Depth-Based Novelty Detection and its Application to Taxonomic Research

Yixin Chen[1], Henry L. Bart, Jr.[2], Xin Dang[3], Hanxiang Peng[4]
[1]Dept. of Computer Science, [2]Tulane University Museum of Natural History, [3,4]Dept. of Mathematics
[1,3,4]University of Mississippi, University, MS 38677, USA. [2]Belle Chasse, LA 70037, USA.
[1]ychen@cs.olemiss.edu, [2]hank@museum.tulane.edu, {[3]xdang,[4]mmpeng}@olemiss.edu

Abstract

It is estimated that less than 10 percent of the world's species have been described, yet species are being lost daily due to human destruction of natural habitats. The job of describing the earth's remaining species is exacerbated by the shrinking number of practicing taxonomists and the very slow pace of traditional taxonomic research. In this article, we tackle, from a novelty detection perspective, one of the most important and challenging research objectives in taxonomy – new species identification. We propose a unique and efficient novelty detection framework based on statistical depth functions. Statistical depth functions provide from the "deepest" point a "center-outward ordering" of multidimensional data. In this sense, they can detect observations that appear extreme relative to the rest of the observations, i.e., novelty. Of the various statistical depths, the spatial depth is especially appealing because of its computational efficiency and mathematical tractability. We propose a novel statistical depth, the kernelized spatial depth (KSD) that 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. Observations with depth values less than a threshold are declared as novel. The proposed algorithm is simple in structure: the threshold is the only one parameter for a given kernel. We give an upper bound on the false alarm probability of a depth-based detector, which can be used to determine the threshold. Experimental study demonstrates its excellent potential in new species discovery.

1. Introduction

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 decline has fueled the current biodiversity crisis [20], 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. The science of taxonomy has also been suffering from dwindling numbers of experts over the past few decades [25]. 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.

We believe that the pace of data gathering and analysis in taxonomy can be greatly increased through the integration of machine learning and data mining techniques into taxonomic research. In this paper, we tackle one of the most important and challenging research objectives in taxonomy – new species discovery.

1.1. Novelty Detection as a One-Class Learning Problem

From a machine learning perspective, new species discovery is closely related to novelty detection. Novelty detection is one of the most challenging problems in data mining [8]. When "normal" observations are given as a training data set, novelty detection can be formulated as finding observations that significantly deviate from the training data, which is essentially a one-class learning problem.

A statistically natural tool for quantifying the deviation is the probability density of the normal observations. Roberts and Tarassenko [24] approximated the distribution of the training data by a Gaussian mixture model. For every observation, an novelty score is defined as the maximum of the likelihood that the observation is generated by each Gaussian component. An observation is identified as novel if the score is less than a threshold. Schweizer and Moura [29] 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 [18] 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 novelty in the unlabeled samples. Parra et al. [19] proposed a class of volume conserving maps 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. [28] 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 novel observations 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) [34] from an optimization viewpoint, the method is called 1-class SVM. Along the line of 1-class SVM, Campbell and Bennett [5] 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. [22] developed a boosting algorithm for one-class classification based on connections between boosting and SVMs. Banerjee et al. [3] applied 1-class SVM for anomaly detection in hyperspectral images and demonstrated improved performance compared with the method described in [23].

There is an abundance of prior work that applies standard supervised learning techniques to tackle novelty detection [1, 11, 17, 32]. 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 novel observations. In [17], a neural network-based novelty detector is trained based on normal observations and artificial novel examples generated by a uniform distribution. Han and Cho [11] 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 novelty 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. [32] provide an interesting justification for the above heuristic by converting novelty detection to a problem of finding level sets of data generating density.

1.2. An Overview of Our Approach

In this paper, we propose a new novelty detection framework based on the notion of *statistical depths*. Novelty detection methods that are based on statistical depths have been studied in statistics and computational geometry [21, 27]. These methods provide a center-outward ordering of observations. Novel observations 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 [21, 27], 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 [33] and simplicial depth [16] is $O(\ell^{d-1}\log\ell)$ [26]; for projection depth [36], it is $O([\binom{2(d-1)}{d-1}/d]^2\ell^3)$ [9].

