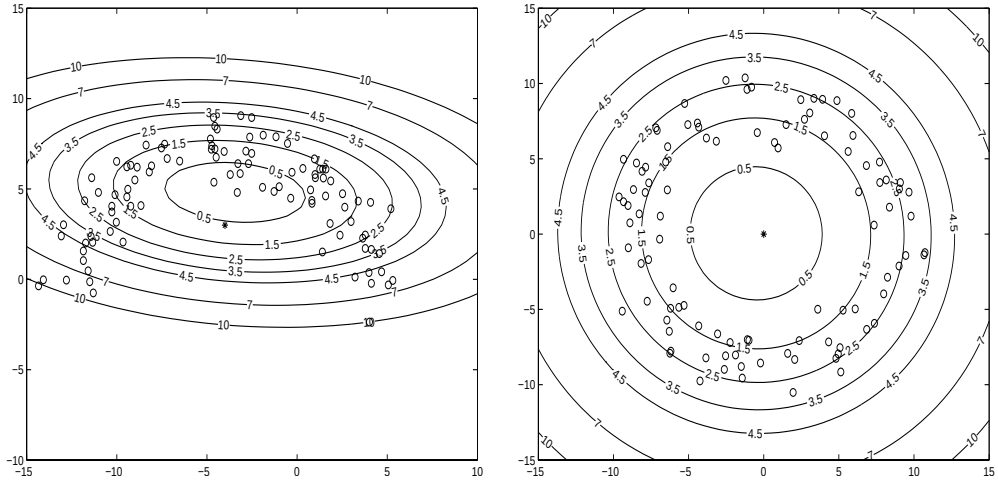
(a) Half-moon data (b) Ring data

Fig. 2. Contour plots of the sample spatial depths based on 100 random observations (denoted by ○'s) of (a) a half-moon shaped distribution and (b) a ring shaped distribution. The depth values are indicated on the contours. The observation (denoted by $*$) at the center of each plot represents a possible outlier. The depth values for the $*$ observations in (a) and (b) are 0.5155 and 0.9544, respectively.

center. Since Mahalanobis distance based outlier detection is a very traditional approach [56], [54], [55], we demonstrate the contours of Mahalanobis distance in Figure 3. [1] These contours are also constrained to be elliptical, which do not follow the shape of the distribution unless the underlying model is elliptically symmetric. Note that unlike the spatial depth-based outlier detection, a larger MD value indicates a higher likelihood of being an outlier.

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 sum of the unit vectors, each of which represents the direction from the point to an observation. This definition 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 some distance measurement, resulting in certain loss of the measurement of *similarity* of the data points. The distance of a point from the data cloud plays an important role in revealing the structure of the data cloud. In the following, we propose 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*.

[1]Robust minimum covariance determinant (MCD) estimator of multivariate location and covariance are calculated using the "mcdcov" function provided at http://www.wis.kuleuven.ac.be/stat/robust/libra.html.



(a) Half-moon data (b) Ring data

Fig. 3. Contour plots of Mahalanobis distance based on 100 random observations (denoted by $\circ$'s) of (a) a half-moon shaped distribution and (b) a ring shaped distribution. The MD values are indicated on the contours. The observation (denoted by $*$) at the center of each plot represents a possible outlier. The MD values for the $*$ observations in (a) and (b) are 0.6123 and 0.0741, respectively.

III. 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 [62]. 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[2], 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.

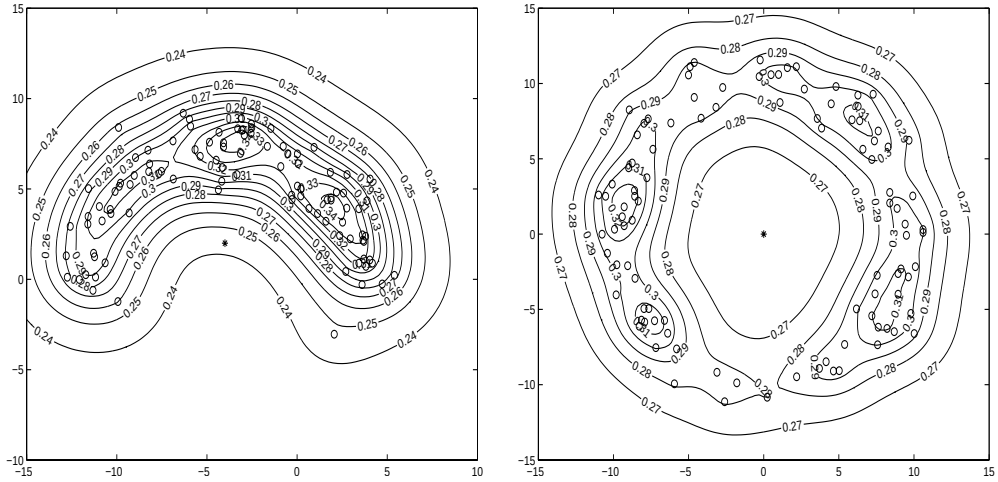
The basic idea of the *kernelized spatial depth* is to evaluate the spatial depth in a feature space induced by a positive definite kernel. Noticing that

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

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

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

[2]See [21] for a thorough discussion on stationary kernels along with other popular positive definite kernels.



(a) Half-moon data (b) Ring data

Fig. 4. Contour plots of KSD functions based on 100 random observations (marked with ∘'s) from (a) a half-moon 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 $*$ observations in (a) and (b) are 0.2495 and 0.2651 respectively.

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

$$D_\kappa(\mathbf{x},\mathcal{X}) = 1 - \frac{1}{|\mathcal{X}\cup\{\mathbf{x}\}|-1}\left(\sum_{\mathbf{y},\mathbf{z}\in\mathcal{X}} \frac{\kappa(\mathbf{x},\mathbf{x})+\kappa(\mathbf{y},\mathbf{z})-\kappa(\mathbf{x},\mathbf{y})-\kappa(\mathbf{x},\mathbf{z})}{\delta_\kappa(\mathbf{x},\mathbf{y})\delta_\kappa(\mathbf{x},\mathbf{z})}\right)^{1/2}, \mathbf{x}\in\mathbb{R}^n \quad (4)$$

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 $\mathbf{x}=\mathbf{z}$. Note that KSD is a spatial depth function in $\mathbb{F}$, but in general is no longer a depth function in $\mathbb{R}^n$ because its center in $\mathbb{F}$ does not necessarily have a preimage in $\mathbb{R}^n$. Even if we define a new center as the location in $\mathbb{R}^n$ that maximizes the KSD, the KSD value in general does not decrease monotonically for points moving away from the new center.

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 4 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 4.a) and the ring-shaped distribution (Figure 4.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 shape of the data clouds. Moreover, the depth values are small for the possible outliers.

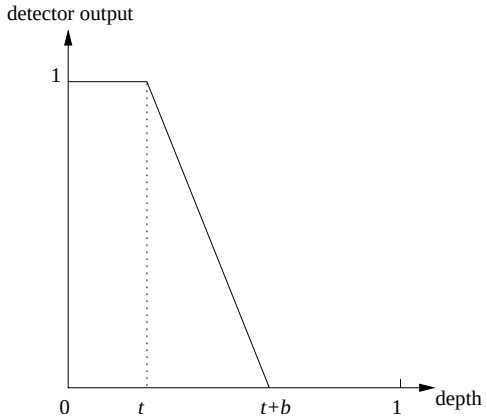


Fig. 5. A depth-based outlier detector. An output value of 1 indicates an outlier, i.e., an observation with depth smaller than t is classified as an outlier.

The depth values at the location of the $*$'s, 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.

IV. 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.

A. 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 5 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 $\mathbb{E}_{F|\mathcal{X}}$ the expectation calculated under cdf F for a given $\mathcal{X}$. We have the following theorem for the bound of the FAP.

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 $1-\delta$:

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

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 $P_{FA}(g_\kappa) \le \mathbb{E}_{F|\mathcal{X}}\left[g_\kappa(\mathbf{x}, \mathcal{X})\right]$ where the equality holds when $b = 0$. This suggests that (6) provides us an upper bound on the FAP. A proof of Theorem 1 is given in the Appendix.

