Kernel Principal Subspace Mahalanobis Distances for Outlier Detection

Cong Li, Michael Georgiopoulos, and Georgios C. Anagnostopoulos

***Abstract*— Over the last few years, Kernel Principal Component Analysis (KPCA) has found several applications in outlier detection. A relatively recent method uses KPCA to compute the reconstruction error (RE) of previously unseen samples and, via thresholding, to identify atypical samples. In this paper we propose an alternative method, which performs the same task, but considers Mahalanobis distances in the orthogonal complement of the subspace that is utilized to compute the reconstruction error. In order to illustrate its merits, we provide qualitative and quantitative results on both artificial and real datasets and we show that it is competitive, if not superior, for several outlier detection tasks, when compared to the original RE-based variant and the One-Class SVM detection approach.**

I. Introduction

OUTLIER detection is also referred to as *novelty detection* or as *one-class classification*. Within this setting, the task amounts to designing a useful model to recognize atypical data (outliers) by using only *normal* data in the design phase, which are not considered to be outliers. Recent comprehensive surveys of the relevant field include [1], [2], [3] and [4].

Furthermore, in the context of machine learning, *kernel*-based methods [5] are computational techniques that, conceptually, involve an implicit transformation $\boldsymbol{\phi} : \mathbb{F} \rightarrow \mathbb{H}$ of the data from the original input space $\mathbb{F}$ to a new feature space $\mathbb{H}$ as a pre-processing step, such that $\boldsymbol{x} \mapsto \boldsymbol{\phi}_{\boldsymbol{x}}$. It is assumed that $\mathbb{H}$ is a Hilbert space, not necessarily finite-dimensional, equipped with a suitably defined inner product $\langle ., . \rangle_{\mathbb{H}} : \mathbb{H} \times \mathbb{H} \rightarrow \mathbb{R}$. Rather than requiring knowledge of the precise representations/images of the data in $\mathbb{H}$, these methods operate solely on the basis of inner products in $\mathbb{H}$, which are represented and computed by a *kernel* function $k : \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{R}$, such that $k(\boldsymbol{x}, \boldsymbol{y}) \equiv \langle \boldsymbol{\phi}_{\boldsymbol{x}}, \boldsymbol{\phi}_{\boldsymbol{y}} \rangle_{\mathbb{H}}$ and, thus, the mapping $\boldsymbol{\phi}$ is only used implicitly. Furthermore, it is typical with kernel-based methods to directly choose a function $k(.,.)$ that has the kernel property, *i.e.*, given k there is a $\boldsymbol{\phi}$, so that k is an inner product in $\mathbb{H}$. Obviously, the particular choice of kernel determines the underlying mapping $\boldsymbol{\phi}$, save an (improper, in general) rotation, and, therefore, the nature of space $\mathbb{H}$.

Cong Li is with the Department of EE & CS, University of Central Florida, Orlando, Florida , US (email: licong@knights.ucf.edu).

Michael Georgiopoulos is with the Department of EE & CS, University of Central Florida, Orlando, Florida , US (phone: +1 407 8235338; email: michaelg@mail.ucf.edu).

Georgios C. Anagnostopoulos is with the Electrical & Computer Engineering Department, Florida Institute of Technology, Melbourne, Florida , US (phone: +1 321 6747125; email: georgio@fit.edu).

Many traditional methods, whose computations involve information about the data only in the form of ordinary inner product values, have been extended to non-linear, kernel-based variants via the so called *kernel trick* (first utilized in [6]), which amounts to merely substituting the use of the ordinary Euclidean inner product in these methods with arbitrary kernel functions, thus, in effect, pre-transforming the data through the induced/implicit mapping. Such implicit transformations can prove desirable in certain cases, as they may favorably influence the data distributions involved. A classic example in pattern recognition is, when data, that are originally not linearly-separable in $\mathbb{F}$, are rendered linearly-separable in the induced space $\mathbb{H}$ through a proper choice of kernel function. Most remarkably, the kernel trick adaptation also allows these inner product reliant methods to be directly applied to non-numeric or mixed-type data, once appropriate kernels have been defined for these data types. As examples, here we can mention outlier detection techniques for categorical or mixed-attribute data such as [7] and [8].

Principal Component Analysis (PCA; for example, see [9]) is one of the methods that has been extended to a kernel-based PCA variant (KPCA; see [10]) thanks to an alternative way of computing the principal axes through the use of inner product evaluations [11]. KPCA has been used in several applications, such as face detection [12], image segmentation [13], feature extraction [14], data de-noising [15] and voice recognition [16], etc. Also recently, KPCA has found application in novelty detection [17][18]. In [17], KPCA is applied to the data using a Gaussian kernel and, subsequently, test data are projected to the resulting principal subspace, whose dimensionality is found experimentally. Test samples are identified as outliers, when their reconstruction error exceeds a certain threshold, which is also established experimentally. The reconstruction error itself is a measure of deviation from the principal subspace and, therefore, it assumes that the principal subspace represents the normal data. We claim that this assumption may not always be applicable and we offer some intuitive arguments, as well as experimental evidence supporting our thesis. Therefore, we advocate the use of Mahalanobis distance within the principal subspace as an alternative to the reconstruction error. It is worth noting that our approach is different with the method in [18], which detects outliers in the orthogonal complementary subspace of the principle subspace of our approach. In order to show the merit of our proposal, we conducted a series of experiments that considered both artificial and real datasets. We used the former ones to illustrate qualitative results and provide some insight into the past and novel approaches. For

the latter ones, which we obtained from the UCI Machine Learning Repository, we draw comparisons in the form of tables and figures based on the Area Under the Receiver Operating Characteristic (ROC) Curve [19].