Of the various depths the *spatial depth* is especially appealing because of its computational efficiency and mathematical tractability [30]. Its computational complexity is of magnitude $O(\ell^2)$, independent of dimension d. Because each observation from a data set contributes equally to the value of depth function, spatial depth takes a global view of the data set. Consequently the novelty can be called as "globally" novel observations. Nevertheless, many data sets from real-world applications exhibit more delicate structures that entail identification of novelty relative to its neighborhood, i.e., "locally" novel observations.

We develop a novelty detection framework that avoids the above limitation of spatial depth. Specifically, 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. 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 novel. For a given kernel, the threshold on the depth value is the only parameter of the algorithm. We provide an upper bound on the false alarm probability of the detector, i.e., the probability of misclassifying a normal observation as novel. The upper bound can be used to determine the threshold. We apply the proposed novelty detector 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.

The remainder of the paper is organized as follows. Section 2 motivates spatial depth-based novelty detection via the connection between spatial depth and spatial median. Section 3 introduces kernelized spatial depth. Section 4 presents an upper bound on the false alarm probability of the proposed kernelized spatial depth-based novelty detector and provides an algorithmic view of the approach. In Section 5, we explain the experimental studies conducted and demonstrate the results. We conclude and discuss possible future work in Section 6.

2. Medians, Spatial Depth, and Novelty Detection

As Barnett and Lewis described [4], *"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 novel than observations that are closer to the center. This suggests a simple novelty detection approach based on the distance between an observation and the center of a distribution.

2.1. 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 $[\mathbf{x}_{(\ell/2)}, \mathbf{x}_{((\ell+1)/2)}]$ can be defined to be the sample median. A convenient choice is the average $\frac{\mathbf{x}_{(\ell/2)}+\mathbf{x}_{((\ell+1)/2)}}{2}$. Next, we present an equivalent definition that can be naturally generalized to a higher dimensional setting.

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

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

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

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

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* or *the unit vector* [6], 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

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

The median defined in (2) is named as the *spatial median* or the L_1 *median* [35]. Next we give another equivalent definition of the spatial median that motivates the depth-based novelty detection.

2.2. The Spatial Depth

The concept of spatial depth was formally introduced by Serfling [30] based on the notion of spatial quantiles proposed by Chaudhuri [7], while a similar concept, L_1 depth, was first described by Vardi and Zhang [35]. 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*:

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

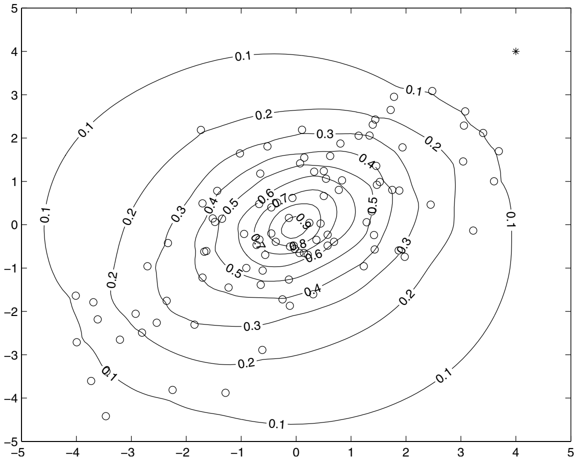


Figure 1. A contour plot of the sample spatial depth based on 100 random observations (represented by $\circ$'s) from a bi-variate Gaussian distribution. The depth values are indicated on the contours. The example marked with $*$ represents a possible novel observation. It has a very low depth value of 0.0219.

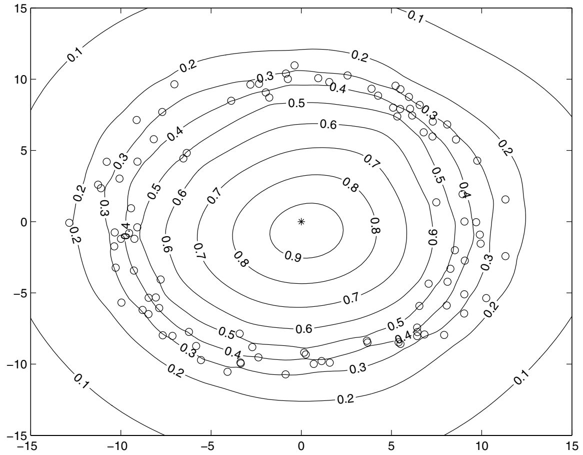


Figure 2. Contour plot of the sample spatial depths based on 100 random observations (denoted by $\circ$'s) of a ring shaped distribution. The depth values are indicated on the contours. The example (denoted by $*$) at the center represents a possible novel observation. It is depth value is 0.9544.