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 (of $g_\kappa(\mathbf{x}, \mathcal{X})$, to be precise), 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* [37]. Next we derive 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 $1-\delta$:

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

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 $E_{F|\mathcal{X}}[g_\kappa(\mathbf{x}, \mathcal{X})] = P_{FA}(g_\kappa)$ 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. A proof of Theorem 2 is given in the Appendix.

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.

B. 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 $\mathbb{E}_{F|\mathcal{X}}[g_\kappa(\mathbf{x}, \mathcal{X})]$ – an upper bound on $\mathbb{E}_{F|\mathcal{X}}[g_\kappa(\mathbf{x}, \mathcal{X})]$ 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

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

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

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

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.

V. An Algorithmic View

We summarize the above discussion in pseudo code. The input is a collection of observations $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\} \in \mathbb{R}^n$, a kernel κ, and parameter t. These observations are generated by either F_{good} (in a one-class learning problem) or $(1-\alpha)F_{good} + \alpha F_{outlier}$ (in a missing label problem). Note that the threshold t is the key parameter in determining whether an observation is an outlier. The parameter b is needed only when the training set bound (6) is used to select t. The following pseudo codes determine whether an observation $\mathbf{x}$ is an outlier. In terms of the number of kernel evaluations and multiplications, the cost of computing the kernelized spatial depth for a given observation is $O(\ell^2)$.

Algorithm 1: Learning an Outlier Detector

```
FOR (every pair of observations x_i and x_j in 𝒳)
    K_ij = κ(x_i, x_j)
END
given input x
FOR (every observation x_i in 𝒳)
    ζ_i = κ(x, x_i)
    δ_i = sqrt(κ(x, x) + K_ii − 2ζ_i)
    IF δ_i = 0
        z_i = 0
```

10 ELSE

11 $z_i = \frac{1}{\delta_i}$

12 END

13 END

14 FOR (every pair of observations $\mathbf{x}_i$ and $\mathbf{x}_j$ in $\mathcal{X}$)

14 $\widetilde{K}_{ij} = \kappa(\mathbf{x}, \mathbf{x}) + K_{ij} - \zeta_i - \zeta_j$

15 END

16 $D_\kappa(\mathbf{x}, \mathcal{X}) = 1 - \frac{1}{|\mathcal{X} \cup \{\mathbf{x}\}| - 1} \sqrt{\mathbf{z}^T \widetilde{K} \mathbf{z}}$

17 OUTPUT ($\mathbf{x}$ is an outlier if $D_\kappa(\mathbf{x}, \mathcal{X}) \leq t$)

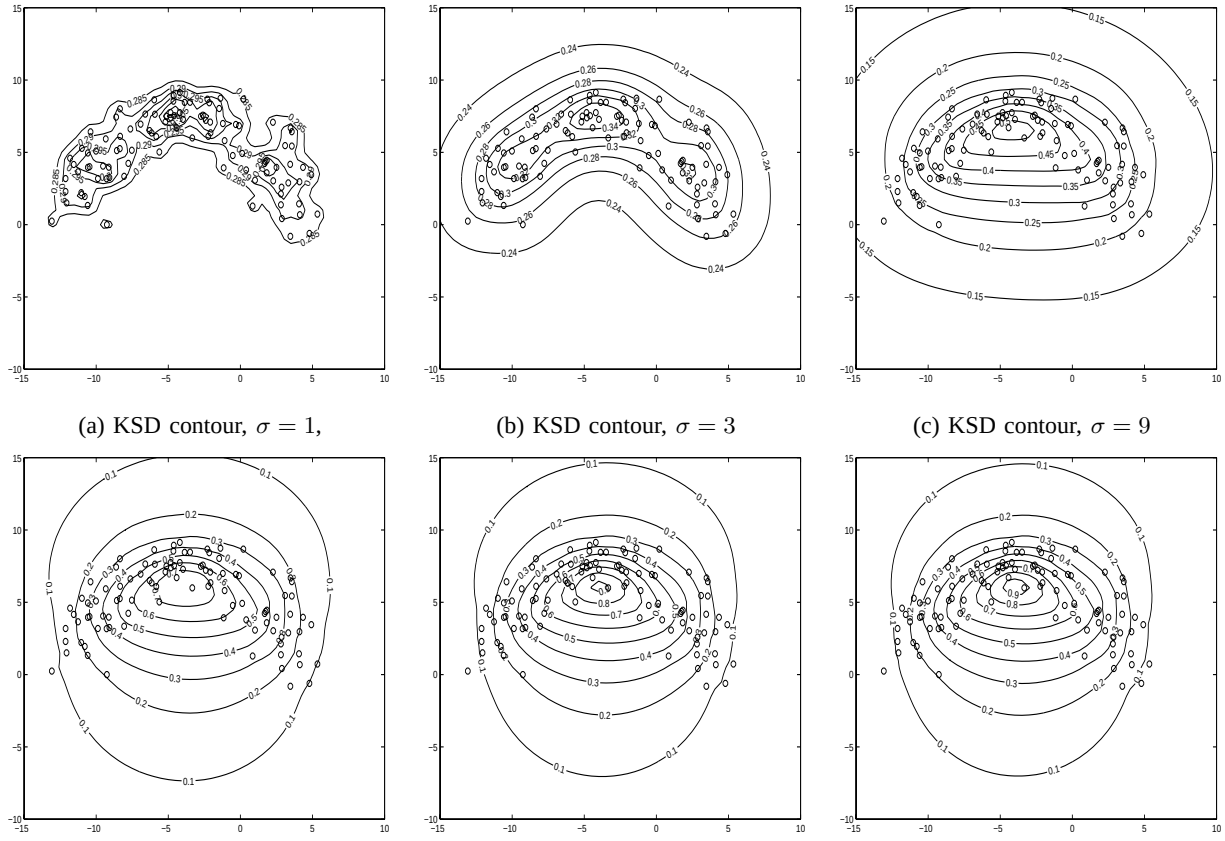
The above pseudo code assumes that the kernel κ is given. The choice of kernel is very important in every kernel method. For Gaussian kernel, which is used in our experimental study, σ determines the size of the neighborhood that is used to compute KSD for an observation. On one extremity, it can be proven that KSD converges to the spatial depth when σ goes to ∞. In this case, at any point $\mathbf{x}$, all observations in the data set contribute equally to the KSD value at $\mathbf{x}$ because each observation contributes a unit vector representing the direction from $\mathbf{x}$ to the observation. On the other extremity, when σ approaches 0, the KSD tends to the same constant depth value, $1 - \frac{\sqrt{2}}{2}$, for every point in the original feature space[3]. As this constant is independent of the observations in the data set, i.e., $D_{\sigma=0}(\mathbf{x}, \mathcal{X}) = 1 - \frac{\sqrt{2}}{2}$ for every $\mathbf{x} \in \mathbb{R}^n$ and every $\mathcal{X} \subset \mathbb{R}^n$, we can essentially view $\mathcal{X}$ as non-informative in defining KSD. In other words, none of the observations in the data set contributes to KSD when $\sigma = 0$. Figure 6 demonstrates the variation of the shape of KSD contours for the half-moon data with $\sigma = 1$, 3, 9, 27, and 81. For comparison, we also include the spatial depth contour in Figure 6(f). It is clear that the KSD contours approaches the spatial depth contour as σ increases.

The σ parameter determines the tradeoff between the global and local behaviors of KSD. A properly chosen σ should result in the contours of KSD following the geometric shape of the underlying model. We consider a generalized Gaussian kernel,

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-(\mathbf{x} - \mathbf{y})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})\right)$$

where $\mathbf{\Sigma} = \mathrm{Diag}[\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]$ is a diagonal matrix. We propose to choose the componentwise scale parameter σ_k in accordance with the dispersion of the data along the k-th dimension. Hence we suggest the following methods to estimate $\mathbf{\Sigma}$.

[3]For the uninteresting case where the data set contains only one observation, the value of KSD (and spatial depth) at that observation is by definition always 1 and 0 everywhere else.