The remainder of the manuscript is organized as follows: Section II provides some background material regarding KPCA and the use of the reconstruction error for outlier detection. Section III discusses our novel approach, while Section IV showcases the obtained experimental results. Finally, Section V summarizes the main outcomes of the paper.

II. Kernel PCA & Outlier Detection

A. KPCA Fundamentals

Assuming a set of N training data $\{\mathbf{x}_n\}_{n=1,\ldots,N}$ in $\mathbb{F}$ and an appropriately parameterized Mercer kernel $k : \mathbb{F} \times \mathbb{F} \rightarrow \mathbb{R}$, the Kernel Principal Component Analysis (Kernel PCA) computational procedure first entails forming the (symmetric) kernel matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ using the available training set patterns. Its (i, j) element equals $k(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the i^{th} and j^{th} training pattern respectively. Since the chosen kernel matrix implies a particular choice of feature mapping $\boldsymbol{\phi} : \mathbb{F} \rightarrow \mathbb{H}$ as mentioned earlier, the kernel matrix $\mathbf{K}$ represents the Gram matrix of the training data $\{\boldsymbol{\phi}_n\}_{n=1,\ldots,N} \hat{=} \{\boldsymbol{\phi}_{\mathbf{x}_n}\}_{n=1,\ldots,N}$ as they are embedded into space $\mathbb{H}$ via $\boldsymbol{\phi}$. Subsequently, the kernel matrix is centered to produce

$$\widetilde{\mathbf{K}} \hat{=} \mathbf{P}\mathbf{K}\mathbf{P} \tag{1}$$

where the orthogonal projection matrix $\mathbf{P}$ is defined as

$$\mathbf{P} \hat{=} \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T \tag{2}$$

In (2), $\mathbf{I}_N \in \mathbb{R}^{N \times N}$ is an identity matrix and $\mathbf{1}_N \in \mathbb{R}^N$ is an "all-ones" vector. $\boldsymbol{P}$ is also known as the *centering matrix* in the statistics community. Through (1), $\widetilde{\mathbf{K}}$ represents now the Gram matrix of the centered set $\left\{\widetilde{\boldsymbol{\phi}_n}\right\}_{n=1,\ldots,N}$, where we define

$$\widetilde{\boldsymbol{\phi}_n} \hat{=} \boldsymbol{\phi}_n - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\phi}_i \tag{3}$$

Next, the eigen-pairs $\{(\gamma_n, \mathbf{u}_n)\}_{n=1,\ldots,N}$ are obtained via the eigen-value decomposition of the centered kernel matrix $\widetilde{\mathbf{K}}$. We assume here that $\gamma_1 \geq \ldots \geq \gamma_N$ and that all eigenvectors $\mathbf{u}_n$ are normalized to unit L_2 length, i.e. $\|\mathbf{u}_n\|_2 = 1$. Then, it can be shown that the eigen-pairs of the (biased) sample covariance matrix of $\{\boldsymbol{\phi}_n\}_{n=1,\ldots,N}$ are given as

$$(\lambda_n, \mathbf{v}_n) = \left(\frac{\gamma_n}{N}, \pm\frac{1}{\sqrt{\gamma_n}}\widetilde{\boldsymbol{\Phi}}^T\mathbf{u}_n\right) \tag{4}$$

where $n = 1, \ldots, \dim(\mathbb{H})$ and $\widetilde{\boldsymbol{\Phi}}$ is the matrix, whose n^{th} row consists of $\widetilde{\boldsymbol{\phi}_n}$. Now, let us assume we have a test pattern $\mathbf{x}_t$, whose Kernel PCA transform we seek to compute. If we define the following quantities

$$\mathbf{A}_p \hat{=} [\mathbf{v}_1 \ldots \mathbf{v}_p]^T \tag{5}$$

where $p \leq \min\{\dim(\mathbb{F}), N\}$ and

$$\mathbf{k}(\mathbf{x}_t) \hat{=} [k(\mathbf{x}_t, \mathbf{x}_1) \ldots k(\mathbf{x}_t, \mathbf{x}_N)]^T \tag{6}$$

and, finally,

$$\widetilde{\mathbf{k}}(\mathbf{x}_t) \hat{=} \mathbf{P}\left[\mathbf{k}(\mathbf{x}_t) - \frac{1}{N}\mathbf{K}\mathbf{1}_N\right] \tag{7}$$

then the Kernel PCA transform $\mathbf{y}_t$ of $\mathbf{x}_t$ with respect to the p principal Kernel PCA eigen-directions in the feature space H is given as

$$\widetilde{\mathbf{y}}_t = \mathbf{A}_p^T\widetilde{\mathbf{k}}(\mathbf{x}_t) \tag{8}$$

Obviously, the results of the transformation strongly depend on the particular inner product kernel employed, the particular values of its parameters and on the dimensionality p of the principal subspace, on which the data are finally projected. As mentioned earlier in Section I, Kernel PCA has been used in a variety of settings and applications including, of course, outlier detection as a data transformation method. In the next section we briefly discuss a relatively recent Kernel PCA-based approach to outlier detection and lay out the details of our approach.