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" of a data point, which can be used for *novelty detection*. From now on all depths refer to the sample depth.

2.3. Novelty 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 bi-variate 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 novel than a point with a high depth value. For example, the point on the upper right corner on Figure 1 (marked with $*$) has a very low depth value of 0.0219. It is isolated and far away from the rest of the data points. This example motivates a simple novelty detection algorithm: *Identify a data point as novel 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 4 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 [12], 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 novel if its depth is low because a lower depth implies a larger distance from the "center" of the data cloud. However, in general, the relationship between the depth and the novelty 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 ring shaped distribution. From the shape of the distribution, it is reasonable to view the point (marked with $*$) in the center as a novel observation. However, the depth at the location of the $*$ is as high as 0.9544. In fact, all of the 100 normal observations have depth smaller than that of the "novel" observation at the center.

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*.

3. Kernelized Spatial Depth

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

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

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

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

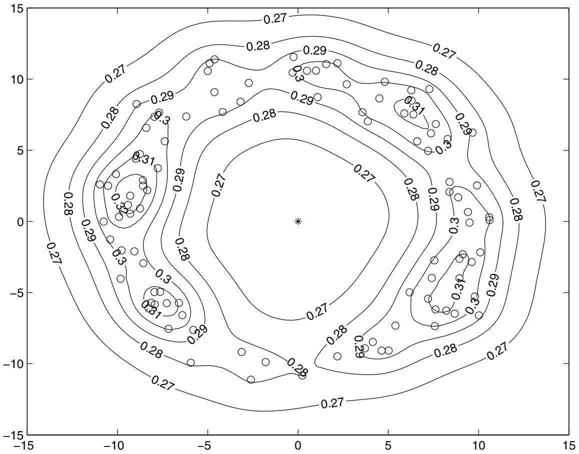


Figure 3. Contour plots of KSD functions based on 100 random observations (marked with $\circ$'s) from a ring-shaped distribution. The depth values are marked on the contours. The depth is kernelized with a Gaussian kernel with $\sigma = 3$. The example (marked with $*$) at the center represents a possible novel observation. It has a depth value of 0.2651.

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

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

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

The KSD (4) is defined for any positive definite kernels. Here we shall be particularly interested in *stationary kernels* (e.g., the Gaussian kernel), because of their close relationship with similarity measures. Figure 3 shows the contour plot of the KSD based on the same 100 random observations generated from the ring shaped distribution in Figure 2. 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 data cloud. Moreover, the depth values are small for the possible novelty. The depth values at the location of the $*$ is 0.2651. A threshold of 0.27 can separate the novel observation from the rest of the ring data. The remaining question is how we determine the threshold. This is addressed in the following section.

4. A Bound 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 novel. In the following, we derive a probabilistic bound on FAP.

Novelty 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. Based on the observations $\mathcal{X}$, a given datum $\mathbf{x}$ is classified as *normal* or *novel* according to whether or not it is generated from F. Let $g : \mathbb{R}^n \rightarrow [0, 1]$ be a novelty detector where $g(\mathbf{x}) = 1$ indicates that $\mathbf{x}$ is novel. The FAP of a novelty detector g, $P_{FA}(g)$, is the probability that an observation generated from F is classified by the detector g as novel, i.e.

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

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

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

For a given data set $\mathcal{X}$ and kernel κ, we define a novelty detector $g_\kappa(\mathbf{x}, \mathcal{X})$ by

$$g_\kappa(\mathbf{x}, \mathcal{X}) = \begin{cases} 1, & \text{if } D_\kappa(\mathbf{x}, \mathcal{X}) \leq t, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $t \in [0, 1]$ is a threshold. An observation $\mathbf{x}$ is classified as novel according to $g_\kappa(\mathbf{x}, \mathcal{X}) = 1$. Denote $\mathbb{E}_F$ the expectation calculated under cdf F. It follows that

$$P_{FA}(g_\kappa) = \mathbb{E}_F\left[g_\kappa(\mathbf{x}, \mathcal{X})\right].$$

We have the following theorem for the bound of the FAP.