(a) KSD contour, $\sigma = 1$, (b) KSD contour, $\sigma = 3$ (c) KSD contour, $\sigma = 9$

(d) KSD contour, $\sigma = 27$ (e) KSD contour, $\sigma = 81$ (f) spatial depth contour

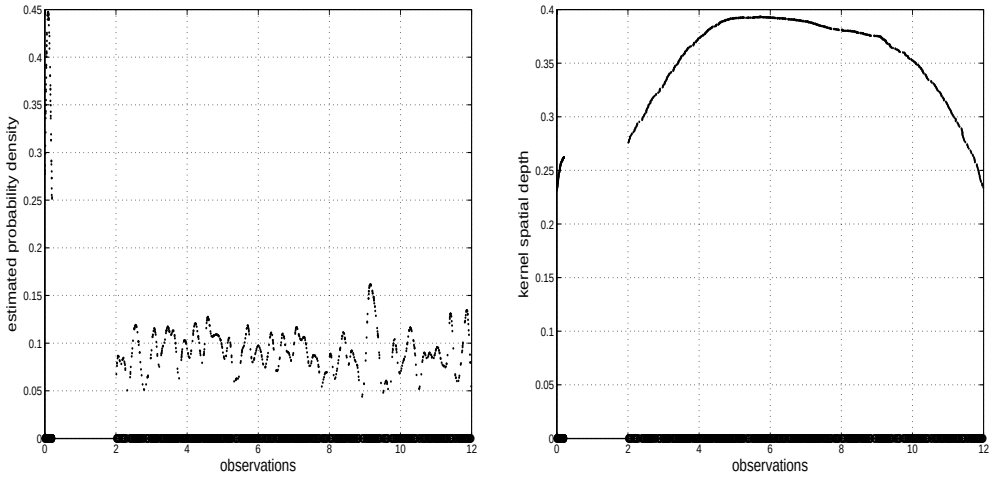
Fig. 6. (a)-(e): Contour plots of KSD functions with different values of σ based on 100 random observations (marked with $\circ$'s) from a half-moon distribution. (f) Contour plot of the spatial depth function.

- $\mathbf{\Sigma}_1$: $\sigma_k = \text{mean}_{i,j=1,\ldots,\ell} \, |x_{ik} - x_{jk}|$ where x_{ik} and x_{jk} represent the k-th component of the observation $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively.
- $\mathbf{\Sigma}_2$: $\sigma_k = \text{median}_{i,j=1,\ldots,\ell} \, |x_{ik} - x_{jk}|$.
- $\mathbf{\Sigma}_3$: $\sigma_k = \text{mean}_{i=1,\ldots,\ell} \, |x_{ik} - \text{mean}_{j=1,\ldots,\ell} \, x_{jk}|$.
- $\mathbf{\Sigma}_4$: $\sigma_k = \text{median}_{i=1,\ldots,\ell} \, |x_{ik} - \text{median}_{j=1,\ldots,\ell} \, x_{jk}|$.

The σ_k in $\mathbf{\Sigma}_1$ is the well-known mean difference also called Gini difference. It is less sensitive to outliers than the sample standard deviation. The σ_k in $\mathbf{\Sigma}_2$ is the more robust version of Gini difference by replacing mean by median. It is discussed in [15]. The σ_k in $\mathbf{\Sigma}_3$ and $\mathbf{\Sigma}_4$ are also robust dispersion estimates, commonly referred to as MAD (mean/median absolution deviation). In Section VII, we provide emprical results for all the above estimators.

VI. A Comparison of KSD and Density Based Outlier Detection

In the above discussion of KSD, we focus our choice of kernel on stationary kernels, in particular, the Gaussian kernel. Stationary kernels have been widely used in kernel density esti-



(a) Kernel density estimates (b) Kernelized spatial depth

Fig. 7. Kernel density estimates and kernelized spatial depth of 800 observations from a mixture of two uniform distributions: $0.1U_{[0,0.2]} + 0.9U_{[2,12]}$. Observations are marked along the horizontal axis.

mation and the related density-based outlier detection methods. Next, we discuss the distinctions between KSD- and density-based outlier detection.

- A KSD function is distinct from a density function.
  While a density describes a likelihood, the KSD measures the outlyingness of a point with respect to the whole population. A density has a range $[0, \infty)$; KSD has range $[0, 1]$.
- A sample KSD function is different from kernel density estimate.
  In kernel density estimation, the bandwidth parameter, (e.g., σ in Gaussian kernel) has to decrease to zero as the sample size increases to infinity in order to have consistency. In a sample KSD function, each diagonal element of $\mathbf{\Sigma}$ converges to the true dispersion of the data along the corresponding dimension, which is in general greater than 0. Moreover, KSD can be constructed from nonstationary kernels, such as the polynomial kernels, which cannot be used in density estimation.
- The underlying assumption of depth-based outlier detection approaches is different from that of density-based methods.
  Density-based outlier detection assumes that outliers mainly appear in low density regions. While in depth-based outlier detection, outliers are defined as those observations that are distant from the majority of the population (measured by depth values). Observations from a high density region may be separated from the majority of the population, which resides in a low density area. For one example, Figure 7 shows 800 observations generated by a distribution $F = 0.1U_{[0,0.2]} + 0.9U_{[2,12]}$ where $U_{[a,b]}$ denotes a uniform distribution over the interval $[a, b]$. Among the 800 observations, only around 80 are generated by $U_{[0,0.2]}$; the rest

of them are from $U_{[2,12]}$. However, the density function has a value of 0.5 on the interval $[0, 0.2]$ and 0.09 on the interval $[2, 10]$. Figure 7(a) shows the estimated probability density using Guassian kernel with $\sigma = 0.06$. Figure 7(b) shows the KSD with $\mathbf{\Sigma} = \mathbf{\Sigma}_3$ (other choices of $\mathbf{\Sigma}$ produce similar results). In this example, a density-based approach would classify all the observations from $U_{[2,12]}$ as outliers before it could identify any observation from $U_{[0,0.2]}$ as an outlier. In contrast, with a threshold 0.2632, KSD outlier detection would claim all observations from $U_{[0,0.2]}$ as outliers together with 24 observations that are in the right end of the interval $[2, 12]$.

VII. Experimental Results

We present systematic evaluations of the proposed outlier detector. In the first experiment, we test kernelized spatial depth outlier detection on several synthetic data sets. Next we apply the proposed outlier detection method to a problem in taxonomic research, *new species discovery*. Finally on several real life data sets we compare the performance of the proposed method with that of three well-established outlier detection algorithms, the LOF [34], the feature bagging [38], and the active learning [1].

A. Synthetic Data

For the synthetic data, we consider the following four models.

- Synthetic 1: $F_{outlier}$ is uniform over the region $[-10, 10] \times [-10, 10]$. 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.
- Synthetic 2: $F_{outlier}$ is a 2-dimensional Gaussian distribution, $N([0, 6]^T, 4I)$. F_{good} is identical to that in Synthetic 1.
- Synthetic 3: $F_{outlier}$ is identical to that of Synthetic 1. F_{good} is a mixture of three Gaussian distributions (with equal weights): $N_1 \sim N\left([-3, 1]^T, \begin{bmatrix} 1.750 & -1.299 \\ -1.299 & 3.250 \end{bmatrix}\right)$, $N_2 \sim N\left([4, -1]^T, \begin{bmatrix} 3.938 & 2.923 \\ 2.923 & 7.313 \end{bmatrix}\right)$, $N_3 \sim N\left([-6, -4]^T, \begin{bmatrix} 0.293 & 0.117 \\ 0.117 & 0.158 \end{bmatrix}\right)$.
- Synthetic 4: $F_{outlier}$ is identical to that of Synthetic 2. F_{good} is identical to that of Synthetic 3.