B. Outlier Detection using Kernel PCA Reconstruction Error

Recently, in [17] a Kernel PCA-based outlier detection method was introduced and its performance was showcased in comparison to other established, kernel-based methods. The particular approach chooses a training set and a suitable projection dimensionality p, proceeds to compute the Kernel PCA transform of another set of test patterns and, finally, computes the *reconstruction error* (RE) for each of these test patterns. The squared reconstruction error $r^2(\mathbf{x}_t)$ for a test pattern $\mathbf{x}_t$ is defined as

$$r^2(\mathbf{x}_t) \hat{=} \left\|\widetilde{\boldsymbol{\phi}_t}\right\|_2^2 - \|\widetilde{\mathbf{y}}_t\|_2^2 \tag{9}$$

where $\widetilde{\boldsymbol{\phi}_t} \hat{=} \widetilde{\boldsymbol{\phi}}(\mathbf{x}_t)$. Again, the norms involved in these expressions are of the L_2 (Euclidean) variety and it holds that $\|\mathbf{z}\|_2^2 = \langle \mathbf{z}, \mathbf{z} \rangle_H$ for any vector $\mathbf{z} \in H$. In light of (3) and noticing that $\left\|\widetilde{\boldsymbol{\phi}_t}\right\|_2^2 = \left\langle \widetilde{\boldsymbol{\phi}_t}, \widetilde{\boldsymbol{\phi}_t} \right\rangle_{\mathbb{H}}$ we obtain that

$$\left\langle \widetilde{\boldsymbol{\phi}_n}, \widetilde{\boldsymbol{\phi}_n} \right\rangle_{\mathbb{H}} = k(\mathbf{x}_t, \mathbf{x}_t) - \frac{2}{N}\mathbf{1}_N^T\mathbf{k}_t + \frac{1}{N^2}\mathbf{1}_N^T\mathbf{K}\mathbf{1}_N \tag{10}$$

Thus, (10) and (8) make it possible to calculate the reconstruction error via (9). Given the projection dimensionality p, outliers are identified as data points, whose RE exceeds an appropriately established threshold value r_{thres}.

In [17], the optimal combination of dimensionality and threshold are determined simultaneously by trial-and-error through, essentially, cross-validation: a hold-out set is used to identify the pair of values that maximize the detection

rate. When using a Gaussian kernel, $\dim \mathbb{H} = \infty$ and, therefore $p \in 1, 2, \ldots, N$. Furthermore, if N_t hold-out set patterns are used to optimize (p, r_{thres}), then $N \times N_t$ reconstruction errors have to be computed and compared to different threshold values in order to assess detection rates. Experimental results on both artificial and real datasets using this approach have been shown to be very competitive to other kernel-based outlier detection methods considered in the same paper. For example, it showed that the method is more robust to noise than One-Class SVM [20].

III. KPCA Mahalanobis Distances for Outlier Detection

It is of particular interest to express the reconstruction error as a suitably scaled distance of transformed input patterns from the sample mean of all transformed training patterns. When we combine (9) and (8), after some algebraic manipulations we can re-express the squared reconstruction error as

$$r^2(\mathbf{x}_t) = \left\| \mathfrak{T}_p^{\perp} \widetilde{\phi}_t \right\|_2^2 \tag{11}$$

where $\mathfrak{T}_p^{\perp}$ is an operator in $\mathbb{H}$ that orthogonally projects onto span $\{\mathbf{v}_1, \ldots \mathbf{v}_p\}^{\perp}$, *i.e.* the orthogonal complement of the p-dimensional principal subspace; as a reminder, the $\mathbf{v}_i$'s are the eigen-vectors spanning the principal subspace. Thus, given a test pattern and assuming a pre-determined dimension p for the principal subspace, its reconstruction error could be computed by first orthogonally projecting the pattern onto the orthogonal complement of the principal subspace and then calculating the distance from the sample mean of all transformed training patterns. Apparently, the intuition behind the reported success of this particular approach in measuring deviation from normalcy and, thus, detecting outliers touches on a simple assumption: normal (non-outlier) data lie on the principal subspace or, equivalently, the principal subspace represents normal data, while anything not belonging to it is deemed to be an outlier. Using a non-zero threshold on the reconstruction error allows for a relaxation of this condition and may be used for fine-tuning the detection rates.

Nevertheless, the aforementioned assumption may not always be suitable for a given distribution of normal data in $\mathbb{H}$. First of all, the implicit transform to $\mathbb{H}$ via a kernel, obviously, preserves the intrinsic dimensionality of the original data. More precisely, the transformed data are going to be mapped on a q-dimensional manifold embedded in $\mathbb{H}$, where $q \leq \dim(\mathbb{F})$. Even samples that in the original feature space were considered as outliers are going to be mapped onto the same manifold. Since the role of the Kernel PCA-derived principal subspace is to approximate this manifold, it seems as if the usage of the reconstruction error as an outlier detection device is not appropriate, at least, in most cases. An example of such a scenario is given in Figure 1, where, after being transformed via ϕ, data points are embedded in a 3-dimensional feature space, but occupy an almost linear and almost 2-dimensional manifold. The dimensions here are deliberately chosen small to allow visualization of the relevant distance concepts. Points with reconstruction error less than r fall between the the two outer planes. Points with Mahalanobis distance less than d are in the interior of the indicated ellipse. As the data points are mapped very close to the middle plane, using a threshold-based outlier detection rule that relies on Mahalanobis distances rather than reconstruction errors seems to be more appropriate here. Of course, this example is conceptual, but the approach seems to be the of merit for other data sets and outlier detection problems, as is demonstrated in this paper's experimental findings.