Theorem 1 *Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\ell_{train}}\} \subset \mathbb{R}^n$ 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 a novelty detector defined in (5). Fix $\delta \in (0, 1)$. For a new random observation $\mathbf{x}$ from cdf F, the following bound holds with probability at least $1 - \delta$:*

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

It is worthwhile to note that there are two sources of randomness in the above inequality: the random sample $\mathcal{Y}$ and the random observation $\mathbf{x}$. For a specific $\mathcal{Y}$, the above bound is either true or false, i.e., it is not random. For a random sample $\mathcal{Y}$, the probability that the bound is true is at least $1 - \delta$. 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 $\frac{1}{\ell_{test}} \sum_{i=1}^{\ell_{test}} g_\kappa(\mathbf{y}_i, \mathcal{X})$, which is the false alarm rate – the sample version of FAP.

Note that the detector is constructed from the training set $\mathcal{X}$ and evaluated using an independent test set $\mathcal{Y}$. A bound as such is usually called a *test set bound* [14]. 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. For a given desired FAP, we should choose the threshold to be the maximum value of t such that the right-hand side of (6) does not exceed the desired FAP. A proof of Theorem 1 is given in the Appendix.

We summarize the above discussion in pseudo code. The input is a set of observations $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\ell_{train}}\} \subset \mathbb{R}^n$ from an unknown cdf F and a kernel κ. The following pseudo codes determine whether an observation $\mathbf{x}$ is novel.

Algorithm 1 Learning a Novelty Detector

```
FOR (every pair of $\mathbf{x}_i$ and $\mathbf{x}_j$ in $\mathcal{X}$)
    $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$
END
given input $\mathbf{x}$
FOR (every $\mathbf{x}_i$ in $\mathcal{X}$)
    $\zeta_i = \kappa(\mathbf{x}, \mathbf{x}_i)$
    $\delta_i = \sqrt{\kappa(\mathbf{x}, \mathbf{x}) + K_{ii} - 2\zeta_i}$
    IF $\delta_i = 0$
        $z_i = 0$
    ELSE
        $z_i = \frac{1}{\delta_i}$
    END
END
FOR (every pair of $\mathbf{x}_i$ and $\mathbf{x}_j$ in $\mathcal{X}$)
    $\widetilde{K}_{ij} = \kappa(\mathbf{x}, \mathbf{x}) + K_{ij} - \zeta_i - \zeta_j$
END
$D_\kappa(\mathbf{x}, \mathcal{X}) = 1 - \frac{1}{|\mathcal{X}\cup\{\mathbf{x}\}|-1}\sqrt{\mathbf{z}^T \widetilde{K} \mathbf{z}}$
OUTPUT ($\mathbf{x}$ is novel if $D_\kappa(\mathbf{x}, \mathcal{X}) \leq t$)
```

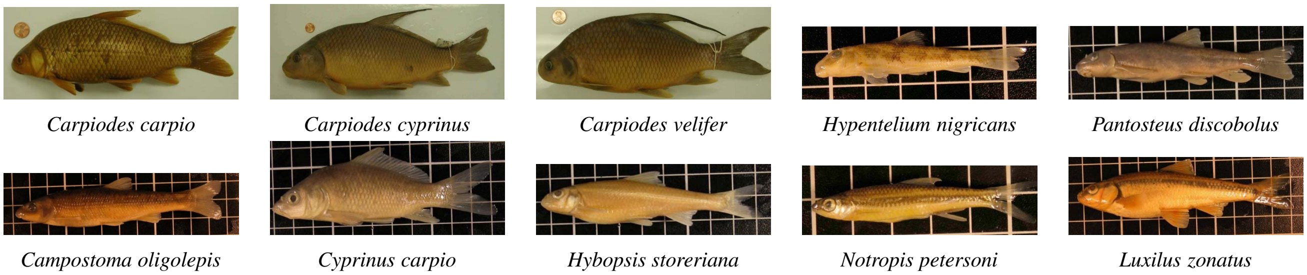
In terms of the number of kernel evaluations and multiplications, the cost of computing the KSD depth for a given observation is $O(\ell^2)$. The above pseudo code assumes that the kernel κ is given. Specifically, for Gaussian kernel, which is used in our experimental study, this requires the knowledge of σ value. Finding an optimal kernel for a given problem is an interesting research issue for its own sake, but is out of the scope of this paper. We propose the following method to select the σ parameter for a given set of observations.