For each synthetic data, we first simulate the one-class learning scenario. A training set and a validation set, each consists of 600 i.i.d. observations, are generated from F_{good}. The KSD

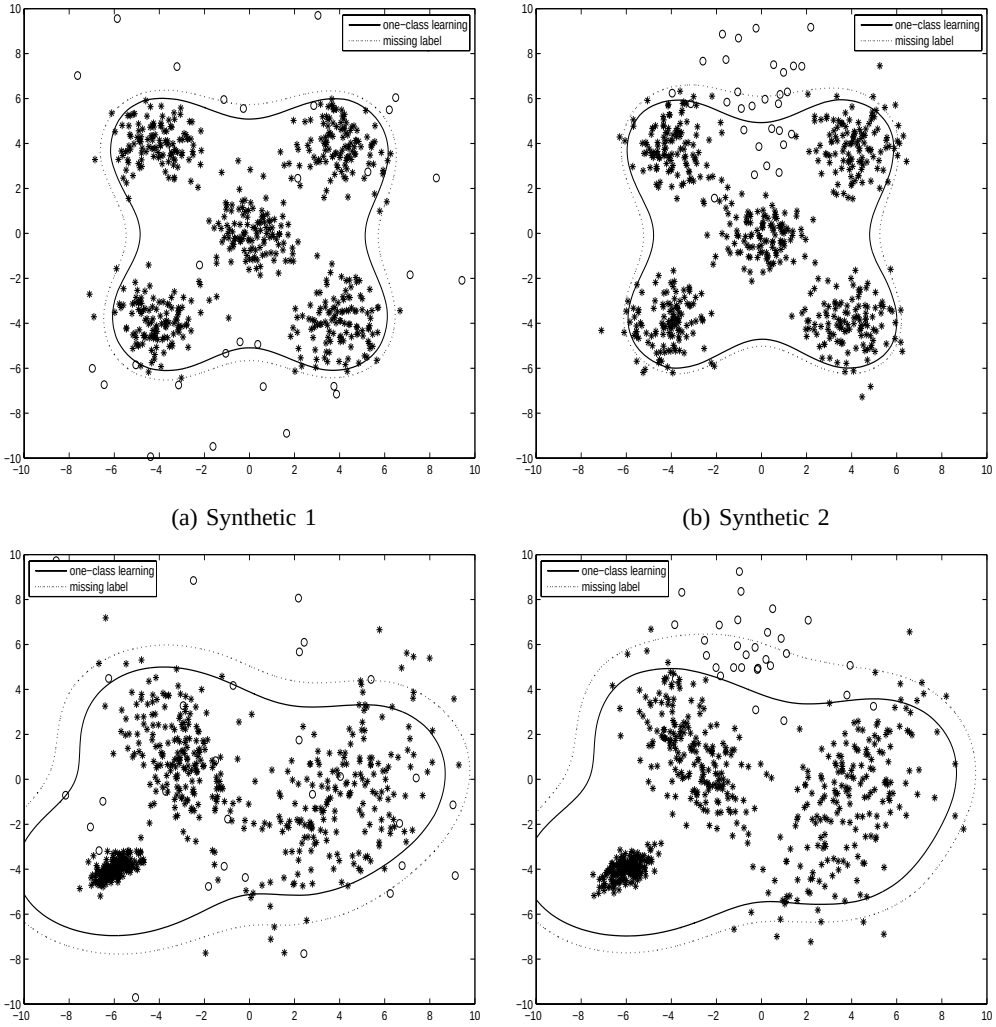
TABLE I

THRESHOLD t, FALSE ALARM RATE, AND DETECTION RATE UNDER ONE-CLASS LEARNING AND MISSING LABEL SCENARIOS. FALSE ALARM RATE IS THE PERCENTAGE OF NORMAL SAMPLES IN THE TEST SET THAT ARE MISCLASSIFIED AS OUTLIERS. DETECTION RATE IS THE PERCENTAGE OF OUTLIERS IN THE TEST SET THAT ARE IDENTIFIED CORRECTLY.

	One-class Learning			Missing Label		
	Threshold t	False Alarm Rate	Detection Rate	Threshold t	False Alarm Rate	Detection Rate
Synthetic 1	0.275	0.027	0.767	0.269	0.013	0.667
Synthetic 2	0.280	0.068	0.633	0.274	0.035	0.433
Synthetic 3	0.242	0.048	0.467	0.234	0.018	0.333
Synthetic 4	0.241	0.042	0.867	0.234	0.017	0.367

function is constructed based on the 600 training observations using Gaussian kernel with $\mathbf{\Sigma} = \mathbf{\Sigma}_2$. We suppose that FAP should be controlled under 0.1. To achieve this, we apply the test set bound (7) with $\delta = 0.05$ to select the threshold t, i.e., t is chosen such that with probability at least 0.95 FAP is less than 0.1. Specifically, we search for the maximum value of t that makes the false alarm rate, evaluated from the validation set, no greater than $0.1 - \sqrt{\frac{\ln(1/0.05)}{2\times 600}} = 0.050$. All observations with KSD value less than t are identified as outliers. We then apply the detector to a test set of 630 i.i.d. observations, among which 600 are generated from F_{good} and the remaining 30 from $F_{outlier}$. Figure 8 shows, for each synthetic data, 630 test observations superimposed with the contour of the KSD at value t (the solid curve). 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. Table I (columns 2-4) shows the false alarm rates and the detection rates of our detector along with the threshold values.

Next, we simulate the missing label scenario. Each of the training and validation set contains 630 i.i.d. observations, of which 600 are generated from F_{good} and 30 from $F_{outlier}$. Hence the data can be viewed as being generated from a mixture distribution $F = (1-\alpha)F_{good} + \alpha F_{outlier}$ where $\alpha = 0.0476$. The kernelized spatial depth function is built upon the training set using Gaussian kernel with $\mathbf{\Sigma} = \mathbf{\Sigma}_2$. Same as the one-class learning scenario, we assume that FAP should be kept below 0.1. So we apply the inequality (9) with $\delta = 0.05$ and $\alpha \leq r = 0.05$ to select the threshold t. Specifically, we search for the maximum value of t that makes the false alarm rate, evaluated from the validation set, no greater than $(1-r)0.1 - \sqrt{\frac{\ln(1/0.05)}{2\times 630}} = 0.046$. We apply the detector to the same test sets as in the one-class learning scenario. Figure 8 shows, for each synthetic data, 630 observations and the contour of KSD at the selected threshold (the dotted curve). Table I (columns 5-7) shows the selected threshold value, the false alarm rate, and the detection rate of our detector.



(a) Synthetic 1 (b) Synthetic 2

(c) Synthetic 3 (d) Synthetic 4

Fig. 8. Decision boundaries of the proposed outlier detectors in one-class learning scenario (solid curves) and missing label scenario (dotted curves) based on 630 i.i.d. test observations in which 600 (marked with $*$'s) were generated from F_{good} and 30 (marked with $\circ$'s) from $F_{outlier}$. Observations outside each contour are classified as outliers.

Compared with the one-class learning setting, the detection rate is lower in the missing label case across all four data sets. This is because we need to be more conservative in selecting the threshold under missing label scenario (the $1 - r$ effect in (9)), which leads to a smaller false alarm rate and a smaller detection rate.

B. New Species Discovery in Taxonomic Research

Approximately 1.4 million species are currently known to science. However, estimates based on the rate of new species discovery place the total number of species on planet earth at 10 to 30 times this number. Human population expansion and habitat destruction are causing extinctions of both known and yet to be discovered species. The accelerated pace of species

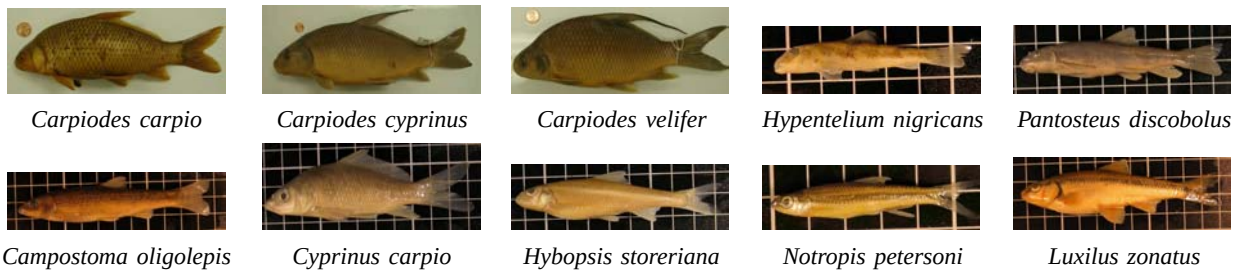


Fig. 9. Sample specimens from ten species of the family *Catostomidae* (suckers) and *Cyprinidae* (minnows).

decline has fueled the current biodiversity crisis, in which it is feared large percentage of the earth's species will be lost before they can be discovered and described. The job of discovering and describing new species falls on taxonomists. Moreover, the pace of taxonomic research, as traditionally practiced, is very slow. In recognizing a species as new to science, taxonomists use a gestalt recognition system that integrates multiple characters of body shape, external body characteristics, and pigmentation patterns. They then make careful counts and measurements on large numbers of specimens from multiple populations across the geographic ranges of both the new and closely related species, and identify a set of external body characters that uniquely diagnoses the new species as distinct from all of its known relatives. The process is laborious and can take years or even decades to complete, depending on the geographic range of the species.