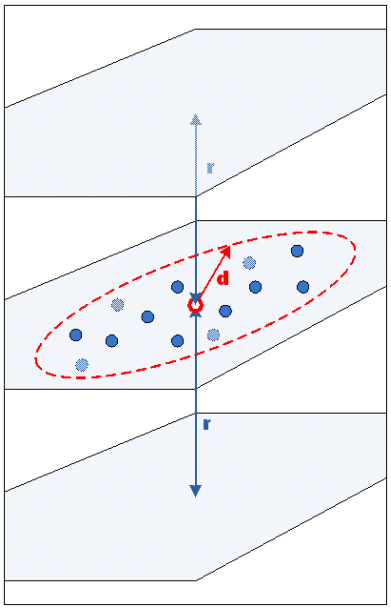


Fig. 1. Conceptual comparison of Reconstruction Error and Mahalanobis distance in the principal subspace.

In this paper we advocate the use of the Mahalanobis distance in the principal subspace as a normalcy indicator for KPCA-transformed data and, of course, as a complementary method to the use of the associated reconstruction error. As we show in Section IV, there are outlier detection problems that benefit from this alternative view. Using the fact that the squared Mahalanobis distance for a p-dimensional principal subspace would be defined as

$$\left\| \widetilde{\phi}_t \right\|_M^2 \hat{=} \sum_{n=1}^{p} \frac{1}{\lambda_n} \left\langle \mathbf{v}_n, \widetilde{\phi}_t \right\rangle_{\mathbb{H}}^2 \tag{12}$$

and based on (4) and (8), we derive that

$$\left\| \widetilde{\phi}_t \right\|_M^2 = N \mathbf{y}_t^T \mathbf{\Gamma}^{-1} \mathbf{y}_t \tag{13}$$

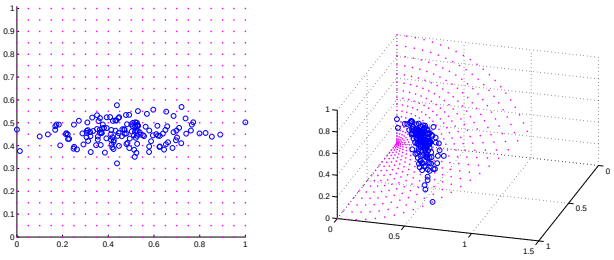
where $\mathbf{\Gamma} \hat{=} diag\{\gamma_1, \ldots, \gamma_p\}$. In other words, the Mahalanobis distance in question can be calculated purely in terms of kernel evaluations and knowledge of the implicit mapping is not required. The particular choice of kernel and/or kernel parameters, as well as the threshold to use for outlier detection, is again a matter resolved through cross-validation.

IV. Experimental Results

A. Qualitative Results

In order to qualitatively illustrate the potential merit of our proposed outlier detection method, we first created an artificial dataset of 150 2-D patterns drawn from the Gaussian distribution with mean $[0.5 \;\; 0.5]^T$ and 2×2 diagonal covariance matrix with elements 1 and 0.1. The patterns were subsequently scaled to fit in $[0,1]^2$. Moreover, we formed a square grid of test data points in $[0,1]^2$. The distribution of the training and test patterns are shown in Figure 2a.

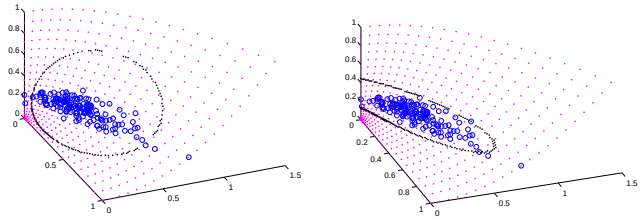
Next, the polynomial kernel of degree $d = 2$, *i.e.* $\boldsymbol{k}(\boldsymbol{x},\boldsymbol{y}) = \left(\boldsymbol{x}^T\boldsymbol{y}\right)^2$, is employed to map both the training and test data into the feature space. The particular kernel maps 2-dimensional data $\boldsymbol{x} = [x_1 \;\; x_2]^T$ into a 3-dimensional space via the implicit mapping $\boldsymbol{\phi}(\boldsymbol{x}) = \left[x_1^2 \;\; \sqrt{2}x_1x_2 \;\; x_2^2\right]^T$ (*e.g.* see [21]). We show the distribution of both training patterns and test data in the 3-D feature space in Figure 2b.



(a) (b)

Fig. 2. Artificial dataset and its implicit mapping via a 2^{nd}-order polynomial kernel.

We then compare the use as outlier discriminants of the Mahalanobis Distance (MD) in the principal subspace as described by (13) and the Reconstruction Error (RE) in the principal subspace's orthogonal complement as described in (11). In this example, the principal subspace is of dimension 2. In Figure 3a, the contour with RE value $RE_{max}/2$ is shown, where RE_{max} is the largest RE of the training points. On the other hand, in Figure 3b, the contour with MD value $MD_{max}/2$ is indicated, where MD_{max} is the largest Mahalanobis distance in the principal subspace of the training points. This side-by-side comparison reveals that, under certain circumstances, the RE may not be an effective measure of deviation from normalcy, when compared to using the MD. In the case depicted, we see that RE produces a decision boundary that is overly broad and, thus, one that does not satisfactorily fit the normal (training) data; many potential outliers would not be detected. However, the MD-induced boundary seems to capture much better the overall structure of the normal data. While this particular example is not necessarily typical of one encountered in practice, we believe that it explains the advantages of our proposed MD-based method over the RE-based method and the outcomes we have observed in our experiments.