Algorithm 2 Deciding σ for Gaussian Kernel

```
FOR (every observation $\mathbf{x}_i$ in $\mathcal{X}$)
    $d_i = \min_{j=1,\ldots,\ell, j\neq i} \|\mathbf{x}_i - \mathbf{x}_j\|$
END
OUTPUT ($\sigma = \text{median}(d_1, d_2, \ldots, d_\ell)$)
```

5. Experimental Results

We apply the proposed novelty detector method to a small group of cypriniform fishes, comprising five species


Carpiodes carpio *Carpiodes cyprinus* *Carpiodes velifer* *Hypentelium nigricans* *Pantosteus discobolus*

Campostoma oligolepis *Cyprinus carpio* *Hybopsis storeriana* *Notropis petersoni* *Luxilus zonatus*

Figure 4. Sample specimens from ten species of the family *Catostomidae* **(suckers) and** *Cyprinidae* **(minnows).**

of suckers of the family *Catostomidae* and five species of minnows of the family *Cyprinidae*. In all the experiments, the KSD is computed using the Gaussian kernel with the σ parameter being determined from Algorithm 2.

5.1. 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 4.

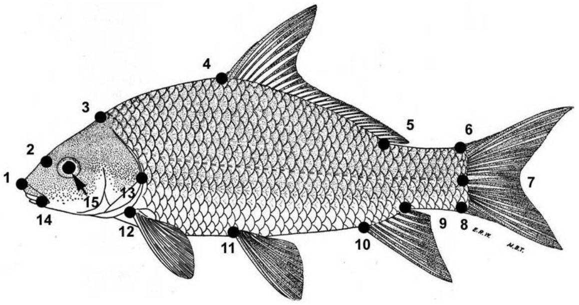


Figure 5. Digitized 15 homologous landmarks using TpsDIG Version 1.4 (©2004 by F. James Rohlf).

Over the past decade, digital landmarking techniques have been widely used to analyze body shape variation, in a procedure called Geometric Morphometrics [15, 2, 31]. The landmarks (LM) are biologically definable points along the body outline, which are arguably related by evolutionary descent. The LM of each specimen are saved as two dimensional coordinates. Non-shape related variation in LM coordinates can be removed using techniques such as Generalized Procrustes Analysis [10, 13]. Figure 5 shows 15 homologous LM digitized on a fish specimen using the TpsDIG software tool developed by F. James Rohlf of SUNY Stony Brook [1]. Various body shape characters can be extracted from these LM 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$_7$).

Digital images of all specimens are uploaded into the TpsDIG software tool, and 15 homologous LM are digitized along the body outline of each specimen (Figure 5). The LM of each specimen are saved as 2-dimensional coordinates. Next, Generalized Procrustes Analysis [13] is used to remove non-shape related variation in LM coordinates. Specifically, the centroid of each configuration (based on the 15 LM associated with each specimen) is translated to the origin, and configurations are scaled to a common unit size. We then compute 12 features, $x_1, \ldots, x_{12}$, for each specimen using the 15 LM. The description of each feature is given in Table 1.

5.2. Results

In the first experiment, we held specimens from one of the 10 species as "unknown" specimens and specimens of the other 9 species as known. 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 (6) for $\delta = 0.05$. The parameter t is chosen such that the upper bound on the FAP is equal to one minus the detection rate evaluated from the "unknown" specimens. We denote this critical value of the upper bound on the FAP by e^*. The detection rate is therefore $1 - e^*$. Loosely speaking, e^* implies

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

Table 1. Features describing shape characters. LM_i denotes the coordinates of the i-th landmark. Non-shape related variation has been removed from the landmarks.