Here we formulate new species discovery as an outlier detection problem. We apply the proposed outlier detection method to a small group of cypriniform fishes, comprising five species of suckers of the family *Catostomidae* and five species of minnows of the family *Cyprinidae*, in order to demonstrate its excellent potential in new species discovery.

1) Data Set and Shape Features: The data set consists of 989 specimens from Tulane University Museum of Natural History (TUMNH). The 989 specimens include 128 *Carpiodes carpio*, 297 *Carpiodes cyprinus*, 172 *Carpiodes velifer*, 42 *Hypentelium nigricans*, 36 *Pantosteus discobolus*, 53 *Campostoma oligolepis*, 39 *Cyprinus carpio*, 60 *Hybopsis storeriana*, 76 *Notropis petersoni*, and 86 *Luxilus zonatus*. We assign identifiers 1 to 10 to the above species. The first five species belong to the family *Catostomidae* (suckers). The next five species belong to the family *Cyprinidae* (minnows). Both families are under the order *Cypriniformes*. Sample images of specimens from the above 10 known species are shown in Figure 9.

Over the past decade, digital landmarking techniques have been widely used to analyze body shape variation, in a procedure called Geometric Morphometrics [39], [2], [63]. These landmarks (LMs) are biologically definable points along the body outline, which are arguably related by evolutionary descent. The LMs of each specimen are saved as two dimensional coordinates. Non-shape related variation in LM coordinates can be removed using techniques such as Generalized

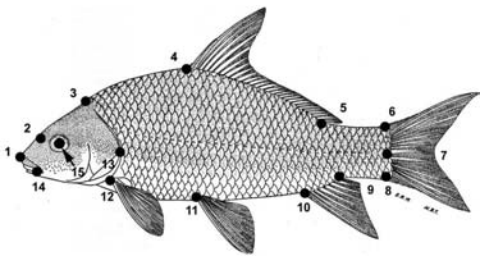


Fig. 10. Digitized 15 homologous landmarks using TpsDIG Version 1.4 (©2004 by F. James Rohlf).

Procrustes Analysis [25], [32]. Figure 10 shows 15 homologous LMs digitized on a fish specimen using the TpsDIG software tool developed by F. James Rohlf of SUNY Stony Brook [4]. Various body shape characters can be extracted from these LMs and expressed in a fairly simple language of lengths, angles, areas, and ratios of these. For example, "*the length of the snout*" is directly related to the slope of the line connecting the tip of the snout (LM 1) and the naris (LM 2), which can be computed as the angle between the vertical axis and the line connecting LM 1 and LM 2. The "*slenderness of the body*" can be defined as the ratio of the body depth (computed as the distance between LM 4 and LM 11) to the body length (computed as the distance between LM 13 and LM 7).

Generalized Procrustes Analysis [32] is used to remove non-shape related variation in LM coordinates. Specifically, the centroid of each configuration (based on the 15 LMs associated with each specimen) is translated to the origin, and configurations are scaled to a common unit size. We then compute 12 features for each specimen using the 15 LMs. A detailed description of these features is given in [14].

2) Results: In the first experiment, we held specimens from one of the 10 species as the "unknown" specimens and specimens of the other 9 species as known. The specimens from the 9 known species are then randomly divided into two groups of roughly equal size. One group is used to build the KSD function. The other group is used to compute the upper bound on the false alarm probability based on (7) for $\delta = 0.05$. The parameter t is chosen such that the upper bound on the false alarm probability is equal to one minus the detection rate evaluated from the unknown specimens. We denote this critical value of the upper bound on the false alarm probability by e^*. The detection rate is therefore $1 - e^*$. Loosely speaking, e^* implies that the false alarm probability of the outlier detector is less than e^* when its detection rate is $1 - e^*$. Therefore, a smaller value of e^* indicates that a larger percentage of the unknown specimens are outliers with respect to the known species, which in turn suggests the possibility that the unknown specimens represent a new species.

[4]http://life.bio.sunysb.edu/morph/

TABLE II

WITH PROBABILITY AT LEAST 0.95, THE FALSE ALARM PROBABILITY IS LESS THAN e^*, AND THE DETECTION RATE IS $1-e^*$. A SMALLER VALUE OF e^* INDICATES A SMALLER FALSE ALARM PROBABILITY AND A LARGER DETECTION RATE. $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$, $\mathbf{\Sigma}_3$, AND $\mathbf{\Sigma}_4$ DENOTE KSD OUTLIER DETECTORS WITH FOUR CHOICES OF THE KERNEL PARAMETER. M_{dist} DENOTES A MAHALANOBIS DISTANCE BASED OUTLIER DETECTOR.

Unknown Species	e^*					Unknown Species	e^*				
	$\mathbf{\Sigma}_1$	$\mathbf{\Sigma}_2$	$\mathbf{\Sigma}_3$	$\mathbf{\Sigma}_4$	M_{dist}		$\mathbf{\Sigma}_1$	$\mathbf{\Sigma}_2$	$\mathbf{\Sigma}_3$	$\mathbf{\Sigma}_4$	M_{dist}
Carp. carp.	0.289	0.258	0.242	0.234	0.523	*Camp. olig.*	0.302	0.264	0.245	0.208	0.396
Carp. cypr.	0.212	0.205	0.202	0.209	0.313	*Cypr. carp.*	0.051	0.051	0.051	0.051	0.077
Carp. veli.	0.198	0.186	0.174	0.180	0.308	*Hybo. stor.*	0.533	0.467	0.433	0.367	0.533
Hype. nigr.	0.048	0.071	0.071	0.071	0.048	*Notr. pete.*	0.605	0.579	0.553	0.487	0.421
Pant. disc.	0.083	0.056	0.083	0.056	0.139	*Luxi. zona.*	0.547	0.535	0.523	0.512	0.384

The results are reported in Table II. As one can see, the proposed outlier detector produces comparable results across all four choices of the kernel parameter. The KSD outlier detector identifies most of the unknown species as outliers, i.e., "new" with high detection rate and low false alarm probabilities. For example, when $\mathbf{\Sigma}_1$ is selected, the detection rate of *Hypentelium nigricans* is 0.952 and its false alarm probability is less than 0.048; the detection rate of *Cyprinus carpio* is 0.949 and its false alarm probability is less than 0.051; *Pantosteus discobolus* has a detection rate 0.917 and false alarm probability less than 0.083; *Carpiodes velifer* has a detection rate 0.802 and false alarm probability less than 0.198; *Carpiodes cyprinus* has a detection rate 0.788 and false alarm probability less than 0.212; *Carpiodes carpio* has a detection rate 0.711 and false alarm probability less than 0.289; and *Campostoma oligolepis* has a detection rate 0.698 and false alarm probability less than 0.302. On the other hand, the method does not produce good detection rate for *Hybopsis storeriana*, *Notropis petersoni*, and *Luxilus zonatus*. The detection rate for *Notropis petersoni* is especially low at 0.395. We also compared KSD outlier detector with a more traditional technique based on Mahalanobis distance (M_{dist}) where a larger M_{dist} value indicates a higher likelihood of being an outlier. On 7 out of the 10 species, this traditional approach produces a detection rate lower than that of the KSD approach (regardless of the choice of the kernel parameter). In addition, it predicts poorly for five species, *Carpiodes carpio*, *Campostoma oligolepis*, *Hybopsis storeriana*, *Notropis petersoni*, and *Luxilus zonatus*. Hence the proposed approach seems to be more competitive on this data set.

C. Comparison with Other Approaches

We compare the performance of the proposed approach with three existing outlier detection algorithms: the well-known LOF method [34], the recent feature bagging method [38], and

TABLE III

PERFORMANCE COMPARISON OF KSD, LOF, FEATURE BAGGING, AND ACTIVE LEARNING OUTLIER DETECTION METHODS. THE AREA UNDER THE ROC CURVE (AUC) FOR EACH METHOD AND EACH DATA SET IS SHOWN. A LARGER AUC VALUE (CLOSER TO 1) INDICATES BETTER PERFORMANCE. $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$, $\mathbf{\Sigma}_3$, AND $\mathbf{\Sigma}_4$ DENOTE THE FOUR PARAMETER SELECTION STRATEGIES PROPOSED IN SECTION V FOR GAUSSIAN KERNEL. POLY2 AND POLY3 REPRESENT POLYNOMIAL KERNELS WITH DEGREE 2 AND 3, RESPECTIVELY.