(a) (b)

Fig. 3. Constant RE (left figure) and MD (right figure) contours based on the first two principle components.

B. Quantitative Results

In this section, we experimentally examine the outlier detection ability of our proposed MD-based method by comparing it with the RE-based method and One-class SVM (OCSVM). The Area Under the Receiver Operating Characteristic curve (AUC) is employed as a comparison criterion as in [19]. The Receiver Operating Characteristic (ROC) curve is a 2-D curve, which captures the relationship between the false positive and the true positive rate. An ideal outlier detector achieves 0 false positive rate and 1 true positive rate, which results to an AUC value equal to 1. In practice though, the AUC lies in the interval $[0,1]$, and the larger AUC is achieved, the better the corresponding outlier detector performs.

For our comparisons, we choose both numerical and categorical datasets. For the numerical datasets, the Gaussian kernel is applied, while the Hamming kernel [7] is used for categorical datasets. For each of the three methods, we used a variety of parameter settings. In specific, when using MD and RE, the parameters include the dimension of the principle subspace p and the kernel parameters (σ for Gaussian kernel and λ for Hamming kernel). Also, the OCSVM features the ν parameter from the One-class SVM formulation and the aforementioned kernel parameters. The dimension of the principle subspace p is searched from 1 to the number of available training samples N_{tr}. The Gaussian kernel's spread parameter σ is searched from 0.2 to 30 with step size of 0.2. The Hamming kernel's λ parameter is searched from 0.02 to 0.98 with a step size of 0.02 and, finally, OCSVM's ν parameter is searched from 0.01 to 0.99 with a step size of 0.01.

1) Datasets: Eight datasets are considered in our experiments. Four of them are numerical datasets, and the other four are categorical datasets.

- *SPECTF Heart (SPECTF)*. The dataset consists of 267 instances with 45 integer attributes. The attributes describe the characteristics of different Region of Interest (ROI) of the Single Proton Emission Computed Tomography (SPECT) images. There are 55 abnormal instances in class 0 and 212 normal instances in class 1. In the experiment, 100 class 1 samples are used for training as normal data, while the test set consisted of 55 class 1 data and 55 class 0 outliers.

- *Thyroid Disease (Thyroid)*. This dataset features 3772 instances, which belong to 3 different classes: hyperfunction, subnormal, and normal. Each instance has 21 real-valued attributes. 200 out of 3488 class 3 are considered normal and, therefore, are used for training. 93 class 3 data and all the 93 class 1 (hyperfunction) samples comprise the test set.
- *Wine Quality (Wine)*. It contains two parts: Red wine quality and White wine quality. We only use Red wine samples in our experiment. 1599 instances with 11 real attributes are separated into 10 levels of quality, which constitute 10 classes. Instances of level 5 and 6 form the majority group. Therefore, we use 100 out of 681 class 5 instances for training, while the test set consists with 53 class 5 instances and all the 53 class 4 as outliers.
- *Yeast*. *Yeast* dataset samples consist of 9 real-valued attributes. Each of the 1484 instances comes from one of 10 different classes, which indicate the different localization site of a given protein. Most data are from the 3 classes, which represent the majority group, while the other 7 classes have only a few patterns. Therefore, in our experiment, 100 patterns from class 3 are treated as normal data and are used in the training phase, and 50 class 3 patterns as well as 50 class 5 outliers are used for testing.
- *Balance*. 4 categorical attributes are recorded for each of the 625 instances which model psychological experimental results. The dataset consists of 288 points of class *L*, 288 points of class *R*, and 49 points of class *B*. 100 samples from class *R* are used in the training process, while the test set consists of 49 class *R* data and 49 class *B* data. Our goal is to detect class *B* instances, which are treated as outliers.
- *Chess (King-Rook vs. King-Pawn) (Chess)*. 3196 instances with 36 categorical attributes are recorded to describe the state of a chess board, and two outcomes, *i.e.* white-can-win "won" and white-cannot-win "nowin", which form 2 classes. In our experiments, 300 samples from class "won" are used in the training process, and 150 samples from class "won" and 150 data from class "nowin" are used for testing. The goal in this particular problem is to detect the "nowin" patterns in the test set.
- *Tic-Tac-Toe Endgame (TTT)*. This data set has 958 instances. 9 categorical attributes represent the 9 positions on tic-tac-toe board, and the final configurations of the board at the end of games are recorded. The two results, *i.e.* "positive" and "negative", indicate 2 classes. To detect the "negative" data, 300 "positive" patterns are used for training, while the test set consists of 150 "positive" and 150 "negative" samples.
- *SPECT Heart (SPECT)*. This data set is similar to *SPECTF* data set. It also contains 267 instances. But unlike *SPECTF*, each instance of *SPECT* data set has 22 binary attributes, which summarize the original SPECT image, instead of the numerical attributes contained in the *SPECTF* data set. 100 class 1 data, which belong to the normal class, are used in training. On the other hand, the test set consists of 55 class 1 and 55 class 0 data points, while class 0 samples are considered as abnormal instances.