x_1	The distance between the tip of the snout and the naris, computed as the distance between LM_1 and LM_2.
x_2	The slope of the line connecting the tip of the snout and the naris, computed as the angle between the vertical axis and the line connecting LM_1 and LM_2.
x_3	The distance between the naris and the back of the mouth, computed as the distance between LM_2 and LM_{14}.
x_4	The slope of the line connecting the naris and the back of the mouth, computed as the angle between the vertical axis and the line connecting LM_2 and LM_{14}.
x_5	The size of head in proportion of the size of the body, computed as the area of the head polygon (vertices defined in sequence by LM_1, LM_2, LM_3, LM_{13}, LM_{12}, and LM_{14}) divided by the area of the body polygon (vertices defined in sequence by LM_3, LM_4, LM_5, LM_6, LM_7, LM_8, LM_9, LM_{10}, LM_{11}, LM_{12}, and LM_{13})
x_6	The length of the head in proportion of the length of the body, computed as the distance between LM_1 and LM_{13} divided by the distance between LM_{13} and LM_7.
x_7	The distance between LM_7 and LM_8.
x_8	The sum of the distance between LM_3 and LM_{13}, the distance between LM_{12} and LM_{13}, and the distance between LM_1 and LM_{13} divided by the distance between LM_{13} and LM_7.
x_9	The distance between the naris and the tip of the snout in proportion to the distance between the naris and the eye, computed as the distance between LM_1 and LM_2 divided by the distance between LM_2 and LM_{15}
x_{10}	The distance between LM_4 and LM_{11} divided by the distance between LM_{13} and LM_7.
x_{11}	The distance between LM_3 and LM_4 divided by the distance between LM_{13} and LM_7.
x_{12}	The angle between the vertical axis and the line connecting LM_{10} and LM_5.

Table 2. With probability at least 0.95, the FAP is less than e^*, and the detection rate is $1-e^*$. A smaller value of e^* indicates a smaller FAP and a larger detection rate.

Unknown Species	e^*	Unknown Species	e^*
Carpiodes carpio	0.258	*Campostoma oligolepis*	0.302
Carpiodes cyprinus	0.202	*Cyprinus carpio*	0.051
Carpiodes velifer	0.192	*Hybopsis storeriana*	0.517
Hypentelium nigricans	0.071	*Notropis petersoni*	0.592
Pantosteus discobolus	0.083	*Luxilus zonatus*	0.547

that the FAP of the novelty 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 novel with respect to the known species, which in turn suggests the possibility that the unknown specimens represent a new species.

The results are given in Table 2. As you can see, the proposed novelty detector identifies most of the "unknown" species as novel, i.e., "new" with high detection rates and low FAPs: the detection rate of *Cyprinus carpio* is 0.949 and its FAP is less than 0.051, the detection rate of *Hypentelium nigricans* is 0.929 and its FAP is less than 0.071, *Pantosteus discobolus* has a detection rate 0.917 and FAP less than 0.083, *Carpiodes velifer* has a detection rate 0.808 and FAP less than 0.192, *Carpiodes cyprinus* has a detection rate 0.798 and FAP less than 0.202, *Carpiodes carpio* has a detection rate 0.742 and FAP less than 0.258, and *Campostoma oligolepis* has a detection rate 0.698 and FAP 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 0.408.

We interpret the low detection rates for some species as a consequence of a "masking" effect as illustrated in Figure 6. The 20 novel observations, marked with $*$'s, are i.i.d. observations from a uniform distribution over $[-1, 1] \times [-1, 1]$. The 400 known observations come from one of the following Gaussian distributions: $N_1 \sim N([2, 2]^T, I)$ (marked with $\circ$'s), $N_2 \sim N([-2, 2]^T, I)$ (marked with $\triangleleft$'s), $N_3 \sim N([2, -2]^T, I)$ (marked with $\diamond$'s), $N_4 \sim N([-2, -2]^T, I)$ (marked with $\triangleright$'s). Clearly, the novel observations are submerged into (or masked by) the known observations. If we construct the KSD function using 200 known observations and evaluate the upper bound on the FAP from the remaining 200 known observations, we obtain $e^* = 0.85$, i.e., a detection rate 0.15 when the FAP is less then 0.85 (a decrease in the upper bound on the FAP will further reduce the detection rate). However, if we consider one Gaussian at a time, we get (1) N_1: detection rate 0.9 and FAP less than 0.1; (2) N_2: detection rate 0.75 and FAP less than 0.25; (3) N_3: detection rate 0.85 and FAP less than 0.15; (4) N_4: detection rate 0.75 and FAP less than 0.25. This suggests that one may reduce the masking effect via a pairwise test, i.e., testing the specimens from the unknown species against each known species separately.