Data Set			Ann-Thyroid 1	Ann-Thyroid 2	Shuttle	KDD-Cup'99
Outlier Class			Class 1	Class 2	Class 2, 3, 5–7	U2R
Size of Data Set			3428	3428	14500	60839
KSD	one-class learning	$\mathbf{\Sigma}_1$	0.9782 ± 0.0068	0.8575 ± 0.0095	0.9970 ± 0.0006	0.9797 ± 0.0031
		$\mathbf{\Sigma}_2$	0.9902 ± 0.0024	0.9330 ± 0.0074	0.9969 ± 0.0005	0.9230 ± 0.0121
		$\mathbf{\Sigma}_3$	0.9760 ± 0.0066	0.8805 ± 0.0081	0.9974 ± 0.0005	0.9789 ± 0.0030
		$\mathbf{\Sigma}_4$	0.9864 ± 0.0029	0.9299 ± 0.0076	0.9891 ± 0.0023	0.7224 ± 0.0354
		Poly2	0.9373 ± 0.0135	0.6168 ± 0.0123	0.9902 ± 0.0021	0.9911 ± 0.0004
		Poly3	0.9393 ± 0.0130	0.6106 ± 0.0130	0.9896 ± 0.0022	0.9911 ± 0.0004
KSD	missing label	$\mathbf{\Sigma}_1$	0.9356 ± 0.0131	0.7426 ± 0.0121	0.9322 ± 0.0087	0.9114 ± 0.0017
		$\mathbf{\Sigma}_2$	0.9747 ± 0.0045	0.8785 ± 0.0071	0.8704 ± 0.0030	0.7905 ± 0.0077
		$\mathbf{\Sigma}_3$	0.9271 ± 0.0145	0.7557 ± 0.0111	0.8898 ± 0.0100	0.9009 ± 0.0019
		$\mathbf{\Sigma}_4$	0.9595 ± 0.0070	0.8492 ± 0.0067	0.7972 ± 0.0079	0.6927 ± 0.0350
		Poly2	0.9295 ± 0.0150	0.6076 ± 0.0127	0.9902 ± 0.0021	0.9908 ± 0.0004
		Poly3	0.9330 ± 0.0142	0.6048 ± 0.0133	0.9896 ± 0.0022	0.9908 ± 0.0004
LOF			0.869	0.761	0.825	0.61 ± 0.1
Feature Bagging			0.869	0.769	0.839	0.74 ± 0.1
Active Learning			0.97 ± 0.01	0.89 ± 0.11	0.999 ± 0.0006	0.935 ± 0.04

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 [38] and [1], one of the rare classes is chosen as the outlier class in our experiment. The outlier classes are listed in Table III. In [38], the smallest intrusion class, U2R, was chosen as the outlier class. We found that the outlier class in [38] 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 and Shuttle data sets are selected as training data. For the KDD-Cup 1999 data set, the training set contains $10,000$ randomly chosen observations and the test set has the remaining $50,839$ 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 [38] 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 III along with the standard deviation. The AUC values of the LOF, the feature bagging, and the active learning methods are obtained from [38] 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 III, it is clear that the KSD based outlier detection (one-class learning) using Gaussian kernel consistently outperforms the LOF and the feature bagging methods on all four data sets. The performance of KSD with Gaussian kernel is comparable with that of the active learning on all four data sets (except for $\mathbf{\Sigma}_4$ on KDD-Cup'99 data). We observed that polynomial kernel generates the best performance on the KDD-Cup'99 data.

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.

VIII. Conclusions and Future Work

We have proposed the kernelized spatial depth (KSD) and an outlier detection method using the KSD function. The KSD is a generalization of the spatial depth [61], [13], [74]. 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.

The proposed method has some limitations.

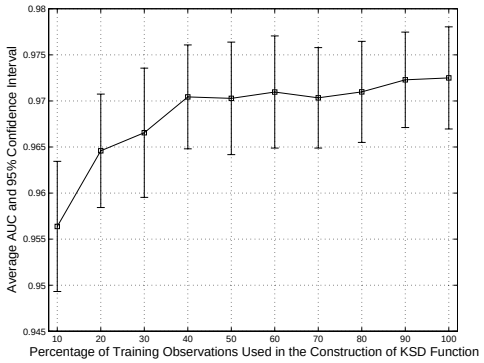


Fig. 11. Plot of average AUC values on Ann-Thyroid 1 data set under random sampling of the training set. Only a portion of randomly selected training observations are used to construct the KSD function. The average AUC value and the corresponding 95% confidence intervals are computed over 10 random runs.

- The implementation of the KSD requires the storage of all ℓ training observations. The required storage space could be prohibitive for applications with large training sets. Furthermore, the rate of the detector can be slow for large scale applications because the computational complexity of evaluating the KSD for an observation is $O(\ell^2)$.
- As currently formulated, the proposed KSD function cannot directly handle symbolic features. In some applications, however, features are symbolic. For example, the 'protocol_type' feature in the KDD-Cup'99 data set takes values of 'udp', 'tcp', or 'icmp.' In our experiments, a symbolic feature is mapped to discrete numbers, e.g., 'udp' $\rightarrow 0$, 'tcp' $\rightarrow 1$, and 'icmp' $\rightarrow 2$. But this mapping inevitably introduces a bias: two symbols are "similar" if they are numerically close.

Continuations of this work could take several directions.

- Using selective sampling to reduce storage space and computational cost. We tested a random sampling method to reduce the computational complexity of the kernelized spatial depth function. Figure 11 shows the AUC values for Ann-Thyroid 1 data set where only a portion of randomly selected training observations were used to construct the KSD function. This simple method seems to perform very well: when 10% of the training observations were used to build the KSD function, the AUC value merely decreased from 0.9725 to 0.9564. It will be promising to investigate other selective sampling approaches.
- Kernel selection. In the current work, the Gaussian kernel is applied in the empirical study. The proposed algorithm for choosing σ parameter for a Gaussian kernel is simple and seems to be effective, but is by no means "optimal." It will be interesting to explore other alternative methods to select σ. It will also be interesting to test other types of kernels. In particular, a kernel defined for symbolic features might provide us a way to integrate

symbolic features into a KSD function.

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 No. DMS-0707074. Henry L. Bart, Jr. is supported by the US National Science Foundation under Grant No. DEB-0237013. The authors would also like to thank Kory P. Northrop for preparing the fish data, Yuanyuan Ding for discussing statistical depth functions, and Huimin Chen for discussing research issues in new species discovery.

APPENDIX

In order to prove the theorems, we need an inequality attributed to McDiarmid.

*Lemma 1 (***McDiarmid):** *Let $X_1, X_2, \ldots, X_n$ be independent random variables taking values in a set $\mathbb{X}$. Suppose that $f : \mathbb{X}^n \to \mathbb{R}$ satisfies*

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\hat{\mathbf{x}}_i\in\mathbb{X}} |f(\mathbf{x}_1,\ldots,\mathbf{x}_n) - f(\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},\hat{\mathbf{x}}_i,\mathbf{x}_{i+1},\ldots,\mathbf{x}_n)| \le c_i$$

for constants $c_i, 1 \le i \le n$. Then for every $\epsilon > 0$,

$$\Pr[f(X_1,\ldots,X_n) - \mathbb{E}f(X_1,\ldots,X_n) \ge \epsilon] \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof of Theorem 1: We break $\mathbb{E}_{F|\mathcal{X}}[g_\kappa(\mathbf{x},\mathcal{X})] - \frac{1}{\ell}\sum_{i=1}^{\ell} g_\kappa(\mathbf{x}_i,\mathcal{X})$ into $A + B + C$:

$$\begin{aligned}
A &= \mathbb{E}_{F|\mathcal{X}}[g_\kappa(\mathbf{x},\mathcal{X})] - \mathbb{E}_{F|\mathcal{X}}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} g_\kappa(\mathbf{x},\mathcal{X}(i))\right], \\
B &= \mathbb{E}_{F|\mathcal{X}}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} g_\kappa(\mathbf{x},\mathcal{X}(i))\right] - \mathbb{E}[g_\kappa(\mathbf{x}_1,\mathcal{X}(1))], \\
C &= \mathbb{E}[g_\kappa(\mathbf{x}_1,\mathcal{X}(1))] - \frac{1}{\ell}\sum_{i=1}^{\ell} g_\kappa(\mathbf{x}_i,\mathcal{X}),
\end{aligned}$$