2) Observations: We provide the experimental results for the 4 numerical datasets in Table I and the 4 categorical datasets in Table II. For each dataset and each outlier detection method, we calculate the AUCs for all parameter settings, *i.e.* the combinations of kernel parameter and model parameter values that are mentioned in the beginning of this section. Then, we report the best AUC value in the first table row for each dataset. Also, we report the minimum, 25% percentile, the median, and the 75% percentile of the AUC distributions in the remaining four rows respectively.

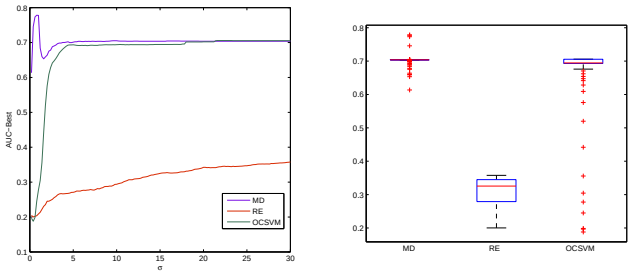
TABLE I
EXPERIMENTAL RESULTS FOR MD-KPCA, RE-KPCA, AND OCSVM ON THE NUMERICAL DATASETS.

	MD	RE	OCSVM
SPECTF	**0.7782**	0.3577	0.7058
	0.2322	0.1867	0.1834
	0.2575	0.2055	**0.4432**
	0.2686	0.2224	**0.5093**
	0.2859	0.2658	**0.6409**
Thyroid	**0.9820**	0.9670	0.9307
	0.0435	0.2060	**0.6064**
	0.9027	0.8533	0.6152
	0.9651	0.8997	0.6242
	0.9760	0.9285	0.6396
Wine	**0.7957**	0.7818	0.7911
	0.2129	0.4978	**0.6727**
	0.7106	**0.7246**	0.7176
	0.7292	**0.7376**	0.7283
	0.7412	**0.7462**	0.7429
Yeast	0.8966	0.8711	**0.9083**
	0.1183	0.2009	**0.4101**
	0.7669	0.7045	**0.8616**
	0.7957	0.7525	**0.9083**
	0.8183	0.7913	**0.8796**

In Table I, by observing the first row of each dataset, our MD-based method achieves the highest AUC on *SPECTF*, *Thyroid*, and *Wine* datasets. For the *Yeast* dataset, OCSVM achieves the best solution with a small advantage over the proposed MD-based method. It is interesting to note that in the other four rows of the results for the *SPECTF* dataset, the performances of the MD-based approach are not good for most parameter settings, since the AUC values are below the 75% percentile, while the best are much higher than most other AUC values. This can also be seen in Figure 4 for the *SPECTF* dataset. In Figure 4a, the x-axis represents the kernel spread values σ, and the y-axis is the highest AUC value that is obtained under all possible model parameter values (p for MD-KPCA and RE-KPCA and ν for OCSVM) as σ changes. Figure 4b is the Box plot of the AUCs in Figure 4a. It can be seen from Figure 4b that the midrange, *i.e.* from the 25% to the 75% percentiles, is small for all the three methods, while the best AUC value of the MD-based

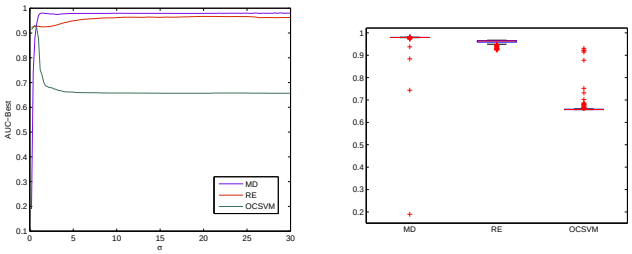
method is identified as an outlier in the Box plot. Figure 4a shows that our method performs better for most σ's.

Unlike the results of *SPECTF* dataset, the last three rows of *Thyroid*, *Wine* and *Yeast* datasets show that the AUC values do not change that much at different percentiles. In other words, all the three methods have a small AUC range for varying kernel parameter σ. This can be seen from Figure 5 to Figure 7, which depict the results for the *Thyroid*, *Wine* and *Yeast* datasets respectively. It can be seen from these figures that, for these three datasets, the AUC values change slightly with changes in σ for all the three approaches (except the result of OCSVM for the *Thyroid* dataset), which implies detection robustness for numerical datasets, when a Gaussian kernel is used. By observing these four figures, the minimum AUC values are all attained, when the parameter σ is very small, which implies that, in practice, care should be taken when choosing small σ values.



(a) (b)

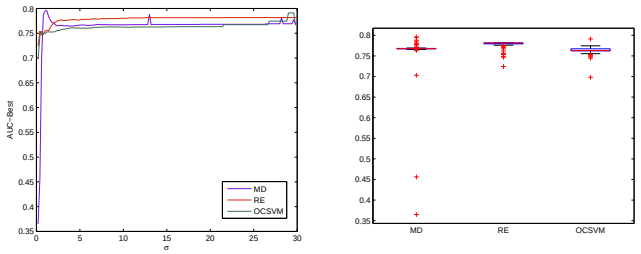
Fig. 4. *SPECTF* dataset: Maximum AUC value versus kernel parameter value and corresponding Box plot.