We summarize the pair-wise test results of the above 10 species in Table 3. Since the sample size of several species is rather small, the upper bound derived using (6) is very loose. Instead of reporting e^* values, we present equal error rates, the value at which the false alarm rate (the $\frac{1}{\ell_{test}} \sum_{i=1}^{\ell_{test}} g_\kappa(\mathbf{y}_i, \mathcal{X})$ term in (6)) is identical to one minus the detection rate. The ij-th entry of Table 3 presents the value of equal error rate at testing the unknown species i against the known species j (the numerical identifier of species is given at the beginning of Section 5.1). The three

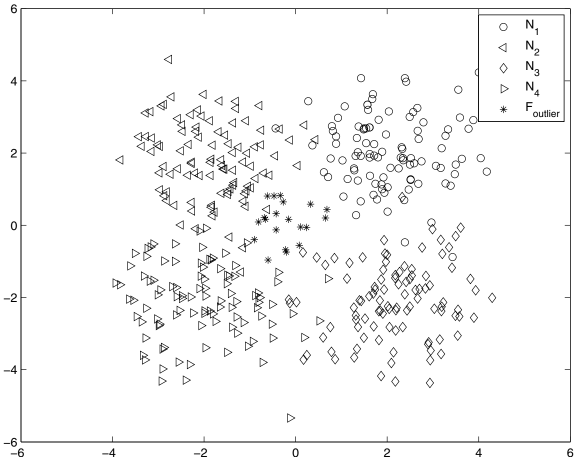


Figure 6. An example of the masking effect. With high detection rate and low FAP, novel observations (marked with $*$'s) can be detected from a group of known observations generated by any one of the four Gaussian distributions (marked with $\circ$'s, $\triangleleft$'s, $\diamond$'s, and $\triangleright$'s, respectively), but are masked by the union of them.

species, *Hybopsis storeriana*, *Notropis petersoni*, and *Luxilus zonatus*, which are masked when each is compared against the remaining species, are easily distinguished in the pair-wise tests (the last three rows in Table 3). Among all 90 comparisons, the top two largest equal error rates occur between *Notropis petersoni* and *Luxilus zonatus*: the detection rate for *Notropis petersoni* is 0.7791 (FAP is 0.2209) when it is tested against *Luxilus zonatus*; the detection rate for *Luxilus zonatus* is 0.75 (FAP is 0.25) when it is tested against *Notropic petersoni*. The above results demonstrate high potential for applying the proposed novelty detection algorithm in taxonomic research, specifically, to problems of new species discovery.

6. Conclusions

We have proposed a new statistical depth function, the kernelized spatial depth (KSD), and a novelty detection method using the KSD function. The KSD is a generalization of the spatial depth [30, 7, 35]. 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 novelty detection algorithm that identifies an observation as novel if its KSD value is smaller than a threshold. We derived an upper bound for the false alarm probability of a novelty detector, which can be applied to determine the threshold. Experimental results demonstrate high potential for applying the proposed novelty detection algorithm in taxonomic research, specifically, to problems of new species discovery.

Appendix

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 $f : \mathbb{X}^n \rightarrow \mathbb{R}$ satisfies*

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\hat{\mathbf{x}}_i \in \mathbb{X}} |f(\mathbf{x}_1,\ldots,\mathbf{x}_n) - f(\mathbf{x}_1,\ldots,\hat{\mathbf{x}}_i,\ldots,\mathbf{x}_n)| \leq c_i$$

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

$$\Pr[f(X_1,\ldots,X_n) - \mathbb{E}f(X_1,\ldots,X_n) \geq \epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}}.$$

Proof of Theorem 1: 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 $\frac{1}{\ell_{test}}$. Thus an application of the McDiarmid's inequality yields

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

Setting $\delta = \exp\left(-2\ell\epsilon^2\right)$ and solving for ϵ, we obtain $\epsilon = \sqrt{\ln(1/\delta)/2\ell}$. This completes the proof. □

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*, 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*, 97:4106–4111, 2000.

[3] A. Banerjee, P. Burlina, and C. Diehl. A support vector method for anomaly detection in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 44(8):2282–2291, 2006.