where $\mathcal{X}(i) = \mathcal{X} - \{\mathbf{x}_i\}$. It is readily checked that

$$\begin{aligned}
|D_\kappa(\mathbf{x},\mathcal{X}) - D_\kappa(\mathbf{x},\mathcal{X}(i))| &= \left|\frac{1}{\ell}\left\|\sum_{j=1}^{\ell} S(\phi(\mathbf{x}) - \phi(\mathbf{x}_j))\right\| - \frac{1}{\ell-1}\left\|\sum_{j=1,j\ne i}^{\ell} S(\phi(\mathbf{x}) - \phi(\mathbf{x}_j))\right\|\right| \\
&\le \frac{1}{\ell}\left\|S(\phi(\mathbf{x}) - \phi(\mathbf{x}_i)) - \frac{1}{\ell-1}\sum_{j=1,j\ne i}^{\ell} S(\phi(\mathbf{x}) - \phi(\mathbf{x}_j))\right\| \le \frac{2}{\ell}
\end{aligned}$$

for $1 \le i \le \ell$, hence

$$|g_\kappa(\mathbf{x},\mathcal{X}) - g_\kappa(\mathbf{x},\mathcal{X}(i))| \le \frac{1}{b}|D_\kappa(\mathbf{x},\mathcal{X}) - D_\kappa(\mathbf{x},\mathcal{X}(i))| \le \frac{2}{\ell b}, \quad \mathbf{x} \in \mathbb{R}^n.$$

Therefore,

$$A \leq \mathbb{E}_{F|\mathcal{X}} \left[\left| g_\kappa(\mathbf{x}, \mathcal{X}) - \frac{1}{\ell} \sum_{i=1}^{\ell} g_\kappa(\mathbf{x}, \mathcal{X}(i)) \right| \right] \leq \frac{2}{\ell b} . \tag{10}$$

Next, we derive bounds for B. It is straightforward to verify that

$$\mathbb{E} \left\{ \mathbb{E}_{F|\mathcal{X}} \left[\frac{1}{\ell} \sum_{i=1}^{\ell} g_\kappa(\mathbf{x}, \mathcal{X}(i)) \right] \right\} = \mathbb{E}[g_\kappa(\mathbf{x}_1, \mathcal{X}(1))] .$$

For a change of one $\mathbf{x}_i$ to $\hat{\mathbf{x}}_i$, denote $\hat{\mathcal{X}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \hat{\mathbf{x}}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_\ell\}$. For fixed $\mathbf{x}$ and i, and any $j \neq i$, we have $|g_\kappa(\mathbf{x}, \mathcal{X}(j)) - g_\kappa(\mathbf{x}, \hat{\mathcal{X}}(j))| \leq \frac{2}{(\ell-1)b}$. Therefore,

$$\begin{aligned} &\sup_{\mathbf{x}_1, \ldots, \mathbf{x}_\ell, \hat{\mathbf{x}}_i \in \mathbb{R}^n} \left| \mathbb{E}_{F|\mathcal{X}} \left[\frac{1}{\ell} \sum_{j=1}^{\ell} g_\kappa(\mathbf{x}, \mathcal{X}(j)) \right] - \mathbb{E}_{F|\hat{\mathcal{X}}} \left[\frac{1}{\ell} \sum_{j=1}^{\ell} g_\kappa(\mathbf{x}, \hat{\mathcal{X}}(j)) \right] \right| \\ = &\sup_{\mathbf{x}_1, \ldots, \mathbf{x}_\ell, \hat{\mathbf{x}}_i \in \mathbb{R}^n} \frac{1}{\ell} \left| \sum_{j=1, j\neq i}^{\ell} \mathbb{E}_{F|\mathcal{X}, \hat{\mathcal{X}}} \left[g_\kappa(\mathbf{x}, \mathcal{X}(j)) - g_\kappa(\mathbf{x}, \hat{\mathcal{X}}(j)) \right] \right| \leq \frac{2}{\ell b} . \end{aligned} \tag{11}$$

By (11), we apply the McDiarmid's inequality to get

$$\Pr(B > \epsilon_1) \leq \exp\left(-\frac{\ell b^2 \epsilon_1^2}{2} \right) . \tag{12}$$

Finally, we look at C. Similar to (11), we have

$$\begin{aligned} &\sup_{\mathbf{x}_1, \ldots, \mathbf{x}_\ell, \hat{\mathbf{x}}_i \in \mathbb{R}^n} \left| \frac{1}{\ell} \sum_{j=1}^{\ell} g_\kappa(\mathbf{x}_j, \mathcal{X}) - \frac{1}{\ell} \sum_{j=1}^{\ell} g_\kappa(\mathbf{x}_j, \hat{\mathcal{X}}) \right| \\ \leq &\sup_{\mathbf{x}_1, \ldots, \mathbf{x}_\ell, \hat{\mathbf{x}}_i \in \mathbb{R}^n} \frac{1}{\ell} \left| \sum_{j=1, j\neq i}^{\ell} \left[g_\kappa(\mathbf{x}_j, \mathcal{X}(j)) - g_\kappa(\mathbf{x}_j, \hat{\mathcal{X}}(j)) \right] \right| + \frac{1}{\ell} \leq \frac{2}{\ell b} + \frac{1}{\ell} . \end{aligned} \tag{13}$$

Hence by McDiarmid's inequality, we obtain

$$\Pr(C > \epsilon_2) \leq \exp\left(-\frac{2\ell \epsilon_2^2}{\left(\frac{2}{b} + 1\right)^2} \right) . \tag{14}$$

Setting $\frac{\delta}{2} = \exp\left(-\frac{\ell b^2 \epsilon_1^2}{2}\right) = \exp\left(-\frac{2\ell \epsilon_2^2}{\left(\frac{2}{b}+1\right)^2}\right)$, and solving for ϵ_1 and ϵ_2, we complete the proof by combining (10), (12), and (14). □

Proof of Theorem 2: Because $\mathbf{y}_i \notin \mathcal{X}$ and g_κ is bounded by 1, a change of one $\mathbf{y}_i$ in $\frac{1}{\ell_{test}} \sum_{i=1}^{\ell_{test}} g_\kappa(\mathbf{y}_i, \mathcal{X})$ results in at most a change of $1/\ell_{test}$. Thus an application of the McDiarmid's inequality yields

$$\Pr\left[\mathbb{E}_{F|\mathcal{X}}[g_\kappa(\mathbf{y}_1, \mathcal{X})] - \frac{1}{\ell_{test}} \sum_{i=1}^{\ell_{test}} g_\kappa(\mathbf{y}_i, \mathcal{X}) > \epsilon \middle| \mathcal{X} \right] \leq \exp\left(-2\ell\epsilon^2\right) .$$

Therefore

$$\Pr\left[\mathbb{E}_{F|\mathcal{X}}[g_\kappa(\mathbf{y}_1,\mathcal{X})] - \frac{1}{\ell_{test}}\sum_{i=1}^{\ell_{test}} g_\kappa(\mathbf{y}_i,\mathcal{X}) > \epsilon\right]$$
$$= \mathbb{E}\left\{\Pr\left[\mathbb{E}_{F|\mathcal{X}}[g_\kappa(\mathbf{y}_1,\mathcal{X})] - \frac{1}{\ell_{test}}\sum_{i=1}^{\ell_{test}} g_\kappa(\mathbf{y}_i,\mathcal{X}) > \epsilon \,\middle|\, \mathcal{X}\right]\right\} \leq \exp\left(-2\ell\epsilon^2\right) .$$

Setting $\delta = \exp\left(-2\ell\epsilon^2\right)$ and solving for ϵ, we complete the proof. □

Proof of Theorem 3: From $F = (1-\alpha)F_{good} + \alpha F_{outlier}$ we have

$$\mathbb{E}_{F|\mathcal{X}}[g_\kappa(\mathbf{x},\mathcal{X})] = (1-\alpha)\mathbb{E}_{F_{good}|\mathcal{X}}[g_\kappa(\mathbf{x},\mathcal{X})] + \alpha\mathbb{E}_{F_{outlier}|\mathcal{X}}[g_\kappa(\mathbf{x},\mathcal{X})] .$$

Therefore in view of $g_\kappa \geq 0$ we have

$$(1-\alpha)\mathbb{E}_{F_{good}|\mathcal{X}}[g_\kappa(\mathbf{x},\mathcal{X})] \leq \mathbb{E}_{F|\mathcal{X}}[g_\kappa(\mathbf{x},\mathcal{X})]$$

Thus the desired proof follows from $\alpha \leq r$. □

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] D. C. Adams and F. J. Rohlf, "Ecological Character Displacement in Plethodon: Biomechanical Differences Found from a Geometric Morphometric Study," *Proceedings of the National Academy of Sciences*, vol. 97, pp. 4106–4111, 2000.