(a) (b)

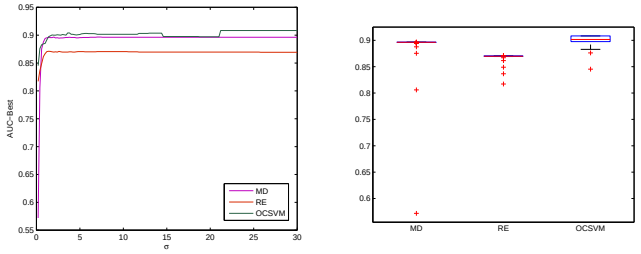
Fig. 5. *Thyroid* dataset: Maximum AUC value versus kernel parameter value and corresponding Box plot.

In Table II, it can be seen that the MD-based method performs very well on the *Balance*, *Chess*, and *TTT* categorical datasets. Not only does it attain the highest AUC value among the three methods, but its AUC distribution is superior to the other corresponding distributions, as witnessed when drawing comparisons among the best 25% percentile, median, and 75% percentile outcomes. Figure 8 through Figure 10 shows the relationship between the AUC and the kernel parameter λ for the three datasets. Like Figure 4, the left plots show the highest AUC value obtained with all possible model parameters as λ changes, while the right plots



(a) (b)

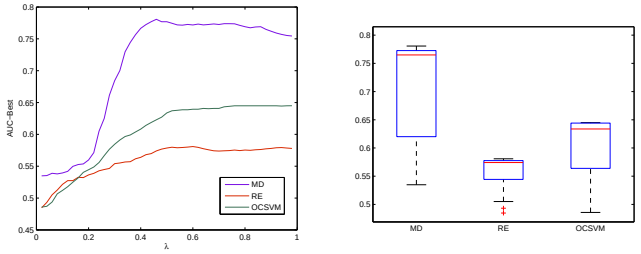
Fig. 6. *Wine* dataset: Maximum AUC value versus kernel parameter value and corresponding Box plot.



(a) (b)

Fig. 7. *Yeast* dataset: Maximum AUC value versus kernel parameter value and corresponding Box plot.

are the Box plots for the corresponding AUC distributions in the left counterparts. It can be seen that KPCA-based detection using Mahalanobis Distances outperforms the other two for most λ values in the three datasets. Our proposed approach slightly underperformed for the *SPECT* dataset for some λ values, as shown in Figure 11, while the AUC values obtained via our method are still high in terms of the 25% percentile and the median comparing to the RE-based method. Based on these results for the four categorical datasets, it can be seen that the AUC range for all three methods is wide, which obviously implies that the three approaches are sensitive to the specific choice of the kernel parameter λ and, thus, care needs to be taken, when choosing values for it.



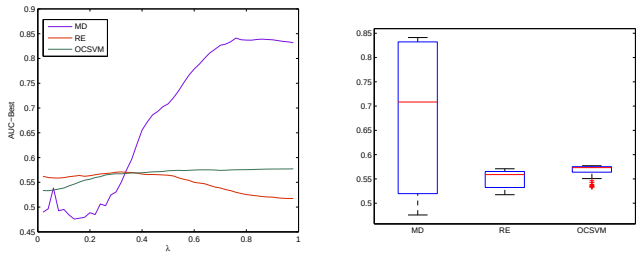
(a) (b)

Fig. 8. *Balance* dataset: Maximum AUC value versus kernel parameter value and corresponding Box plot.

TABLE II

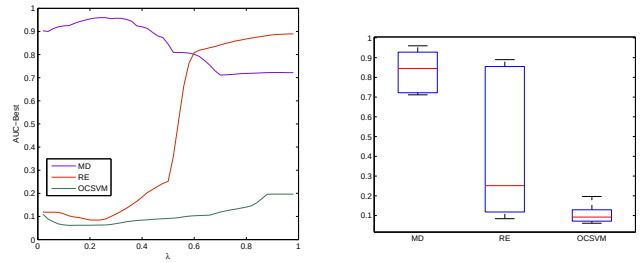
EXPERIMENTAL RESULTS FOR MD-KPCA, RE-KPCA, AND OCSVM ON THE CATEGORICAL DATASETS.

	MD	RE	OCSVM
Balance	**0.7807**	0.5808	0.6449
	0.3320	0.1999	**0.4666**
	0.5342	0.2227	**0.5459**
	0.6568	0.3625	0.5820
	0.7091	0.4504	0.5983
Chess	**0.8412**	0.5709	0.5772
	0.2559	0.2028	**0.4118**
	0.4637	0.2809	**0.5068**
	0.6171	0.4110	0.5391
	0.7520	0.5217	0.5660
TTT	**0.9597**	0.8895	0.1966
	0.0045	0.0139	**0.0341**
	0.1833	0.0782	0.0538
	0.5614	0.2085	0.0721
	0.8671	0.8476	0.0849
SPECT	0.7858	**0.8402**	0.3751
	0.1550	**0.1778**	0.1291
	0.1960	**0.2004**	0.1478
	0.5854	0.2539	0.1650
	0.7278	**0.8119**	0.1806



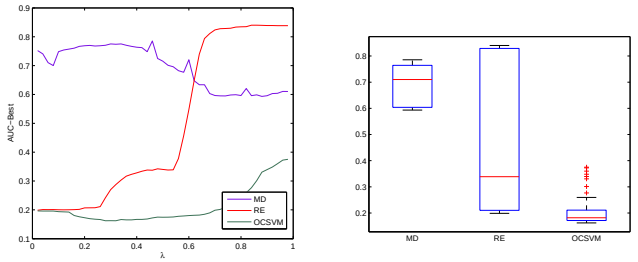
(a) (b)

Fig. 9. *Chess* dataset: Maximum AUC value versus kernel parameter value and corresponding Box plot.