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

[5] C. Campbell and K. P. Bennett. A linear programming approach to novelty detection. *Advances in Neural Information Processing Systems*, 13:395–401, 2001.

[6] P. Chaudhuri. Multivariate location estimation using extension of R-estimates through U-statistics type approach. The Annals of Statistics, 20(2):897–916, 1992.

[7] P. Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872, 1996.

[8] L. Geng and H. J. Hamilton. Interestingness measures for data mining: a survey. *ACM Computing Surveys*, 38(3), 2006.

Table 3. Equal error rates for the pairwise test, i.e., the value that false alarm rate is equal to one minus the detection rate. A smaller value indicates a higher detection rate and a lower false alarm rate. The ij-th entry is the result for testing the unknown species i against the known species j.

	0.1313	0.1919	0.0000	0.1389	0.0943	0.0256	0.0000	0.0132	0.0116
0.1797		0.0523	0.0000	0.1389	0.0943	0.0256	0.0000	0.0132	0.0116
0.2109	0.0404		0.0000	0.1389	0.0943	0.0256	0.0000	0.0132	0.0116
0.0234	0.0034	0.0058		0.1389	0.0943	0.0256	0.0000	0.0132	0.0116
0.0234	0.0034	0.0058	0.0000		0.0943	0.0256	0.0000	0.0132	0.0116
0.0234	0.0034	0.0058	0.0000	0.1667		0.0256	0.0167	0.0789	0.0698
0.0234	0.0101	0.0058	0.0000	0.1389	0.0943		0.0000	0.0132	0.0116
0.0234	0.0034	0.0058	0.0000	0.1389	0.1321	0.0256		0.1316	0.1512
0.0234	0.0034	0.0058	0.0000	0.1389	0.1509	0.0256	0.1000		**0.2209**
0.0313	0.0034	0.0058	0.0000	0.1389	0.1321	0.0256	0.1000	**0.2500**	

[9] A. K. Ghosh and P. Chaudhuri. On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, 11(1):1–27, 2005.

[10] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

[11] 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*, 36(3):559–570, 2006.

[12] 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)*, 51–64, 2006.

[13] D. G. Kendall. Shape-manifolds, procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society*, 16:81–121, 1984.

[14] J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

[15] 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*, 86:415–427, 1991.

[16] R. Y. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

[17] M. Markou and S. Singh. A neural network-based novelty detection for image sequence analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1664–1677, 2006.

[18] 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*, 25(11):1468–1483, 2003.

[19] L. Parra, G. Deco, and S. Miesbach. Statistical independence and novelty detection with information preserving non-linear maps. *Neural Computation*, 8(2):260–269, 1996.

[20] S. L. Pimm and J. H. Lawton. Ecology–Planning for biodiversity. *Science*, 279:2068–2069, 1998.

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

[22] 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*, 24(9):1184–1199, 2002.

[23] 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*, 38(10):1760–1770, 1990.

[24] S. Roberts and L. Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6(2):270–284, 1994.

[25] J. E. Rodman and J. H. Cody. The taxonomic impediment overcome: NSF's partnerships for enhancing expertise in taxonomy (PEET) as a model. *Systematic Biology*, 52:428–435, 2003.

[26] P. J. Rousseeuw and I. Ruts. Algorithm AS 307: bivariate location depth. *Applied Statistics*, 45(4):516–526, 1996.

[27] I. Ruts and P. Rousseeuw. Computing depth contours of bivariate point clouds. *Computational Statistics and Data Analysis*, 23(1):153–168, 1996.

[28] 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*, 13(7):1443–1471, 2001.

[29] S. M. Schweizer and J. M. F. Moura. Hyperspectral imagery: clutter adaptation in anomaly detection. *IEEE Transactions on Information Theory*, 46(5):1855–1871, 2000.

[30] 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.), 25–38, 2002.

[31] D. E. Slice. Landmark coordinates aligned by procrustes analysis do not lie in Kendall's shape space. Systematic Biology, 50:141–149, 2001.

[32] I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.

[33] J. W. Tukey. Mathematics and picturing data. *Proc. 1975 Int'l Congress of Mathematics*, 2:523–531, 1974.

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

[35] Y. Vardi and C.-H. Zhang. The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1436, 2000.

[36] Y. Zuo and R. Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482, 2000.