[3] 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.

[4] 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.

[5] A. C. Atkinson, "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1329–1339, 1994.

[6] 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.

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

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

[10] M. J. Carlotto, "A Cluster-Based Approach for Detecting Man-Made Objects and Changes in Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 2, pp. 374–387, 2005.

[11] D. Castaño and A. Kunoth, "Robust Regression of Scattered Data with Adaptive Spline-Wavelets," *IEEE Transactions on Image Processing*, vol. 15, no. 6, pp. 1621–1632, 2006.

[12] 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.

[13] 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.

[14] Y. Chen, H. L. Bart, Jr., S. Huang, and H. Chen, "A Computational Framework for Taxonomic Research: Diagnosing Body Shape within Fish Species Complexes," *Proc. 5th IEEE Int'l Conf. on Data Mining*, pp. 593–596, 2005.

[15] C. Croux and P. J. Rousseeuw, "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273–1283, 1993.

[16] X. Dang and R. Serfling, "Nonparametric Depth-Based Multivariate Outlier Identifiers, and Robustness Properties," *submitted for journal publication*, 2006.

[17] G. Danuser and M. Stricker, "Parametric Model Fitting: From Inlier Characterization to Outlier Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 263–280, 1998.

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

[19] S. Fidler, D. Skočaj, and A. Leonardis, "Combining Reconstructive and Discriminative Subspace Methods for Robust Classification and Regression by Subsampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 337–350, 2006.

[20] 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.

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

[22] A. K. Ghosh and P. Chaudhuri, "On Data Depth and Distribution-Free Discriminant Analysis Using Separating Surfaces," *Bernoulli*, vol. 11, no. 1, pp. 1–27, 2005.

[23] A. K. Ghosh and P. Chaudhuri, "On Maximum Depth Classifiers," *Scandinavian Journal of Statistics*, vol. 32, no. 2, pp. 327–350, 2005.

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

[25] J. C. Gower, "Generalized Procrustes Analysis," *Psychometrika*, vol. 40, pp. 33–51, 1975.

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

[27] 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.

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

[29] 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.

[30] 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.

[31] R. Jörnsten, "Clustering and Classification Based on the L_1 Data Depth," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 67–89, 2004.

[32] D. G. Kendall, "Shape-Manifolds, Procrustean Metrics and Complex Projective Spaces," *Bulletin of the London Mathematical Society*, vol. 16, pp. 81–121, 1984.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] S. Lele and J. T. Richtsmeier, "Euclidean Distance Matrix Analysis: a Coordinate Free Approach for Comparing Biological Shapes Using Landmark Data," *American Journal of Physical Anthropology*, vol. 86, pp. 415–427, 1991.

[40] A. Leonardis and H. Bischof, "Robust Recognition Using Eigenimages," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 99–118, 2000.

[41] R. Y. Liu, "On a Notion of Data Depth Based on Random Simplices," *The Annals of Statistics*, vol. 18, no. 1, pp. 405–414, 1990.

[42] 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.

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

[44] 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.

[45] M. Markou and S. Singh, "A Neural Network-Based Novelty Detection for Image Sequence Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1664–1677, 2006.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.

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

[54] D. M. Rocke and D. L. Woodruff, "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, vol.91, no. 435, pp. 1047–1061, 1996.

[55] P. J. Rousseeuw and K. van Driessen, "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics* vol. 41, no. 3, pp. 212–223, 1999.

[56] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, New York, Wiley, 1987.

[57] P. J. Rousseeuw and I. Ruts, "Algorithm AS 307: Bivariate Location Depth," *Applied Statistics*, vol. 45, no. 4, pp. 516–526, 1996.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[63] D. E. Slice, "Landmark Coordinates Aligned by Procrustes Analysis Do Not Lie in Kendall's Shape Space," Systematic Biology, vol. 50, pp. 141–149, 2001.

[64] C. G. Small, "A Survey of Multidimensional Medians," *International Statistical Review*, vol. 58, no. 3, pp. 263–277, 1990.

[65] D. Song, M. I. Heywood, and A. N. Zincir-Heywood, "Training Genetic Programming on Half a Million Patterns: An Example From Anomaly Detection," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 3, pp. 225–239, 2005.

[66] C. Stauffer and W. E. Grimson, "Learning Patterns of Activity Using Real-Time Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

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

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

[69] J. Takeuchi and K. Yamanishi, "A Unifying Framework for Detecting Outliers and Change Points from Time Series," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 4, pp. 482–492, 2006.

[70] 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.

[71] M. Thottan and C. Ji, "Anomaly Detection in IP Networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2191–2204, 2003.

[72] J. W. Tukey, "Mathematics and Picturing Data," *Proc. 1975 Int'l Congress of Mathematics*, vol. 2, pp. 523–531, 1974.

[73] V. Vapnik, The Nature of Statistical Learning Theory, New York: Springer-Verlag, 1995.

[74] 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.

[75] 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.

[76] L. Wei, E. Keogh, and X. Xi, "SAXually Explicit Images: Finding Unusual Shapes," *Proc. IEEE Int'l Conf. on Data Mining*, pp. 711–720, 2006.

[77] 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," *Manuscript*, 2006 (http://www.utdallas.edu/∼serfling/papers/ZhouSerfling2006.pdf).

[78] Y. Zuo and R. Serfling, "General Notions of Statistical Depth Function," *The Annals of Statistics*, vol. 28, no. 2, pp. 461–482, 2000.

PLACE PHOTO HERE

Yixin Chen (S'99-M'03) received the B.S. degree (1995) from the Department of Automation, Beijing Polytechnic University, the M.S. degree (1998) in control theory and application from Tsinghua University, and the M.S. (1999) and Ph.D. (2001) degrees in electrical engineering from the University of Wyoming. In 2003, he received the Ph.D. degree in computer science from The Pennsylvania State University. From August 2003 to July 2006, he was an Assistant Professor of computer science at University of New Orleans. Since August 2006, he has been an Assistant Professor at the Department of Computer and Information Science, The University of Mississippi. His research interests include machine learning, data mining, computer vision, bioinformatics, and robotics and control. Dr. Chen is a member of the ACM, the IEEE, the IEEE Computer Society, the IEEE Neural Networks Society, and the IEEE Robotics and Automation Society.

PLACE PHOTO HERE

Xin Dang received the PhD degree in statistics from the University of Texas at Dallas, working with Robert Serfling on statistical depth functions. Currently she is an assistant professor of department of mathematics at the University of Mississippi, joining the faculty in 2005. Her research interests include robust and nonparametric statistics, statistical and numerical computing, and multivariate data analysis. In particular, she has focused on data depth and application, machine learning, and robust procedure computation. Dr. Dang is a member of the American Statistical Association and Institute of Mathematical Statistics.

PLACE PHOTO HERE

Hanxiang Peng received the doctoral degree in mathematics from Binghamton University in 2001. He was an assistant professor in the department of mathematics, the University of Mississippi, from 2001 to 2006 and currently holds an associate professorship in this department. His current interests of research include robust and efficient statistical inference, modeling of correlated data, survival analysis, and semiparametric modeling.

PLACE PHOTO HERE

Henry L. Bart, Jr. is Professor of Ecology and Evolutionary Biology at Tulane University, and Director and Curator of Fishes of the Tulane Museum of Natural History. He is Editor of Tulane Studies in Zoology and Botany and Occasional Papers Tulane University Museum of Natural History. He earned BS and MS degrees from University of New Orleans and a Ph.D. (1985) in Zoology from the University of Oklahoma. He held faculty positions at the University of Illinois and Auburn University prior to joining Tulane University in 1992. His area of research specialization is ecology and systematics of freshwater fishes.