(a) (b)

Fig. 10. *TTT* dataset: Maximum AUC value versus kernel parameter value and corresponding Box plot.



(a) (b)

Fig. 11. *SPECT* dataset: Maximum AUC value versus kernel parameter value and corresponding Box plot.

V. CONCLUSIONS

In this paper we discuss two approaches on using Kernel Principal Component Analysis (KPCA) for detection of atypical samples. Both methods rely on first mapping samples, that are considered typical (non-outliers), to a Hilbert (feature) space reproduced by the specific inner-product kernel employed by KPCA. Subsequently, they identify the principal subspace of the aforementioned transformed dataset. Given a test sample, the first method, which has been introduced in [17], measures its reconstruction error (RE) in the feature space and identifies the sample as an outlier, if the RE exceeds an adjustable threshold. The RE measured depends on the particular characteristics of the principal subspace's orthogonal complement. We provide arguments implying that, because of the aforementioned fact, RE may not always be an effective measure of deviation from normalcy. Based on this motivation, we present an alternative approach that utilizes the principal subspace Mahalanobis Distance (MD) of a test sample from the transformed dataset's sample average as such a measure instead. We argue that this approach can be justified by intuition and could be applicable in practice.

In order to showcase the merits of our proposed approach, we performed a number of experiments that compared the capability of detecting outliers in data of the One-Class SVM, the RE-based, and the MD-based KPCA detection methods. The experimental results reveal, as expected, that the performance robustness and detection quality of all three methods compared varies from one dataset to the next and, furthermore, depends on the particular choice of parameterized kernel function that is utilized. Nevertheless, the outcomes indicate that the MD-based KPCA method is competitive, if not superior, in detecting true outliers, when compared to the other two approaches.

ACKNOWLEDGMENT

Cong Li acknowledges partial support from NSF grants No. 0647120, No. 0717680, No. 0806931 and No. 0837332. Moreover, Michael Georgiopoulos acknowledges partial support from NSF grants No. 0525429, No. 0837332 and No. 0717680. Finally, Georgios C. Anagnostopoulos acknowledges partial support from NSF grants No. 0717674 and No. 0647018. Any opinions, findings, and conclusions or

recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Finally, the authors would like to thank the anonymous reviewers of this manuscript for their time and helpful comments.

REFERENCES

[1] M. Markou and S. Singh, "Novelty detection: A review – part 1: Statistical approaches," *Signal Processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[2] ——, "Novelty detection: A review – part 2: Neural network based approaches," *Signal Processing*, vol. 83, no. 12, pp. 2499–2521, 2003.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection : A survey," *ACM Computing Surveys*, vol. 41, no. 3, July 2009.

[4] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, October 2004.

[5] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, ser. Adaptive Computation and Machine Learning, T. Dietterich, Ed. Cambridge, Massachusetts, US: The MIT Press, 2002.

[6] M. Aizerman, E. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, pp. 821–837, 1964.

[7] J. Couto, "Kernel k-means for categorical data," *Lecture Notes in Computer Science*, vol. 3646, pp. 46–56, 2005.

[8] S. Guo, L. Chen, and J. Tsai, "A boundary method for outlier detection based on support vector domain description," *Pattern Recognition*, vol. 42, pp. 77–83, 2009.

[9] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 6, pp. Philosophical Magazine, Vol. 2, No. 6. (1901), pp. 559–572., 1901.

[10] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299 – 1319, July 1998.

[11] J. McLaughlin and J. Raviv, "Nth-order autocorrelations in pattern recognition," *Information and Control*, vol. 12, pp. 121–142, 1968.

[12] K. I. Kim, K. Jung, and H. J. Kim, "Face recognition using kernel principal component analysis," *IEEE Signal Processing Letters*, vol. 9, no. 2, pp. 40–42, 2002.

[13] C. Alzate and J. Suykens, "Image segmentation using a weighted kernel PCA approach to spectral clustering," in *Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007. IEEE Symposium on*, April 2007, pp. 208–213.

[14] R. Rosipal, M. Girolami, L. J. Trejo, and A. Cichocki, "Kernel PCA for feature extraction and de-noising in nonlinear regression," *Neural Computing & Applications*, vol. 10, no. 3, pp. 231–243, 2001.

[15] S. Mika, B. Schölkopf, A. Smola, K.-R. Muller, M. Scholz, and G. Ratsch, "Kernel PCA and de-noising in feature spaces," in *Proceedings of the 1998 conference on Advances in neural information processing systems II*, 1999, pp. 536 – 542.

[16] M.-S. Kim, I.-H. Yang, and H.-J. Yu, "Robust speaker identification using greedy kernel PCA," in *Tools with Artificial Intelligence, 2008. ICTAI '08. 20thIEEE International Conference on*, 2008, pp. 143–146.

[17] H. Hoffmann, "Kernel PCA for novelty detection," *Pattern Recogn.*, vol. 40, no. 3, pp. 863–874, 2007.

[18] Y. Shen and A. J. Izenman, "Outlier detection using the smallest kernel principal components with radial basis function kernel," in *Joint Statistical Meetings*, 2008.

[19] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997.

[20] 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, Jul. 2001.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.